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

Xiaomin Li*
Harvard University

Zhou Yu
Amazon

Ziji Zhang
Amazon

Yingying Zhuang
Amazon

Swair Shah
Amazon

Narayanan Sadagopan
Amazon

Anurag Beniwal
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.

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 et al., 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 et al., 2023, Liu et al., 2024, Quevedo et al., 2024]. Existing methods focus on **internal uncertainty**, which generally

*Work done during internship at Amazon. Correspondence to: Xiaomin Li (email: xiaominli@g.harvard.edu)

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Third Workshop on Structured Probabilistic Inference and Generative Modeling.

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 et al., 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 et al., 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 et al., 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 et al., 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 et al., 2023, Farquhar et al., 2024, Kossen et al., 2024, Lin et al., 2023]. A representative method in this category is the *Semantic Entropy*, which clusters sampled responses into semantic equivalence classes and computes entropy across these clusters [Kuhn et al., 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 [2].
- We provide **theoretical interpretations** of our semantic volume, linking it to differential entropy and generalizing existing sampling-based uncertainty measures.

[2]Code of our experiments: `https://anonymous.4open.science/r/LLMUncertainty-5C61`.

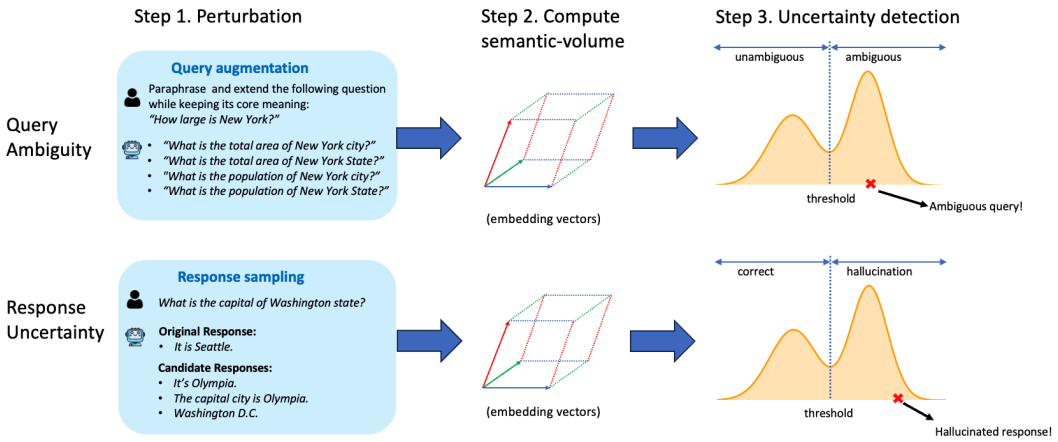


Figure 1: Pipeline for external and internal uncertainty detection using semantic volume. *Step 1.* Generate perturbations. For external uncertainty, we augment each query using an LLM, treating the augmentations as perturbations. For internal uncertainty, perturbations refer to multiple sampled candidate 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).

2 Related Work

External Uncertainty. Query ambiguity detection is typically performed using LLMs with various prompting techniques [Kuhn et al., 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 et al. [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.

Internal Uncertainty.

(1) 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 token probabilities, such as the minimal and averaged token probabilities.

(2) Sampling-based. Kuhn et al. [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 et al. [2023]. The *Lexical Similarity* method [Fomicheva et al., 2020, Grewal et al., 2024] considers the averaged similarity of the sampled answers, and lower similarity indicates higher dispersion.

3 Method

3.1 Definitions and Notations:

Denote $[k] \stackrel{\text{def}}{=} \{1, 2, \ldots, k\}$ 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 $\boldsymbol{V} = [\boldsymbol{v}_1 \boldsymbol{v}_2 \dots \boldsymbol{v}_n]$ (each $\|\boldsymbol{v}_i\| = 1$), we define the squared volume as:

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

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 C.

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 F, 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

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

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

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

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 M, 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 $\tilde{\boldsymbol{V}} = [\tilde{\boldsymbol{v}}_1 \tilde{\boldsymbol{v}}_2 \dots \tilde{\boldsymbol{v}}_n] \in \mathbb{R}^{d \times n}$ for n perturbations. Then compute the semantic volume according to equation 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 [3] (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

[3] 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.

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 L. `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 for internal uncertainty 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 et al., 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 et al., 2023, Fomicheva et al., 2020]: computes the averaged pairwise similarity of perturbed queries. (c) **Semantic Entropy** [Kuhn et al., 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 H). 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 generally 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.

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 $\pm$ std. over three trials (std. as subscript).

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] [4]. 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 I).

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 et al., 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 J). 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

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

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 $\pm$ std. over three trials (std. shown as subscript).

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 J). 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].

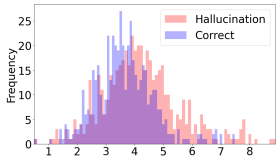
5.3 Distribution Separation

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.

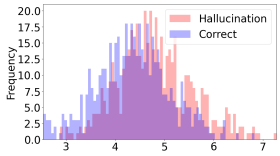
The plots and statistics are shown in Figure 2. Combined the KS statistics and the histograms, we observe that indeed *Last Token Entropy* and *Log Probabilities* struggle to effectively separate the two groups of data, while *Lexical Similarity*, *Semantic Entropy*, and our *Semantic Volume* exhibit 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 Appendix D 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

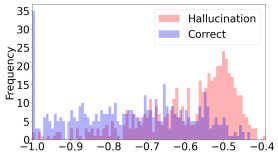
The original embedding dimension from the sentence-transformer is 1536. In the external uncertainty experiment, we reduce the dimensionality using PCA with $d = 10$, while in the internal uncertainty experiment, we set $d = 20$. Ablation studies confirm PCA-projected vectors $\hat{\boldsymbol{V}}$ outperform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



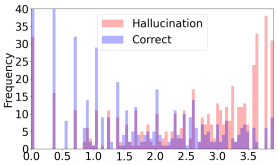
(a) Last Token Entropy (KS Statistic 0.238 with p-value 1e-36).



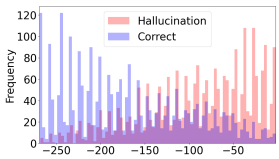
(b) Log Probabilities (KS Statistic 0.226 with p-value 5e-27).



(c) Lexical Similarity (KS Statistic 0.428 with p-value 1e-26).



(d) Semantic Entropy (KS Statistic: 0.376 with p-value 7e-32).



(e) Semantic Volume (KS Statistic 0.446 with p-value 3e-54).

Figure 2: Distribution of subsets for both labels across different uncertainty measures. Two-sample Kolmogorov–Smirnov statistics are computed to quantify the separation of two bulks.

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 H, 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 I and J, 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 Conclusion

One limitation of our current study is that our work considers the same scope as the references in Section 2: 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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Meta AI. Llama 3.2: Advancing ai for vision, edge, and mobile devices. Meta AI Blog, 2024. URL `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`. Accessed: Feb 11, 2025.

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Anthropic. Introducing the Claude 3 family of models, 2023. URL `https://www.anthropic.com/news/claude-3-family`. Accessed: 2025-01-02.

Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*, 2021.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Alexander Bastounis, Paolo Campodonico, Mihaela van der Schaar, Ben Adcock, and Anders C Hansen. On the consistent reasoning paradox of intelligence and optimal trust in ai: The power of'i don't know'. *arXiv preprint arXiv:2408.02357*, 2024.

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

Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. Towards improving faithfulness in abstractive summarization. *Advances in Neural Information Processing Systems*, 35:24516–24528, 2022.

Yizhou Chi, Jessy Lin, Kevin Lin, and Dan Klein. Clarinet: Augmenting language models to ask clarification questions for retrieval. *arXiv preprint arXiv:2405.15784*, 2024.

Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Yashvir S Grewal, Edwin V Bonilla, and Thang D Bui. Improving uncertainty quantification in large language models via semantic embeddings. *arXiv preprint arXiv:2410.22685*, 2024.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

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

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

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

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. Aligning language models to explicitly handle ambiguity. *arXiv preprint arXiv:2404.11972*, 2024.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

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

Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. Asking clarification questions to handle ambiguity in open-domain qa. *arXiv preprint arXiv:2305.13808*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Chang Yue, and Hong Hu. Rule-based data selection for large language models. *arXiv preprint arXiv:2410.04715*, 2024.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*, 2024.

LMSys. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. LMSys Blog, 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`. Accessed: Feb 11, 2025.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.

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

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach. *arXiv preprint arXiv:2405.19648*, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. URL https://arxiv.org/abs/1606.05250.

N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Siddhartha Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b. URL https://arxiv.org/abs/2307.09288.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

Michael JQ Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*, 2023.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models. *arXiv preprint arXiv:2405.12063*, 2024.

Appendix

A More Related Work **13**

B Semantic-Volume Uncertainty Detection Algorithm **13**

C Geometric interpretation as volume **14**

D Theoretical Analysis **14**

E Proofs **15**

E.1 Proof of Proposition 1 15

E.2 Proof of Theorem 1 15

E.3 Gaussian Assumption 18

F Numerical Stability with ϵ **19**

G Variation: Reduced Dimension d **20**

H Variation: Perturbation Sample Size n **20**

I Variation: Embedding Models **21**

J Variation: Response Generation Model **21**

K Hallucination Detection Pipeline using both External and Internal Uncertainty **22**

L Prompts **24**

L.1 Prompt template to extend queries. 24

L.2 Prompt template for query ambiguity detection 24

M Case study **24**

M.1 External Uncertainty (Query Ambiguity) 25

M.2 Internal Uncertainty (Response Uncertainty) 28

N Variation: Query Extension Model **29**

O More Discussion **29**

A More Related Work

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 explicitly trained to handle ambiguous queries and often generate incorrect responses instead, leading to hallucinations [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 et al., 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-based techniques [Wei et al., 2022, Guo et al., 2025, OpenAI, 2024], or simply turning to stronger LLMs or human agents.

Internal Uncertainty. Many studies propose uncertainty measures for hallucination detection [Kuhn et al., 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 et al., 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.

B Semantic-Volume Uncertainty Detection Algorithm

Here we provide the pseudocode of the Semantic-Volume uncertainty detection algorithm.

Algorithm 1 Uncertainty Detection via Semantic-Volume

Require:

- Dataset $\mathcal{D} = \{s_i\}_{i=1}^N$ of queries or query-response pairs.
- A small labeled subset $\mathcal{L} \subseteq \mathcal{D}$ with binary labels.
- SentenceTransformer model $\mathcal{M}$.

for each string $s \in \mathcal{D}$ **do**

  Extend s via query extension/response sampling to obtain: $[s_1, s_2, \ldots, s_N]$

  Normalize embedding vectors: $\boldsymbol{V} = [\boldsymbol{v}_1 \boldsymbol{v}_2 \ldots \boldsymbol{v}_N] \stackrel{\text{def}}{=} \left[\frac{\mathcal{M}(s_1)}{\|\mathcal{M}(s_1)\|}, \frac{\mathcal{M}(s_2)}{\|\mathcal{M}(s_2)\|}, \ldots, \frac{\mathcal{M}(s_N)}{\|\mathcal{M}(s_N)\|}\right]$

  Apply PCA dimension reduction: $\tilde{\boldsymbol{V}} \stackrel{\text{def}}{=} \boldsymbol{P}_{PCA}\boldsymbol{V}$.

  Compute Semantic Volume:

$$m(s) \leftarrow \log \mathbf{Vol}(\boldsymbol{P}_{PCA}\boldsymbol{V}) \stackrel{\text{def}}{=} \log\det\left(\tilde{\boldsymbol{V}}^\top \tilde{\boldsymbol{V}} + \epsilon\, \boldsymbol{I}_n\right)$$

end for

Threshold Tuning: Using the labeled set $\mathcal{L}$, find the threshold τ^* that maximizes the F1 score:

$$\tau^* \leftarrow \arg\max_{\tau\in\mathbb{R}} F_1(\tau) \stackrel{\text{def}}{=} \arg\max_{\tau\in\mathbb{R}} \left(\frac{2\sum_{i=1}^{|\mathcal{L}|} \mathbf{1}_{\hat{y}_i(\tau)=y_i=1}}{\sum_{i=1}^{|\mathcal{L}|} \mathbf{1}_{\hat{y}_i(\tau)=1} + \mathbf{1}_{y_i=1}}\right).$$

for each $(s, m(s))$ **do**

  Predict s with:

$$\hat{y}_s \leftarrow \begin{cases} 1, & \text{if } m(s) > \tau^*, \\ 0, & \text{otherwise.} \end{cases}$$

end for

return The fully labeled dataset: $\{(s, \hat{y}_s) \,:\, s \in \mathcal{D}\}$.

Time Complexity

Most of the computational cost in our algorithm comes from perturbation sampling, similar to other sampling-based methods. Using the standard LLM implementation to generate multiple candidate

sequences, the cost is equivalent to one LLM inference per data point. Embedding generation is highly efficient, taking approximately 5 minutes for our CLAMBER-3K dataset and 3 minutes for the TriviaQA-5K dataset on a single NVIDIA H100-80GB GPU. The computation of Semantic Volume, which involves matrix multiplication and determinant calculation for each set of perturbations, is computationally negligible (typically within seconds for the entire dataset), compared to the sampling step. This results in an overall complexity of $O(N)$ for our algorithm.

C Geometric interpretation as volume

We illustrate why $\det(\boldsymbol{V}^\top \boldsymbol{V})$ represents the squared volume using $n = 2$ and $n = 3$ as examples (see Figure 3). For $n = 2$, where $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2]$, the volume of the parallelepiped (equivalently, the area of the parallelogram) is simply $\sin(\theta)$, where θ is the angle between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. In this case,

$$\boldsymbol{V}^\top \boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1^\top \boldsymbol{v}_1 & \boldsymbol{v}_1^\top \boldsymbol{v}_2 \\ \boldsymbol{v}_2^\top \boldsymbol{v}_1 & \boldsymbol{v}_2^\top \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} 1 & \cos\theta \\ \cos\theta & 1 \end{bmatrix}.$$

Therefore, $\sqrt{\det(\boldsymbol{V}^\top \boldsymbol{V})} = \sqrt{1 - \cos^2\theta} = \sin\theta$. Similarly, for $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \boldsymbol{v}_3]$, the $\sqrt{\det(\boldsymbol{V}^\top \boldsymbol{V})}$ exactly computes $|\boldsymbol{v}_1^\top (\boldsymbol{v}_2 \times \boldsymbol{v}_3)|$, which is the volume of the parallelepiped.

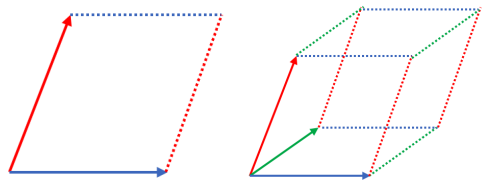


Figure 3: Parallelogram for $n = 2$ (left) and parallelepiped for $n = 3$ (right)

Beyond its geometric interpretation, $\det(\boldsymbol{V}^\top \boldsymbol{V})$ also quantifies the orthogonality of the vectors and is closely related to the determinantal point process (DPP) Kulesza et al. [2012]. DPPs favor diverse or orthogonal subsets by assigning higher probabilities to sets with dissimilar elements, using determinants to model repulsion among points. The Gram matrix $\boldsymbol{V}^\top \boldsymbol{V}$ is commonly used as the kernel matrix in DPPs, where the determinant of its submatrices determines subset probabilities. For a more detailed discussion, see Li et al. [2024] and Kulesza et al. [2012]. Moreover, some prior work has considered a similar volume measure to quantify vector diversity and orthogonality in LLM topics, such as the second orthogonality measure in Li et al. [2024], which motivates our design of semantic volume.

D Theoretical Analysis

In this section, we present theoretical analyses, with proofs and additional supporting lemmas provided in Appendix E. 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 et al., 2023]. Notably, semantic entropy involves a manual clustering step, and only considers the entropy between clusters while ignoring discrepancies 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

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

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 $\boldsymbol{X} \stackrel{\text{def}}{=} [\boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ and the normalized perturbation embeddings as $\boldsymbol{V} \stackrel{\text{def}}{=} [\boldsymbol{v}_1\boldsymbol{v}_2\ldots\boldsymbol{v}_n] \in \mathbb{R}^{d\times n}$ (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 $\mathcal{H}(\boldsymbol{X}) \stackrel{\text{def}}{=} -\mathbb{E}_{x\sim\boldsymbol{X}}\left[\log p_{\boldsymbol{X}}(x)\right]$ 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 E.

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*

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

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

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

E Proofs

E.1 Proof of Proposition 1

Proof. The formula equation 4 directly follows from the definition of F_1 scores. Recall we have

$$\text{Precision}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau)+\text{FP}(\tau)}, \quad \text{Recall}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau)+\text{FN}(\tau)},$$

where $\text{TP}(\tau) = \sum_{i=1}^{M} \mathbf{1}_{\hat{y}_i(\tau)=1\wedge y_i=1}$, $\text{FP}(\tau) = \sum_{i=1}^{M} \mathbf{1}_{\hat{y}_i(\tau)=1\wedge y_i=0}$, and $\text{FN}(\tau) = \sum_{i=1}^{M} \mathbf{1}_{\hat{y}_i(\tau)=0\wedge y_i=1}$. F_1 score is the harmonic mean of Precision and Recall:

$$\begin{aligned}
F_1(\tau) &= 2\cdot\frac{\text{Precision}(\tau)\times\text{Recall}(\tau)}{\text{Precision}(\tau)+\text{Recall}(\tau)} \\
&= \frac{2\cdot\text{TP}(\tau)}{2\cdot\text{TP}(\tau)+\text{FP}(\tau)+\text{FN}(\tau)} \\
&= \frac{2\sum_{i=1}^{M}\mathbf{1}_{\hat{y}_i(\tau)=y_i=1}}{2\sum_{i=1}^{M}\mathbf{1}_{\hat{y}_i(\tau)=1\wedge y_i=1} + \sum_{i=1}^{M}\mathbf{1}_{\hat{y}_i(\tau)=1\wedge y_i=0} + \sum_{i=1}^{M}\mathbf{1}_{\hat{y}_i(\tau)=0\wedge y_i=1}} \\
&= \frac{2\sum_{i=1}^{M}\mathbf{1}_{\hat{y}_i(\tau)=y_i=1}}{\sum_{i=1}^{M}(\mathbf{1}_{\hat{y}_i(\tau)=1} + \mathbf{1}_{y_i=1})}.
\end{aligned}$$

□

E.2 Proof of Theorem 1

First, in Lemma 1, we formally justify that our semantic volume measure for uncertainty detection is invariant under linear transformations applied to the uncertainty measure. This invariance follows directly from the step of searching for optimal threshold τ^* in our method.

Lemma 1 (Invariance of semantic volume method under linear transformation)**.** *Our method is invariant under linear transformation on the semantic volume measure $m(\cdot)$. More precisely, denote τ^* as the decision boundary defined in equation 4, we label*

$$y_i \overset{\text{def}}{=} \mathbf{1}_{m(s_i)<\tau^*}.$$

Given any linear transformation $T(m_i) \overset{\text{def}}{=} \alpha m_i + \beta$ ($\alpha > 0$) applied to the measure $m(\cdot)$, denote the new labels of our method under the new measure $T(m(\cdot))$ as

$$\tilde{y}_i \overset{\text{def}}{=} \mathbf{1}_{T(m(s_i))<\tilde{\tau}^*}.$$

Then we have

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

Proof. If we transform m_i to $\tilde{m}_i = T(m_i)$, the new threshold $\tilde{\tau}^*$ that maximizes the same metric still yields the same partition of the labeled samples. In other words,

$$m_i > \tau^* \iff T(m_i) = \tilde{m}_i > T(\tau^*),$$

a result of the simple fact that $m_i > \tau^* \iff \alpha(m_i) + \beta > \alpha(\tau^*) + \beta$. Thus our new threshold $\tilde{\tau}^*$ is effectively $T(\tau^*)$. Note that here if $\alpha < 0$ in $T(m_i) \overset{\text{def}}{=} \alpha m_i + \beta$, then naturally we should flip the decisions of our method, and then the labels still remain the same. □

Lemma 2 below provides the formula for the differential entropy of Gaussian vectors. This is a known result but we include this lemma and provide proof here for completeness.

Lemma 2 (Differential entropy of Gaussian vectors)**.** *For $X \in \mathbb{R}^d$ that follows a multivariate normal distribution*

$$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

The differential entropy of X is given by

$$\mathcal{H}(\boldsymbol{x}) = \frac{1}{2}(\log\det(\boldsymbol{\Sigma}) + d\log(2\pi) + d).$$

Proof. By the definition of the differential entropy and the probability density function of the multivariate normal distribution, we have

$$\begin{aligned}
\mathcal{H}(X) &= -\int_X p(x)\log p(x)dx = -\mathbb{E}_{x\sim X}\left[\log p(x)\right] \\
&= -\mathbb{E}\log\left[\frac{1}{\sqrt{(2\pi)^d\det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right] \\
&= \mathbb{E}\log\sqrt{(2\pi)^d\det(\boldsymbol{\Sigma})} + \mathbb{E}\left[\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \frac{1}{2}\mathbb{E}\left[d\log(2\pi) + \log\det(\boldsymbol{\Sigma})\right] + \frac{1}{2}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \frac{1}{2}\left(d\log(2\pi) + \log\det(\boldsymbol{\Sigma}) + \mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right)
\end{aligned}$$

By the cyclic property of trace, we have

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] &= \mathbb{E}\,\mathrm{Tr}\left[(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \mathbb{E}\,\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right] \\
&= \mathrm{Tr}\,\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right] \\
&= \mathrm{Tr}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} \\
&= d.
\end{aligned}$$

Combine these steps, we get

$$\mathcal{H}(X) = \frac{1}{2}(\log\det(\boldsymbol{\Sigma}) + d\log(2\pi) + d).$$

□

Lemma 3 (Eigenvalue distribution of rank-one perturbed matrix)**.** *For non-singular matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ and any rank-one perturbation $\boldsymbol{S} \stackrel{def}{=} \boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^\top$, we have*

$$\log\det(\boldsymbol{S}) = \log\det(\boldsymbol{M}) + \log(1 + \boldsymbol{v}^\top \boldsymbol{M}^{-1}\boldsymbol{u}), \tag{5}$$

Proof. By the matrix determinant lemma Harville [1998], we have

$$\begin{aligned} &\det(\boldsymbol{S}) = \det(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^\top) = \det(\boldsymbol{M})(1 + \boldsymbol{v}^\top \boldsymbol{M}^{-1}\boldsymbol{u}) \\ \Longrightarrow &\log\det(\boldsymbol{S}) = \log\det(\boldsymbol{M}) + \log(1 + \boldsymbol{v}^\top \boldsymbol{M}^{-1}\boldsymbol{u}). \end{aligned}$$

□

Next, we combine the results above to derive the proof of Theorem 1.

Proof of Theorem 1. For each s (either a query or response) with embedding $\bar{\boldsymbol{x}} \in \mathbb{R}^d$, we extend it to get n perturbations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. Equivalently, $\{\boldsymbol{v}_i\}_{i\in[n]}$ are perturbations of $\bar{\boldsymbol{v}}$. Assume the perturbation is Gaussian with covariance matrix $\boldsymbol{\Sigma}$. Then $\{\boldsymbol{x}_i\}_{i\in[n]}$ are samples for a Gaussian vector $X \sim \mathcal{N}(\bar{\boldsymbol{x}}, \boldsymbol{\Sigma})$. Denote $\lambda_i(\boldsymbol{A})$ as the i-th largest eigenvalue of matrix $\boldsymbol{A}$. Note that $\boldsymbol{V}^\top\boldsymbol{V}$ and $\boldsymbol{V}\boldsymbol{V}^\top$ are both positive semi-definite matrices and share the same nonzero eigenvalues (can be easily proved using SVD of $\boldsymbol{V}$). Similarly for $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^\top$. We can diagonalize $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = Q\Lambda Q^\top$ where Q is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Define an isotropic Gaussian vector $Z \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Then we can write $X = Q\Lambda^{1/2}Z$. Hence we have

$$\begin{aligned} &\mathbb{E}\|X\|^2 = \mathbb{E}[X^\top X] = \mathbb{E}[Z^\top \Lambda^{1/2} Q^\top Q \Lambda^{1/2} Z] \\ &= \mathbb{E}[Z^\top \Lambda Z] = \mathbb{E}\left[\sum_{i=1}^d \lambda_i Z_i^2\right] = \sum_{i=1}^d \lambda_i \mathbb{E}[Z_i^2] = \mathrm{Tr}(\boldsymbol{\Sigma}). \end{aligned}$$

Note that in high-dimensions, $\|X\|^2$ is concentrated around its mean $\mathbb{E}\|X\|^2 = \mathrm{Tr}(\boldsymbol{\Sigma})$ (particularly for isotropic Gaussian, this is $\sqrt{D}$). Under this regime, we have $\boldsymbol{V} \stackrel{\text{def}}{=} [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n] \approx \frac{1}{\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma})}}\boldsymbol{X}$. Let $\alpha = n/\mathrm{Tr}(\boldsymbol{\Sigma})$. Then

$$\begin{aligned} &\log\det(\boldsymbol{V}^\top\boldsymbol{V}) \approx \log\det\left(\frac{1}{\mathrm{Tr}(\boldsymbol{\Sigma})}\boldsymbol{X}^\top\boldsymbol{X}\right) \\ &= \log\det\left(\alpha \cdot \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) \\ &= \log\det\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) + n\log\alpha. \end{aligned}$$

Since our method sets a threshold τ^* on the log-determinant value, searched using a held-out set $\mathcal{L} \in \mathcal{D}$, the final classification is invariant to shifts (and more generally, to any linear transformation. See Proposition 1). Hence we can ignore the constant term $n\log\alpha$ and only focus on $\log\det\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)$. We have

$$\begin{aligned} &\log\det\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) = \log\prod_{i=1}^n \lambda_i\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) \\ &= \sum_{i=1}^n \log\lambda_i\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) = \sum_{i=1}^n \log\lambda_i\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top\right). \end{aligned}$$

Thus, it suffices to study the eigenvalues of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$. Note that we have

$$\begin{aligned}
\boldsymbol{X}\boldsymbol{X}^\top - n\boldsymbol{\Sigma} &= \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top\right) - \left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top\right) \\
&= \sum_{i=1}^{n}(\bar{\boldsymbol{x}}\boldsymbol{x}_i^\top - \boldsymbol{x}_i\bar{\boldsymbol{x}}^\top + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top) \\
&= \bar{\boldsymbol{x}}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\right)^\top - \left(\sum_{i=1}^{n}\boldsymbol{x}_i\right)\bar{\boldsymbol{x}}^\top + n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top) \\
&\approx n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top + n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top \\
&= n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top.
\end{aligned}$$

This implies that

$$\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma} \approx \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top. \tag{6}$$

Thus $\boldsymbol{\Sigma}$ is essentially a rank-one perturbation of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$. By Lemma 2, we know the differential entropy of X is:

$$\mathcal{H}(X) = \frac{1}{2}(\log\det(\boldsymbol{\Sigma}) + d\log(2\pi) + d).$$

Again due to the invariance of our method to linear transformation, we can only focus on $\log\det(\boldsymbol{\Sigma})$. When d is large, by Lemma 3, let $\boldsymbol{M} \stackrel{\text{def}}{=} \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$, we have $\log\det(\boldsymbol{\Sigma}) = \log\det\boldsymbol{M} + \log(1 + \bar{\boldsymbol{x}}^\top\boldsymbol{M}^{-1}\bar{\boldsymbol{x}}^\top)$. Note here $\bar{\boldsymbol{x}}$ is constant, and recall that $\boldsymbol{M}$ concentrate around a deterministic matrix $\boldsymbol{\Sigma} + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top$ from equation 6 above (one can show $\|\boldsymbol{M} - (\boldsymbol{\Sigma} + \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top)\| \xrightarrow{p} 0$ as $n \to \infty$, using standard random matrix theory tools). Therefore combining all above, we have derived that $\log\det\boldsymbol{M}$ is essentially shifted $\log\det(\boldsymbol{\Sigma})$ in high-dimensional regimes, and hence same for $\log\det\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$. This completes the proof. □

E.3 Gaussian Assumption

Our theoretical analysis assumes a Gaussian distribution for the embedding vectors, particularly after PCA projection. This Gaussian assumption facilitates the interpretation of our proposed semantic volume measure in terms of differential entropy. Here, we provide empirical evidence to support the relative Gaussianity of these projected embedding vectors.

To empirically test Gaussianity, we leverage the fact that if samples follow a multivariate normal distribution with mean and covariance matrix, then the *squared Mahalanobis distances*:

$$D_i^2 = (x_i - \mu)^T\Sigma^{-1}(x_i - \mu)$$

should follow a *chi-square distribution* with d degrees of freedom:

$$D_i^2 \sim \chi_d^2$$

Thus we compute D_i^2 for each sample, using the empirical mean and covariance matrix of the perturbation vectors, and compare their distribution to the χ_d^2 distribution. We analyze the quantiles of Mahalanobis distances and the theoretical χ_d^2 distribution. Better alignment with the line $y = x$ in these plots indicates stronger evidence of Gaussianity. Example Q–Q plots are shown in Figure 4, where the empirical distributions generally align well with the theoretical expectations.

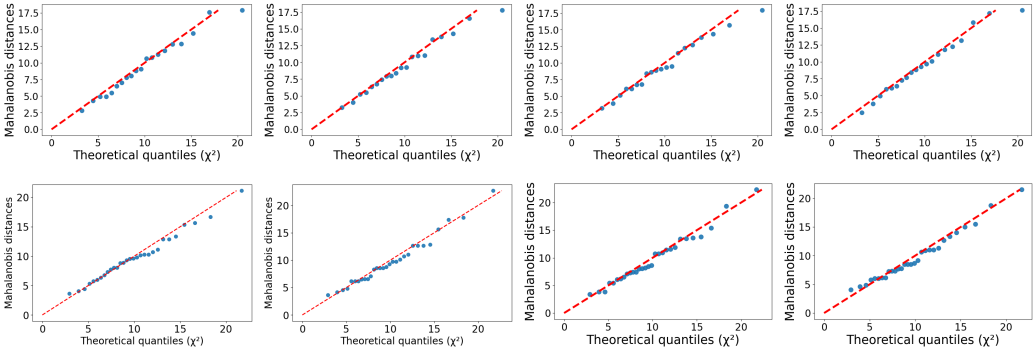


Figure 4: Mahalanobis–chi-square Q–Q plots for four examples each of queries (top row) and responses (bottom row).

To quantitatively evaluate Gaussianity, we compute the R^2 score from a linear regression between the empirical quantiles of squared Mahalanobis distances and the theoretical χ^2_d quantiles:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i are the observed values, $\hat{y}_i$ the theoretical values, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ the mean of the observations. Note that $R^2 \in (-\infty, 1]$ and a larger R^2 value indicates better alignment and thus stronger Gaussianity. The histogram of R^2 values is plotted in Figure 5. If we use 0.8 as a threshold for high Gaussianity, 90.5% of the query embeddings (in CLAMBER) and 71.4% of the response embeddings (in TriviaQA) exceed this threshold, suggesting strong empirical support for our Gaussian assumption.

Although these distributions are not perfectly Gaussian in practice, our theoretical results primarily serve as interpretative tools to understand how semantic embedding measures semantic dispersion. We leave the further theoretical exploration on this to future work.

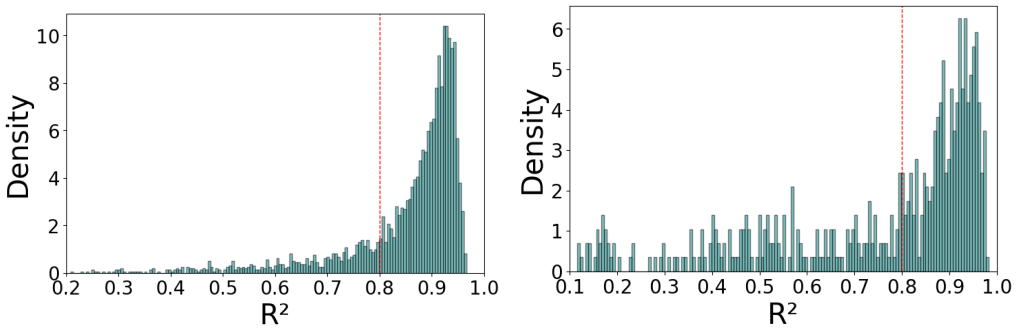


Figure 5: Distribution of R^2.

F Numerical Stability with ϵ

Here, we present the distribution of the matrix norm $\|\boldsymbol{V}^\top \boldsymbol{V}\|$ (spectral norm) across our datasets in Figure 6. The histograms indicate that the chosen value of $\epsilon = 10^{-10}$ is negligible compared to the typical magnitude of $\|\boldsymbol{V}^\top \boldsymbol{V}\|$. This confirms that ϵ only serves to ensure numerical stability rather than influencing the quantification. Moreover, in practice, we observe that exact repetition among sampled perturbations is rare.

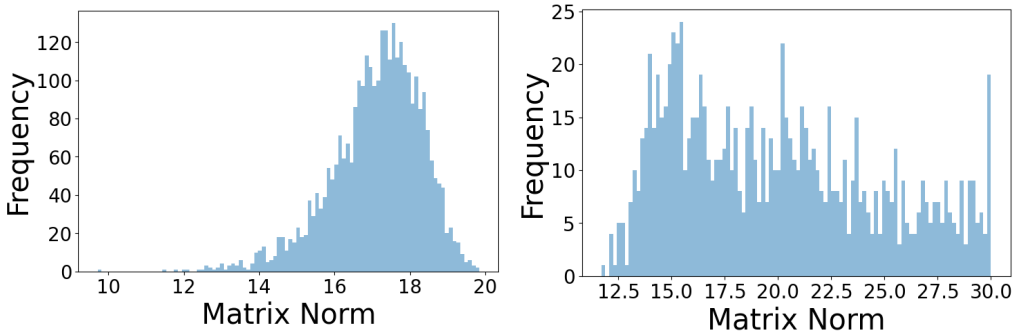


Figure 6: Distribution of the spectral norm $\|\boldsymbol{V}^\top \boldsymbol{V}\|$ for the perturbation embedding vectors (the left histogram corresponds to query perturbations and the right is for response perturbations.

G Variation: Reduced Dimension d

In Figure 7, we illustrate the impact of varying d on performance. Additionally, we compare our method with the Semantic Volume that uses the original embedding vectors instead of the projected ones. Our results show that projecting to a lower-dimensional space improves performance, possibly because PCA projections better separate perturbation vectors. Furthermore, our study suggests that there may be an optimal, task-dependent choice of d. Exploring this further and characterizing the optimal d is left for future work.

Moreover, alternative dimensionality reduction techniques, such as autoencoders, may also be employed. Investigating the impact of different dimensionality reduction methods remains an important direction for future research.

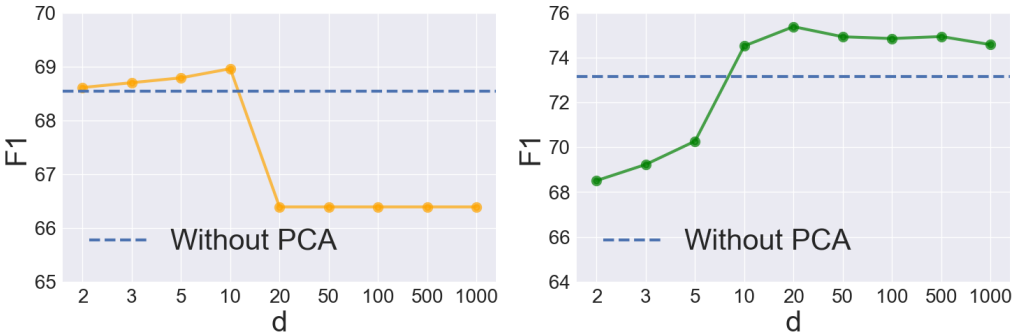
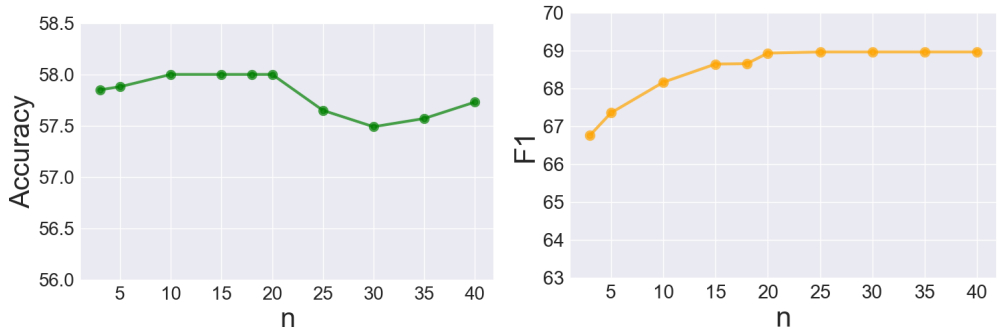


Figure 7: F1 scores for different dimension d in PCA dimension reduction. The dashed line represents the F1 score of the semantic volume method without dimension reduction. The left figure corresponds to the external uncertainty task, while the right figure corresponds to the internal uncertainty task.

H Variation: Perturbation Sample Size n

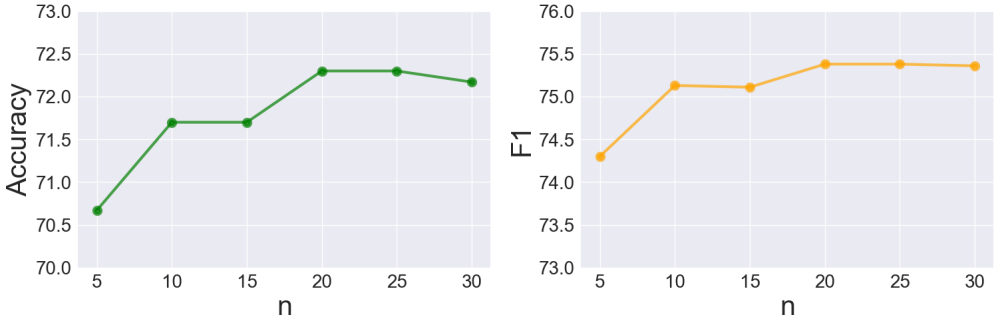
The perturbation sample size n is a hyperparameter in our method. A larger n provides a more comprehensive estimate of semantic dispersion but increases computational cost, while a smaller n may fail to capture sufficient variation, reducing the reliability of uncertainty quantification.

Figure 8 illustrates how varying n affects classification performance. In general, increasing n tends to improve F1 scores, likely due to a more accurate estimation of dispersion. However, beyond a certain point, the improvements diminish while the cost continues to rise. To balance efficiency and performance, we set $n = 20$ for external uncertainty detection and internal uncertainty detection.



(a) External uncertainty: best accuracy for different n.

(b) External uncertainty: best F1-score for different n.



(c) Internal uncertainty: best accuracy for different n.

(d) Internal uncertainty: best F1-score for different n.

Figure 8

I Variation: Embedding Models

To examine the effect of different embedding models on our method, we explore two larger sentence-transformers: `Alibaba-NLP/gte-Qwen2-7B-instruct` (embedding dimension 3584) from the same Qwen family as used in our main results, and `nvidia/NV-Embed-v2` (embedding dimension 4096) from a different model family. Using the external uncertainty task as a case study, we present the results in Table 3. Our findings indicate that models from different families but with similar sizes produce comparable results. Compared to the Qwen-1.5B sentence transformer used in the main results, performance remains largely similar—while the 7B model achieves slightly higher accuracy but a lower F1 score, this difference may not be statistically significant.

	Qwen2-7B		**NV-Embed**	
	Accuracy	**F1**	**Accuracy**	**F1**
$n = 25$	57.8	68.0	57.4	67.5
$n = 20$	58.1	68.0	58.4	67.9
$n = 15$	57.7	68.1	57.9	67.8
$n = 10$	57.2	67.7	57.4	67.5
$n = 5$	56.8	66.9	57.7	67.0

Table 3: Performance with different embedding models.

J Variation: Response Generation Model

To assess the generalizability of our method, we conduct the experiment for internal uncertainty detection for these response models:

Method	TriviaQA			SQuAD		
	Acc.	F1	AUROC	Acc.	F1	AUROC
Prompt Llama3.2-1B-Instruct	50.1	66.5	N/A	50.8	66.6	N/A
Prompt Llama3-8B-Instruct	65.5	64.0	N/A	58.3	51.6	N/A
Prompt Mistral-7B-Instruct	60.4	54.3	N/A	50.1	0.6	N/A
LastTokenEntropy	56.2	55.9	63.9	51.7	35.8	53.0
LogProb	53.4	51.1	54.4	57.1	55.8	61.0
pTrue Llama3.2-1B-Instruct	$50.0_{0.0}$	$66.6_{0.0}$	$55.1_{0.8}$	$50.2_{0.6}$	$54.9_{0.6}$	$51.1_{0.2}$
pTrue Llama3-8B-Instruct	$55.2_{0.2}$	$23.2_{7.4}$	$52.8_{1.0}$	$50.7_{1.3}$	$57.7_{1.2}$	$50.6_{0.5}$
pTrue Mistral-7B-Instruct	$50.8_{1.2}$	$4.3_{3.2}$	$72.7_{1.6}$	$51.4_{0.4}$	$58.5_{0.2}$	$48.9_{0.3}$
LexicalSimilarity	$64.2_{0.2}$	$72.0_{0.4}$	$78.7_{0.5}$	$67.0_{0.3}$	$64.0_{0.9}$	$70.4_{0.2}$
SemanticEntropy	$63.5_{0.5}$	$69.4_{1.8}$	$73.2_{0.4}$	$62.5_{0.1}$	$65.2_{1.3}$	$66.1_{0.3}$
Semantic Volume (ours)	$\mathbf{72.0}_{0.1}$	$\mathbf{74.9}_{0.0}$	$\mathbf{79.7}_{0.2}$	$\mathbf{68.0}_{0.2}$	$\mathbf{71.0}_{0.1}$	$\mathbf{71.0}_{0.9}$

Table 4: Response model: `Llama3-8B-Instruct`. 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 $\pm$ std. over three trials (std. shown as subscript).

Method	TriviaQA			SQuAD		
	Acc.	F1	AUROC	Acc.	F1	AUROC
Prompt Llama3.2-1B-Instruct	50.6	66.7	N/A	50.6	66.7	N/A
Prompt Llama3-8B-Instruct	65.2	62.5	N/A	64.3	62.8	N/A
Prompt Mistral-7B-Instruct	51.2	5.6	N/A	50.8	3.7	N/A
LastTokenEntropy	48.3	29.9	49.3	48.8	13.0	50.2
LogProb	55.5	52.7	58.4	54.2	39.8	58.1
pTrue Llama3.2-1B-Instruct	$53.3_{0.7}$	$52.0_{0.9}$	$50.7_{0.7}$	$51.8_{0.6}$	$49.9_{0.8}$	$50.5_{0.9}$
pTrue Llama3-8B-Instruct	$52.0_{0.1}$	$63.0_{0.1}$	$49.9_{0.3}$	$52.1_{0.2}$	$62.9_{0.3}$	$50.3_{1.1}$
pTrue Mistral-7B-Instruct	$52.8_{0.7}$	$54.0_{0.9}$	$50.5_{0.1}$	$53.6_{0.1}$	$55.4_{0.2}$	$51.8_{0.6}$
LexicalSimilarity	$64.3_{0.4}$	$62.4_{0.7}$	$66.0_{0.1}$	$65.6_{0.2}$	$61.0_{1.8}$	$65.7_{0.2}$
SemanticEntropy	$59.5_{0.7}$	$51.2_{2.2}$	$63.7_{0.0}$	$58.5_{0.6}$	$55.3_{1.2}$	$62.3_{0.5}$
Semantic Volume (ours)	$\mathbf{65.4}_{0.2}$	$\mathbf{68.6}_{0.0}$	$\mathbf{67.6}_{0.8}$	$\mathbf{66.0}_{0.1}$	$\mathbf{69.2}_{0.2}$	$\mathbf{68.7}_{0.7}$

Table 5: Response model: `Mistral-7B-Instruct`. 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 $\pm$ std. over three trials (std. shown as subscript).

- `Llama3.2-1B-Instruct` (Table 2)
- `Llama3-8B-Instruct` (Table 4)
- `Qwen2.5-1.5B-Instruct` (Table 6)
- `Qwen3-8B` (Table 7)
- `Qwen3-14B` (Table 8).
- `Mistral-7B-Instruct` (Table 5).

Notably, our methods consistently outperforms the baselines for all the response model we tested. It is also worth noting that, with a larger model such as `Llama3-8B-Instruct`, the probability-based methods show a lot higher latency. In our experiment, both Last Token Entropy and Log Probabilities took more than 4 hours on a H100-80GB GPU for each response model.

Furthermore, similar to Table 2, we observe that methods relying on prompting the LLM (including pTrue) sometimes exhibit a strong bias toward predicting almost exclusively "Yes" or "No". In Table 4, for example, *Prompt Llama3.2-1B-Instruct* and *pTrue (Llama3.2-1B-Instruct)* predict nearly all samples as "Yes", whereas *pTrue (Mistral-7B-Instruct)* predicts almost all as "No". This instability highlights another drawback of using prompted LLMs for uncertainty prediction.

K Hallucination Detection Pipeline using both External and Internal Uncertainty

In this paper, we conducted separate experiments on external uncertainty detection and internal uncertainty detection to demonstrate that the Semantic Volume is an effective method that can be

Method	TriviaQA			SQuAD		
	Acc.	F1	AUROC	Acc.	F1	AUROC
Prompt Llama3.2-1B-Instruct	50.9	67.0	N/A	51.1	66.8	N/A
Prompt Llama3-8B-Instruct	77.0	76.6	N/A	62.3	59.7	N/A
Prompt Mistral-7B-Instruct	51.8	7.7	N/A	50.4	2.0	N/A
LastTokenEntropy	53.6	33.6	53.8	51.0	15.9	52.4
LogProb	57.6	56.9	62.2	54.1	43.9	57.1
pTrue Llama3.2-1B-Instruct	$55.1_{0.2}$	$58.0_{0.1}$	$47.9_{0.9}$	$50.7_{1.0}$	$53.9_{1.1}$	$50.3_{0.8}$
pTrue Llama3-8B-Instruct	$53.8_{0.5}$	$59.7_{0.6}$	$52.4_{0.2}$	$51.9_{0.4}$	$58.5_{0.4}$	$51.9_{0.5}$
pTrue Mistral-7B-Instruct	$53.6_{0.3}$	$59.2_{0.3}$	$47.1_{0.7}$	$52.1_{0.7}$	$59.2_{0.5}$	$50.8_{0.7}$
LexicalSimilarity	$74.6_{0.2}$	$73.9_{0.6}$	$72.8_{0.1}$	$63.5_{0.1}$	$68.7_{0.6}$	$67.5_{0.1}$
SemanticEntropy	$74.0_{0.5}$	$73.9_{1.3}$	$75.5_{0.1}$	$63.8_{0.4}$	$68.9_{0.6}$	$67.2_{0.2}$
Semantic Volume (ours)	$\mathbf{75.2}_{0.1}$	$\mathbf{77.7}_{0.1}$	$\mathbf{75.6}_{0.8}$	$\mathbf{65.7}_{0.1}$	$\mathbf{71.4}_{0.1}$	$\mathbf{68.1}_{0.7}$

Table 6: Response model: `Qwen2.5-1.5B-Instruct`. 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 $\pm$ std. over three trials (std. shown as subscript).

Method	TriviaQA			SQuAD		
	Acc.	F1	AUROC	Acc.	F1	AUROC
Prompt Llama3.2-1B-Instruct	50.5	66.7	N/A	50.6	66.6	N/A
Prompt Llama3-8B-Instruct	71.9	70.2	N/A	61.1	56.3	N/A
Prompt Mistral-7B-Instruct	50.8	3.5	N/A	50.2	1.1	N/A
LastTokenEntropy	57.7	49.6	60.8	52.1	26.3	51.8
LogProb	56.2	38.0	62.3	55.9	48.8	59.4
pTrue Llama3.2-1B-Instruct	$51.5_{0.6}$	$56.2_{0.4}$	$50.4_{0.6}$	$49.3_{0.7}$	$52.6_{0.6}$	$49.9_{0.7}$
pTrue Llama3-8B-Instruct	$50.9_{0.2}$	$57.7_{0.2}$	$49.7_{0.7}$	$50.2_{0.8}$	$56.8_{0.8}$	$50.3_{0.4}$
pTrue Mistral-7B-Instruct	$52.7_{0.8}$	$57.5_{0.7}$	$51.5_{1.0}$	$51.6_{0.6}$	$57.2_{0.4}$	$50.3_{0.5}$
LexicalSimilarity	$73.7_{0.7}$	$71.6_{1.4}$	$67.8_{0.0}$	$66.2_{0.1}$	$63.6_{1.5}$	$71.9_{0.2}$
SemanticEntropy	$73.4_{0.6}$	$71.5_{1.6}$	$69.8_{0.2}$	$64.1_{0.4}$	$62.5_{0.9}$	$69.1_{0.1}$
Semantic Volume (ours)	$\mathbf{75.2}_{0.2}$	$\mathbf{77.3}_{0.1}$	$\mathbf{75.3}_{0.6}$	$\mathbf{66.7}_{0.4}$	$\mathbf{71.1}_{0.1}$	$66.7_{0.7}$

Table 7: Response model: `Qwen3-8B`. 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 $\pm$ std. over three trials (std. shown as subscript).

Method	TriviaQA			SQuAD		
	Acc.	F1	AUROC	Acc.	F1	AUROC
Prompt Llama3.2-1B-Instruct	50.5	66.8	N/A	50.6	66.5	N/A
Prompt Llama3-8B-Instruct	68.6	65.9	N/A	59.5	53.5	N/A
Prompt Mistral-7B-Instruct	51.0	4.1	N/A	50.3	1.8	N/A
LastTokenEntropy	62.2	64.1	65.8	56.1	48.9	60.5
LogProb	58.2	49.8	62.5	58.0	62.8	60.3
pTrue Llama3.2-1B-Instruct	$52.5_{0.3}$	$56.1_{0.4}$	$51.1_{0.9}$	$48.9_{0.4}$	$51.4_{0.4}$	$51.2_{0.2}$
pTrue Llama3-8B-Instruct	$51.2_{0.8}$	$56.9_{0.8}$	$49.8_{0.5}$	$50.2_{0.6}$	$56.3_{0.6}$	$49.8_{0.3}$
pTrue Mistral-7B-Instruct	$54.3_{0.3}$	$58.2_{0.4}$	$53.2_{0.5}$	$51.3_{0.3}$	$56.7_{0.3}$	$51.6_{0.2}$
LexicalSimilarity	$75.0_{0.1}$	$76.0_{0.1}$	$74.7_{0.1}$	$64.3_{0.2}$	$58.3_{0.6}$	$64.5_{0.2}$
SemanticEntropy	$74.4_{0.2}$	$74.6_{0.7}$	$74.2_{0.1}$	$63.1_{0.5}$	$57.6_{1.5}$	$65.5_{0.2}$
Semantic Volume (ours)	$\mathbf{75.1}_{0.5}$	$\mathbf{76.9}_{0.6}$	$\mathbf{75.1}_{0.5}$	$\mathbf{66.1}_{0.6}$	$\mathbf{69.0}_{0.5}$	$\mathbf{66.0}_{0.6}$

Table 8: Response model: `Qwen3-14B`. 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 $\pm$ std. over three trials (std. shown as subscript).

generally applied to both tasks. In practice, these two tasks can be combined into a unified pipeline for hallucination detection (see Figure 9): First, we perform the internal uncertainty detection. If high uncertainty is detected in the response, then hallucination is likely to happen. Then in the second step, we check the external uncertainty: If the query is detected to be ambiguous, the LLM should ask a clarification question to the user. After clarification is provided or additional information is provided to resolve the ambiguity in the query, the LLM can then generate the answer to the query.

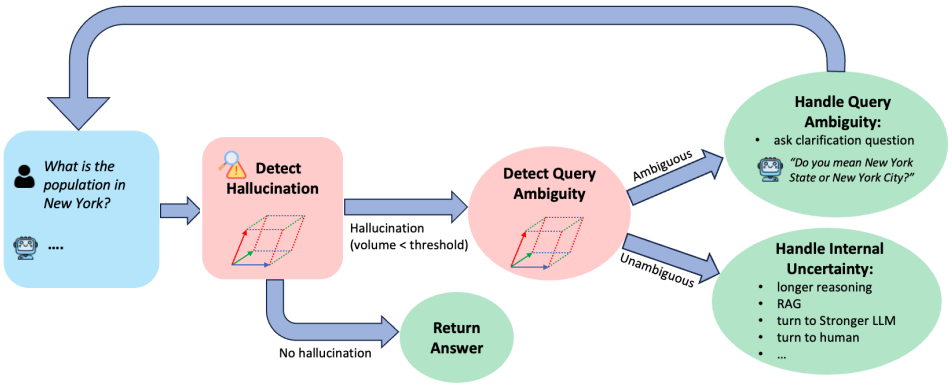


Figure 9: Complete pipeline for hallucination detection utilizing both external and internal uncertainty.

On the other hand, if external uncertainty is ruled out, then the hallucination is likely caused by the internal lack of knowledge of the LLM. This can be addressed through various methods such as retrieval-augmented generation (RAG), reasoning-based approaches, or by leveraging stronger LLMs or human agents.

Moreover, all of our experiments are set up as classification tasks for uncertainty detection, relying on a threshold-based decision. However, these uncertainty measures can also be applied to comparison or ranking tasks.

L Prompts

L.1 Prompt template to extend queries.

```
Provide a paraphrase of the following question with a contextual expansion,
while maintaining its core meaning.  The filled context information should
be diverse but must be concrete and specific (it cannot be a placeholder or
a template).  Only reply with the new version of the question and nothing
else.

Question:

{question}
```

L.2 Prompt template for query ambiguity detection

```
Is the following question ambiguous?  A question is ambiguous if it can
be interpreted in multiple ways or has multiple possible answers.  If the
question is ambiguous, then reply 'Yes', otherwise reply 'No'.  Only reply
with 'Yes' or 'No' and nothing else.

Question:

{question}
```

M Case study

To further illustrate the distinct levels of uncertainty in queries and responses, we present a detailed case study in this section, providing concrete examples and their corresponding perturbations analyzed using our Semantic Volume method. For the queries in CLAMBER, we analyze three unambiguous examples with Semantic Volumes in the bottom 5% percentile of the dataset, and three ambiguous examples with large Semantic Volumes (above the 95% percentile). For demonstration purposes, we present five extended versions (perturbations) for each sample. We observe that the ambiguous queries exhibit high semantic dispersion (as captured by their large Semantic Volumes), while the unambiguous ones have extensions that are quite similar.

Similarly, we present three correct samples and three hallucinated samples with the lowest and highest Semantic Volume values, respectively, from the TriviaQA dataset. For each sample, we provide five additional sampled candidates and show that, for correct answers, the candidates are semantically equivalent, whereas for hallucinated answers, the candidates exhibit substantial diversity.

M.1 External Uncertainty (Query Ambiguity)

For the unambiguous queries, we observe that all extensions are quite similar and preserve the same core meaning. For each unambiguous query, we include the original clarification question from CLAMBER and provide our case analysis following each example.

Examples of Unambiguous Queries

Example 1. Query: Who plays harry in the amazing spider man 2?

Perturbations (extended queries):

1. In the 2014 superhero film "The Amazing Spider-Man 2," directed by Marc Webb and featuring Andrew Garfield as Peter Parker/Spider-Man, which actor portrays Harry Osborn, the childhood friend of Peter who eventually becomes the Green Goblin?
2. In the 2014 superhero film "The Amazing Spider-Man 2," directed by Marc Webb and starring Andrew Garfield as Peter Parker/Spider-Man, which actor portrays Harry Osborn, the childhood friend of Peter who becomes the Green Goblin?
3. In the 2014 superhero film "The Amazing Spider-Man 2," directed by Marc Webb and starring Andrew Garfield as Peter Parker/Spider-Man, which actor portrays the character of Harry Osborn, Peter's childhood friend who becomes the Green Goblin?
4. In the 2014 superhero film "The Amazing Spider-Man 2," directed by Marc Webb and starring Andrew Garfield as Peter Parker, which actor portrays Harry Osborn, the childhood friend of Peter who eventually becomes the Green Goblin?
5. In the 2014 superhero film "The Amazing Spider-Man 2," directed by Marc Webb and starring Andrew Garfield as Peter Parker, which actor portrays Harry Osborn, the childhood friend of Peter who becomes the Green Goblin?

Example 2. Query: Who plays young william in a knight's tale?

Perturbations (extended queries):

1. In the 2001 medieval-themed sports drama "A Knight's Tale," starring Heath Ledger as the adult William Thatcher, which actor portrays the character of William as a child during the film's opening sequences set in 14th-century England?
2. In the 2001 medieval-themed sports drama "A Knight's Tale," starring Heath Ledger as the adult William Thatcher, which young actor portrays the childhood version of the protagonist during the film's opening scenes set in 14th-century England?
3. In the 2001 medieval-themed sports drama "A Knight's Tale," starring Heath Ledger as the adult William Thatcher, which actor portrays the character of William as a child during the film's opening scenes set in 14th-century England?
4. In the 2001 medieval-themed sports drama "A Knight's Tale," starring Heath Ledger as the adult William Thatcher, which actor portrays the childhood version of the protagonist during the film's opening sequences set in 14th-century England?
5. In the medieval-themed 2001 film "A Knight's Tale," starring Heath Ledger as the adult William Thatcher, which actor portrays the character of William as a child during the opening scenes set in 14th-century England?

Example 3. Query: When did david cassidy release i think i love you?

Perturbations (extended queries):

1. In what year did the teen idol and star of "The Partridge Family," David Cassidy, release his breakthrough hit single "I Think I Love You," which topped the Billboard Hot 100 chart and helped solidify his status as a 1970s pop music sensation?
2. In what year did the teen idol and former star of "The Partridge Family," David Cassidy, release his chart-topping hit single "I Think I Love You," which became an anthem for young romance in the early 1970s and helped solidify his status as a pop sensation?
3. In what year did the former teen idol and star of "The Partridge Family," David Cassidy, release his chart-topping hit single "I Think I Love You," which became an anthem for young romance in the early 1970s and helped solidify his status as a pop culture icon?
4. In which year did the teen heartthrob and former star of "The Partridge Family," David Cassidy, release his chart-topping hit single "I Think I Love You," which became an anthem for young romance in the early 1970s and helped solidify his status as a pop idol?
5. In which year did the teen idol and star of "The Partridge Family," David Cassidy, release his chart-topping hit single "I Think I Love You," which became an anthem for young romance in the early 1970s and helped solidify his status as a pop music sensation?

Examples of Ambiguous Queries

Example 4. Query: What is the theme of Togetherness?

Clarifying Question: Are you referring to the TV show or the concept of togetherness in general?

Perturbations (extended queries):

1. How does the concept of unity and shared experiences manifest as the central motif in the 2015 Danish TV series "Togetherness," created by Mark and Jay Duplass, which explores the complexities of marriage, friendship, and personal growth among a group of middle-aged characters in Los Angeles?
2. In what ways does the concept of unity and shared experience manifest itself as a central motif in "Togetherness," the HBO comedy-drama series created by Jay and Mark Duplass? How does the show explore the complexities of human relationships, particularly among its four main characters living in Los Angeles, as they navigate marriage, friendship, and personal growth?
3. How does the theme of togetherness manifest in human relationships, whether it's in the tight-knit bonds of family gatherings during Thanksgiving dinner, the camaraderie among teammates celebrating a championship victory, or the collective spirit of strangers coming together to help their community rebuild after a natural disaster?
4. How does the concept of unity and collective experience manifest as a central motif in the novel "Togetherness" by acclaimed author Sarah Chen, which explores the interconnected lives of residents in a small coastal town facing environmental challenges?
5. How does the central message of unity and human connection manifest throughout the album "Togetherness" by jazz trumpeter Roy Hargrove, particularly in tracks like "Trust" and "Strasbourg/St. Denis" which emphasize collective improvisation and musical dialogue between band members?

Analysis: The extended queries above illustrate distinct interpretations of the ambiguous term "Togetherness." The first two extensions interpret it as the HBO television series Togetherness (2015) by Mark and Jay Duplass. The third treats it as a general conceptual inquiry into human connection, unity, and solidarity. The fourth interprets "Togetherness" as the title of a novel, while the final extension refers to a music album–Togetherness by jazz trumpeter Roy Hargrove. These diverse extensions highlight the semantic ambiguity of the original query, as well as the high variability in meaning captured by the Semantic Volume method.

Example 5. Query: What do you call a ten sided shape?

Clarifying Question: Which one: 2d shape, or 3d shape?

Perturbations (extended queries):

1. What is the geometric term for a polygon with ten sides and ten angles, such as the shape found in the British 50 pence coin or the logo of the Detroit Red Wings hockey team?
2. In geometry class, while studying regular polygons, the teacher introduced various shapes. After discussing triangles, squares, pentagons, and octagons, she asked the students: "Moving on to more complex figures, what is the proper term for a polygon with exactly ten edges and ten vertices?"
3. In the field of geometry, particularly when studying polygons and their properties, what is the specific term used to describe a two-dimensional figure with exactly ten straight sides and ten vertices? This shape is commonly encountered in various applications, from architectural designs to board game dice, and understanding its proper nomenclature is essential for mathematical discussions and practical applications.
4. What is the mathematical term for a polygon with ten edges and ten vertices, like the shape of the US Pentagon building if it had five additional sides or the cross-section of some crystalline minerals found in nature?
5. What is the geometric term for a polygon with ten edges and ten vertices, like the profile of a traditional Chinese pagoda or the cross-section of certain crystals found in mineralogy?

Analysis: While all extensions refer to ten-sided geometric figures, their interpretations vary in subtle ways. The first three primarily describe 2D polygons, consistently referencing shapes with ten sides and vertices in mathematical or educational contexts. The last two, however, introduce 3D analogies, such as architectural structures or crystal cross-sections. This variation highlights the ambiguity in the original query.

Example 6. Query: What is the cure for Cured?

Clarifying Question: Are you referring to a medical condition or a song?

Perturbations (extended queries):

1. How can we effectively treat and reverse the effects of the fictional disease known as "Cured," which paradoxically causes illness despite its name, in the dystopian world of Lauren DeStefano's Chemical Garden trilogy?
2. What medical treatment or intervention has been discovered to effectively eliminate or reverse the effects of Cured syndrome, a rare genetic disorder characterized by progressive muscle weakness and cognitive decline, which was first identified in 2008 by Dr. Elena Cured at Johns Hopkins University?
3. After watching the 2020 psychological horror film "Cured" directed by Sean Ellis, where a woman discovers she has a mysterious condition that makes her immune to diseases but also causes violent outbursts, what medical treatment or intervention ultimately resolves her affliction?
4. How can we treat or reverse "Cured," the psychological horror game released in 2016 that follows a patient trapped in an abandoned mental asylum who must escape while uncovering dark secrets about illegal experiments and supernatural phenomena?
5. What is the medical treatment or remedy for the ironic condition known as Cured Syndrome - a rare autoimmune disorder first identified in Uppsala, Sweden in 1982 that causes patients to experience persistent sensations of perfect health despite having serious underlying symptoms?

Analysis: The term "Cured" is interpreted in five distinct ways: as a fictional disease, a genetic disorder, a psychological horror film, a video game, and an ironically named medical condition. These varied interpretations highlight the ambiguity of the original query and the broad semantic spread, which is captured by our Semantic Volume method.

M.2 Internal Uncertainty (Response Uncertainty)

In the cases where the responses exhibit low Semantic Volumes (the first three examples), the answers are correct and the sampled candidates are nearly identical. This consistency reflects the model's strong confidence and low uncertainty. In contrast, for cases with high Semantic Volumes (last three examples), the sampled candidates show substantial diversity.

Examples of Correct Responses

Example 1. Query: The Ebro River is in which country?
Response: Spain
Ground Truth Answer: Spain

Perturbations (candidate responses):

1. Spain
2. Spain
3. Spain
4. Spain
5. Spain

Example 2. Query: Gorgonzola cheese is from which country?
Response: Italy
Ground Truth Answer: Italy

Perturbations (candidate responses):

1. Italy
2. Italy
3. Italy
4. Italy
5. Italy

Example 3. Query: Tipperary is in which European country??
Response: Ireland
Ground Truth Answer: Ireland

Perturbations (candidate responses):

1. Ireland
2. Ireland
3. Ireland
4. Republic of Ireland
5. Ireland

Examples of Hallucinated Responses

Example 4. Query: In Top Cat, who was the voice of Choo Choo?
Response: John Kassir
Ground Truth Answer: Marvin Kaplan

Perturbations (candidate responses):

1. Ricky Gervais
2. Bob Denver

3. Steve O'Brien
4. Mark Linn-Baker
5. Seth Carwell

Example 5. Query: Who was the artist on the TV series Vision On?
Response: David Hasselhoff
Ground Truth Answer: Tony Hart

Perturbations (candidate responses):

1. Michael Welch
2. Brian Thompson
3. Burt Reynolds
4. Nicole Boyer
5. Jim Parsons

Example 6. Query: Who wrote the novel The Go Between?
Response: Edith Pattou
Ground Truth Answer: L. P. Hartley

Perturbations (candidate responses):

1. Jennifer Weiner
2. Jojo Moyes
3. John Irving
4. Margaret Mitchell
5. Alison Bechdel

N Variation: Query Extension Model

In the external uncertainty task, we investigate the impact of the query extension model by replacing the default `Claude3.7-Sonnet` with smaller models and reporting their corresponding results. Specifically, we explore the following query extension models:

- `Claude3.7-Sonnet` (Tables 1),
- `Qwen2.5-1.5B-Instruct` (Table 9),
- `Llama3-8B-Instruct` (Table 10),
- `Qwen3-8B` (Table 11).

In addition, for baseline methods that rely on token probabilities (e.g., *Last Token Entropy* and *Log Probabilities*), we enhance robustness by evaluating each method with multiple underlying models for token extraction, including `Qwen2.5-1.5B-Instruct`, `Llama3.2-1B-Instruct`, and `Llama3-8B-Instruct`.

Across all tested query extension models, our *Semantic Volume* method consistently outperforms the baselines in general. We also observe that stronger query extension models, such as `Claude3.7-Sonnet`, lead to generally improved performance for all sampling-based approaches.

O More Discussion

To support broader adoption and ensure practical robustness, we discuss several core aspects of our design choices.

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
LastTokenEntropy (Qwen2.5-1.5B-Instruct)	49.4	66.0	51.7	60.4
LastTokenEntropy (Llama3.2-1B-Instruct)	49.5	66.1	50.7	64.7
LastTokenEntropy (Llama3-8B-Instruct)	49.4	66.1	51.5	64.9
LogProb (Qwen2.5-1.5B-Instruct)	49.4	64.9	55.9	65.2
LogProb (Llama3.2-1B-Instruct)	48.9	63.8	52.7	64.2
LogProb (Llama3-8B-Instruct)	49.2	63.6	53.1	65.3
pTrue (Llama3-8B-Instruct)	$49.7_{0.3}$	$41.7_{0.4}$	$53.3_{0.6}$	$52.4_{0.9}$
pTrue (Mistral-7B-Instruct)	$50.0_{0.0}$	$0.0_{0.0}$	$50.0_{0.0}$	$0.0_{0.0}$
Lexical Similarity	$51.1_{0.3}$	$57.0_{0.6}$	$55.0_{0.1}$	$21.3_{0.3}$
Semantic Entropy	$49.4_{0.4}$	$48.7_{1.9}$	$49.9_{0.6}$	$64.6_{0.8}$
Semantic Volume (ours)	$\mathbf{57.8}_{0.7}$	$\mathbf{67.1}_{0.5}$	$\mathbf{57.5}_{0.8}$	$\mathbf{67.0}_{0.5}$

Table 9: Query Extension Model: `Qwen2.5-1.5B-Instruct`. External uncertainty: Accuracy and F1 on CLAMBER and AmbigQA. Sampling-based methods report mean $\pm$ std. over three trials (std. as subscript).

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
LastTokenEntropy (Qwen2.5-1.5B-Instruct)	49.4	66.0	51.7	60.4
LastTokenEntropy (Llama3.2-1B-Instruct)	49.5	66.1	50.7	64.7
LastTokenEntropy (Llama3-8B-Instruct)	49.4	66.1	51.5	64.9
LogProb (Qwen2.5-1.5B-Instruct)	49.4	64.9	55.9	65.2
LogProb (Llama3.2-1B-Instruct)	48.9	63.8	52.7	64.2
LogProb (Llama3-8B-Instruct)	49.2	63.6	53.1	65.3
pTrue (Llama3-8B-Instruct)	$49.1_{0.4}$	$42.0_{0.3}$	$51.9_{0.4}$	$54.7_{0.4}$
pTrue (Mistral-7B-Instruct)	$50.0_{0.0}$	$0.0_{0.0}$	$50.0_{0.0}$	$0.0_{0.0}$
Lexical Similarity	$51.2_{0.2}$	$62.5_{0.5}$	$49.9_{0.3}$	$66.0_{0.3}$
Semantic Entropy	$49.7_{0.2}$	$54.1_{5.2}$	$48.9_{0.5}$	$61.2_{2.2}$
Semantic Volume (ours)	$\mathbf{57.4}_{0.6}$	$\mathbf{66.3}_{0.5}$	$54.8_{0.6}$	$\mathbf{67.2}_{0.6}$

Table 10: Query Extension Model: `Llama3-8B-Instruct`. External uncertainty: Accuracy and F1 on CLAMBER and AmbigQA. Sampling-based methods report mean $\pm$ std. over three trials (std. as subscript).

Coverage of Confident Predictions. Semantic Volume is designed to quantify model uncertainty in cases where elevated dispersion reflects ambiguity or error. A separate class of failure modes involves *confidently incorrect responses*, where the model produces erroneous answers with high certainty. This is fundamentally a different problem: for example, if a model has been trained on faulty data (such as memorizing "1+1=3") then from its perspective, that answer is correct. In such cases, probing the model's token probabilities, hidden states, or sampled responses will not reveal any anomaly, because the internal uncertainty is low by design. Addressing these failures requires techniques that operate *outside* the model, such as factual verification, training data auditing, or external consistency checks. As such, confidently wrong responses lie outside the scope of our current work and require complementary solutions beyond intrinsic uncertainty measures.

Efficiency in Sampling and Computation. Our approach relies on sampling perturbed queries or responses (typically $n = 20$), in line with other sampling-based metrics. The computation of the semantic volume itself is highly efficient, requiring only seconds per dataset split and scaling linearly with the number of examples. Appendix H shows that even with $n = 10$, performance remains

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
LastTokenEntropy (Qwen2.5-1.5B-Instruct)	49.4	66.0	51.7	60.4
LastTokenEntropy (Llama3.2-1B-Instruct)	49.5	66.1	50.7	64.7
LastTokenEntropy (Llama3-8B-Instruct)	49.4	66.1	51.5	64.9
LogProb (Qwen2.5-1.5B-Instruct)	49.4	64.9	55.9	65.2
LogProb (Llama3.2-1B-Instruct)	48.9	63.8	52.7	64.2
LogProb (Llama3-8B-Instruct)	49.2	63.6	53.1	65.3
pTrue (Llama-3-8B-Instruct)	$49.1_{0.3}$	$29.1_{0.6}$	$49.8_{0.3}$	$39.6_{0.4}$
pTrue (Mistral-7B-Instruct)	$50.0_{0.0}$	$0.0_{0.0}$	$50.0_{0.0}$	$0.0_{0.0}$
Lexical Similarity	$51.9_{0.1}$	$51.7_{3.8}$	$51.3_{0.1}$	$60.3_{1.5}$
Semantic Entropy	$49.3_{0.9}$	$44.0_{9.7}$	$49.9_{0.8}$	$43.5_{5.5}$
Semantic Volume (ours)	$55.5_{0.8}$	$\mathbf{66.5}_{0.5}$	$\mathbf{57.2}_{0.7}$	$\mathbf{67.5}_{0.5}$

Table 11: Query Extension Model: `Qwen3-8B`. External uncertainty: Accuracy and F1 on CLAMBER and AmbigQA. Sampling-based methods report mean $\pm$ std. over three trials (std. as subscript).

competitive. This enables a practical accuracy–cost trade-off. Moreover, our method does not require log-probabilities or gradient access, making it compatible with both open and closed LLM APIs.

Calibrated Decision Thresholds. To convert continuous scores into binary predictions (e.g. ambiguous or not), we fit a threshold using a small, labeled subset. This process is non-iterative, uses a closed-form objective (Proposition 1), and requires only a modest annotation effort. Our experiments show that a threshold derived from just 100 examples is sufficient and generalizes well across datasets and response models. This enhances usability in low-label or cross-domain settings.

Distributional Assumptions in Theory. Our theoretical connection between semantic volume and differential entropy relies on the assumption that embedding perturbations are approximately Gaussian after PCA. This provides a useful analytical interpretation, but is not strictly required for the log-determinant to act as a reliable dispersion measure. As shown in Appendix E, PCA-normalized embeddings pass normality checks in the majority of cases, especially for query perturbations, supporting the practical soundness of the theoretical intuition.

Factuality Labeling Strategy. For hallucination detection, we use a ROUGE-L threshold to identify deviations from ground truth, following prior work. Although this heuristic may underestimate paraphrased but correct responses, a blind human validation on 1000 samples confirms over 95% agreement with our automatic labeling (Section 5). This indicates strong alignment between the heuristic and human judgment, offering a high-quality supervision signal at scale.