Efficient Hallucination Detection: Adaptive Bayesian Estimation of Semantic Entropy with Guided Semantic Exploration

Qiyao Sun[1], Xingming Li[1], Xixiang He[1], Ao Cheng[1], Xuanyu Ji[1], Hailun Lu[2], Runke Huang[3], Qingyong Hu[2]*

[1]National University of Defense Technology, Changsha, China
[2]Intelligent Game and Decision Lab, Beijing, China
[3]The Chinese University of Hong Kong, Shenzhen, China
{sunqiyao18, lixingming, hexixiang, chengao18, jixuanyu18}@nudt.edu.cn, Luhailun0728@outlook.com, runkehuang@cuhk.edu.cn, huqingyong15@outlook.com

Abstract

Large language models (LLMs) have achieved remarkable success in various natural language processing tasks, yet they remain prone to generating factually incorrect outputs—known as "hallucinations". While recent approaches have shown promise for hallucination detection by repeatedly sampling from LLMs and quantifying the semantic inconsistency among the generated responses, they rely on fixed sampling budgets that fail to adapt to query complexity, resulting in computational inefficiency. We propose an Adaptive Bayesian Estimation framework for Semantic Entropy with Guided Semantic Exploration, which dynamically adjusts sampling requirements based on observed uncertainty. Our approach employs a hierarchical Bayesian framework to model the semantic distribution, enabling dynamic control of sampling iterations through variance-based thresholds that terminate generation once sufficient certainty is achieved. We also develop a perturbation-based importance sampling strategy to systematically explore the semantic space. Extensive experiments on four QA datasets demonstrate that our method achieves superior hallucination detection performance with significant efficiency gains. In low-budget scenarios, our approach requires about 50% fewer samples to achieve comparable detection performance to existing methods, while delivers an average AUROC improvement of 12.6% under the same sampling budget.

1 Introduction

Large language models (LLMs) (Zhao et al. 2025) have demonstrated remarkable capabilities in language understanding, generation, and reasoning, fundamentally transforming the landscape of natural language processing (Chen et al. 2023; Lai and Nissim 2024). However, these models exhibit a critical limitation: they are prone to generating hallucination contents that appear plausible and coherent but lack factual grounding or contradict verifiable information (Wang et al. 2023a). Unlike traditional natural language generation systems where hallucinations primarily involve inconsistencies with source content, LLM hallucinations encompass a broader spectrum of factual errors and faithfulness issues due to their open-ended nature (Huang et al. 2025). The human-like fluency of LLM responses makes these hallucinations particularly difficult to detect, creating significant risks when deploying these models in real-world scenarios such as education, economics, science and so on (Lee et al. 2024; Wang et al. 2024; Luo et al. 2025). This fundamental challenge poses a substantial barrier to the reliable integration of LLMs into critical information systems where accuracy and trustworthiness are paramount.

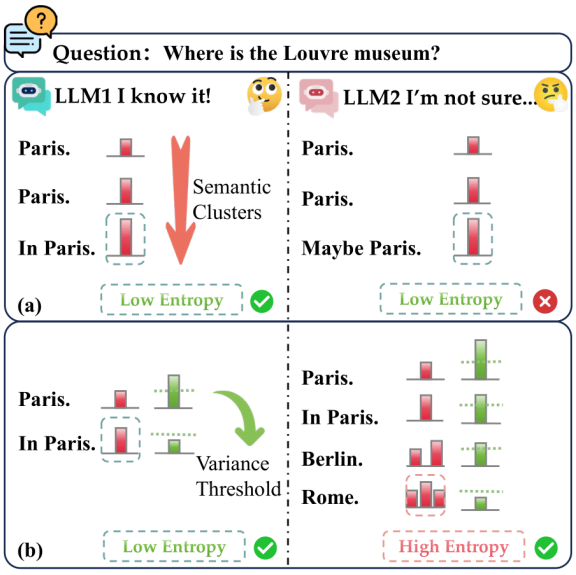


Figure 1: Comparison of fixed sampling (a) versus our adaptive Bayesian approach (b) for hallucination detection. Fixed sampling wastes computational resources on simple queries (LLM1) while failing to discover semantic diversity in complex cases (LLM2). Our method dynamically adjusts sampling based on variance thresholds, enabling efficient and accurate hallucination detection.

Current hallucination detection methods can be broadly categorized into four paradigms based on their underlying mechanisms. (1) External knowledge-based methods leverage retrieval systems to validate LLM outputs against authoritative sources (Choi et al. 2023; Dhuliawala et al. 2024), but face limitations in domain coverage and require com-

*Corresponding author.

plex verification pipelines. (2) Metacognition-based methods prompt LLMs to assess their own confidence (Kadavath et al. 2022), yet struggle with unreliability. (3) Single-sample methods analyze token-level patterns or hidden states within individual generations (Kossen et al. 2024; Zhang et al. 2023), offering computational efficiency but limited accuracy. (4) Multi-sample methods, particularly semantic entropy approaches (Farquhar et al. 2024), estimate uncertainty by clustering semantically equivalent outputs across multiple samples, demonstrating superior detection performance compared to other methods.

Despite their effectiveness, existing multi-sample methods suffer from a critical limitation: they employ fixed sampling budgets that fail to adapt to the inherent complexity of different queries. Simple factual questions may require only a few samples to reliably estimate semantic uncertainty, while complex or ambiguous queries necessitate extensive exploration of the semantic space. This one-size-fits-all approach results in computational waste for straightforward queries and insufficient sampling for challenging ones. Furthermore, current methods rely on multinomial sampling, which may repeatedly generate semantically similar outputs without efficiently exploring the full distribution of possible meanings. The lack of adaptive mechanisms and guided exploration strategies fundamentally limits the practical deployment of these methods in low-budget settings.

To address these limitations, we propose an **Adaptive Bayesian Estimation framework** for Semantic Entropy with **Guided Semantic Exploration**. Our approach introduces a hierarchical Bayesian framework that explicitly models the semantic distribution through a Dirichlet prior and decomposes the semantic entropy expectation via marginalization into two components: the posterior over the number of semantic categories and the conditional entropy given each possible category count. Building on this probabilistic foundation, we implement variance-based adaptive sampling that dynamically adjusts the number of queries based on posterior uncertainty, enabling efficient allocation of computational resources according to problem complexity. To accelerate variance convergence and enhance sampling efficiency, we develop a guided exploration strategy that identifies semantically critical tokens through importance weighting and systematically perturbs them to discover diverse interpretations, while employing importance sampling to maintain unbiased estimates.

Extensive experiments on four QA datasets demonstrate that our method achieves comparable or superior hallucination detection performance while significantly reducing the required number of model queries.

The main contributions can be summarized as follows:

- We introduce an adaptive Bayesian framework for semantic entropy estimation that dynamically adjusts sampling requirements based on observed uncertainty.
- We develop a guided semantic exploration strategy with importance sampling that discovers diverse semantic interpretations by perturbing critical tokens, accelerating convergence compared to random sampling.
- We demonstrate through comprehensive experiments that our method achieves state-of-the-art hallucination detection performance with significantly fewer samples, particularly excelling in resource-constrained scenarios.

2 Related Work

2.1 Hallucination

Hallucination in large language models refers to the generation of content that appears coherent but lacks factual support or deviates from the intended output (Wang et al. 2023a). Current research categorizes LLM hallucinations into two main types: factuality hallucination and faithfulness hallucination (Huang et al. 2025). Factuality hallucination occurs when generated content contains errors that contradict verifiable facts, while faithfulness hallucination manifests as inconsistencies with provided context or internal logic within the generated text. The ability of LLMs to produce highly convincing and human-like responses makes detecting these hallucinations particularly challenging, necessitating the development of robust detection methods.

2.2 Hallucination Detection

The detection of hallucinations in LLMs has emerged as a critical research area, with various approaches developed to identify when models generate factually incorrect or inconsistent content. Current detection methods can be broadly categorized into three main paradigms based on their underlying mechanisms.

External knowledge based methods use external knowledge bases to guide LLM generation and validate outputs (Choi et al. 2023; Dhuliawala et al. 2024; Min et al. 2023; Wang et al. 2023b). However, these approaches face significant challenges as they heavily depend on the accuracy and completeness of external knowledge sources, and struggle with domain-specific or rapidly evolving information.

Metacognition based methods utilize the metacognitive capabilities of LLMs—their ability to reflect on and evaluate their own knowledge states. The fundamental assumption is that LLMs develop implicit metacognitive signals during training that enable them to distinguish between confident, factual generations and uncertain, potentially hallucinated content (Kadavath et al. 2022). These methods face challenges in calibrating self-awareness across diverse domains and struggle when the model's metacognitive judgments are themselves unreliable.

Single-sample based methods require only a single forward pass through the model. FOCUS (Zhang et al. 2023) retrieve relevant tokens from the generation process and analyze token-level confidence patterns. While (Kossen et al. 2024) train specialized classifiers to predict uncertainty metrics directly from the model's hidden states. These methods offer computational efficiency but often face challenges in interpretability and accuracy.

Multi-sample based methods detect hallucinations by analyzing inconsistencies across multiple LLM outputs. Early approaches using Lexical Similarity (Lin, Liu, and Shang 2022) and Predictive Entropy (Kadavath et al. 2022)

often overestimated uncertainty due to diverse surface forms. Semantic Entropy (Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024) marked a breakthrough by clustering outputs into semantic equivalence classes and computing entropy over meaning distributions. While subsequent refinements (Duan et al. 2024; Bakman et al. 2024; Chen et al. 2024; Nikitin et al. 2024) and SDLG's (Aichberger et al. 2025) diversified exploration strategies improved performance, all existing methods rely on fixed sampling budgets that ignore query complexity. Our adaptive framework addresses this limitation by dynamically adjusting sample sizes based on inherent uncertainty, achieving computational efficiency with detection accuracy improvement.

3 Problem Formulation

Language Model Generation Let $\mathcal{X}$ denote the space of all possible prompts that can be presented to a language model. For a given prompt $x \in \mathcal{X}$, let $\mathcal{R}_x$ represent the set of all possible response sequences that the LLM can generate. We model the LLM's generation process as a conditional probability distribution $P_\theta(\mathbf{r}|x)$, where θ represents the model parameters and $\mathbf{r} \in \mathcal{R}_x$ is a response sequence.

Semantic Equivalence We define $\mathcal{M}_x$ as the set of distinct semantic meanings for prompt x. Let $f_x : \mathcal{R}_x \to \mathcal{M}_x$ be a mapping function that assigns each response $r \in \mathcal{R}_x$ to its corresponding meaning class $m \in \mathcal{M}_x$. This mapping induces a partition over the response space, where responses within the same partition are semantically equivalent.

Semantic Entropy Given a prompt x and the LLM's response distribution $P_\theta(\mathbf{r}|x)$, The probability of generating meaning m is:

$$p(m|x) = \sum_{r \in \mathcal{R}_x : f_x(r) = m} P_\theta(r|x) \tag{1}$$

The semantic entropy for prompt x is then defined as the Shannon entropy over the meaning distribution:

$$H_{sem} = -\sum_{m \in \mathcal{M}_x} p(m|x) \log p(m|x) \tag{2}$$

Estimation Problem In practice, we cannot enumerate all possible responses in $\mathcal{R}_x$ to compute the exact semantic entropy. Instead, treating H_{sem} as a random variable, we estimate it from a finite dataset obtained by sampling from the LLM. For a given prompt x, we generate N independent response sequences $r_1, \ldots, r_N \sim P_\theta(\cdot|x)$.

For each sampled sequence r_i, we determine its semantic meaning $m_i = f_x(r_i)$, yielding a corresponding list of meanings $m_1, \ldots, m_N$. Additionally, we obtain the generation probability $P_\theta(r_i|x)$ for each sequence. Thus, the estimation dataset is defined as:

$$\mathcal{D} = \{(r_1, m_1, P_\theta(r_1|x)), \ldots, (r_N, m_N, P_\theta(r_N|x))\} \tag{3}$$

We seek to develop an efficient semantic entropy estimator that dynamically increases N based on the observed sample variance, ensuring reliable yet minimal sampling cost.

Algorithm 1: Bayesian Estimation of Semantic Entropy

Input: Prompt x, LLM P_θ, variance threshold γ
Parameter: Initial samples N_0, top-k alternatives
Output: Semantic entropy $\hat{H}_{sem}$

// Initialize with weighted perplexity prior
Generate initial samples $\{r_1, \ldots, r_{N_0}\} \sim P_\theta(\cdot|x)$
Compute token importance weights $w_{i,j}$ ▷ **Eq. (15)**
Calculate weighted perplexity $\hat{\lambda}$ ▷ **Eq. (17)**
Set prior $p(K) \sim \text{Poisson}(\hat{\lambda})$
// Initialize dataset with initial samples
$\mathcal{D} \leftarrow \{(r_i, m_i, P_\theta(r_i|x)) : i = 1, \ldots, N_0\}$
Compute initial $\mathbb{E}[\mathbf{h}|\mathcal{D}]$ and $\text{Var}[\mathbf{h}|\mathcal{D}]$ ▷ **Eq. (4-5)**
// Adaptive sampling loop
while $\text{Var}[\mathbf{h}|\mathcal{D}] > \gamma$ **do**
  Generate new sample (r', w') via **Guided Semantic Exploration** from random $r \in \mathcal{D}$ using top-k ▷ **Section 4.2**
  Add $(r', m', P_\theta(r'|x))$ to $\mathcal{D}$
  // Bayesian update
  Update effective counts n_j and Dirichlet posterior parameters $\tilde{\alpha}_j$ with importance weight w' ▷ **Eq. (23-24)**
  Update truncated Dirichlet posterior constraints $\mathcal{C}$ ▷ **Eq. (6-7)**
  Compute posterior $p(K|\mathcal{D})$ ▷ **Eq. (11)**
  Recalculate $\mathbb{E}[\mathbf{h}|\mathcal{D}]$ and $\text{Var}[\mathbf{h}|\mathcal{D}]$ ▷ **Eq. (4-5)**
end while
return $\hat{H}_{sem} = \mathbb{E}[\mathbf{h}|\mathcal{D}]$

4 Method

We propose a hierarchical Bayesian framework for efficient semantic entropy estimation that addresses two fundamental challenges: uncertainty in the number of distinct semantic meanings and the computational cost of exhaustive sampling. Our approach decomposes the estimation problem by marginalizing over the unknown cardinality of the semantic space, employing a truncated Dirichlet posterior that incorporates generation probability constraints to provide tighter uncertainty bounds. To adaptively calibrate prior beliefs, we introduce a weighted perplexity metric that captures prompt-specific semantic diversity. Furthermore, we develop a guided semantic exploration strategy using importance sampling, which systematically perturbs semantically critical tokens to discover diverse interpretations while maintaining unbiased estimates. The framework dynamically adjusts sampling based on observed variance, enabling reliable semantic entropy estimation with minimal cost. We detail the hierarchical Bayesian framework (Section 4.1) and the guided exploration mechanism (Section 4.2) below.

4.1 Hierarchical Bayesian Framework

We propose a hierarchical Bayesian framework that stratifies the semantic entropy estimation based on the number of distinct semantic categories $|\mathcal{M}_x|$. Given the dataset $\mathcal{D}$ containing sampled responses and their semantic meanings, the key insight is that the semantic entropy H_{sem} fundamentally

depends on both the cardinality of the meaning space and the probability distribution over these meanings.

Let $K = |\mathcal{M}_x|$ denote the number of distinct semantic meanings in $\mathcal{M}_x$, and let $\mathbf{h}$ be a random variable that represents our belief about the value of H_{sem}. To compute its expected value given the observed dataset $\mathcal{D}$, we apply the law of total expectation to marginalize over the unknown number of semantic categories:

$$\mathbb{E}[\mathbf{h}|\mathcal{D}] = \sum_{K=1}^{\infty} \mathbb{E}[\mathbf{h}|K, \mathcal{D}] \cdot p(K|\mathcal{D}) \tag{4}$$

$$\text{Var}[\mathbf{h}|\mathcal{D}] = \mathbb{E}_K[\text{Var}[\mathbf{h}|K, \mathcal{D}]] + \text{Var}_K[\mathbb{E}[\mathbf{h}|K, \mathcal{D}]] \tag{5}$$

This hierarchical decomposition naturally separates the estimation problem into two components. $\mathbb{E}[\mathbf{h}|K, \mathcal{D}]$ represents the expected entropy conditioned on exactly K distinct semantic meanings being present in $\mathcal{M}_x$. This expectation is taken with respect to the posterior distribution of $\mathbf{p}$ given both K and the observed data. $p(K|\mathcal{D})$ captures our posterior belief about the cardinality of the meaning space after observing the sampled responses.

To enable adaptive sampling, we employ the total variance $\text{Var}[\mathbf{h}|\mathcal{D}]$ from Equation (5) as our stopping criterion. The sampling process terminates when $\text{Var}[\mathbf{h}|\mathcal{D}] < \gamma$, where γ is a predefined threshold that controls the trade-off between estimation accuracy and computational cost.

Calculation of $\mathbb{E}[\mathbf{h}|K, \mathcal{D}]$ and $\text{Var}[\mathbf{h}|K, \mathcal{D}]$ Given a fixed number of semantic categories K, we model the probability distribution over these categories using a Dirichlet prior. Let $\mathbf{p} = (p_1, \ldots, p_K)$ denote the probability vector where p_j represents the probability of generating semantic meaning $j \in \{1, \ldots, K\}$. We adopt a uninformative Dirichlet prior $\mathbf{p} \sim \text{Dir}(\alpha_0, \ldots, \alpha_0)$.

After observing the dataset $\mathcal{D}$, let $n_j = |\{i : m_i = j, i = 1, \ldots, N\}|$ denotes the number of sampled responses that map to semantic meaning j. Under standard Bayesian updating, the posterior would be $\text{Dir}(\alpha_0 + n_1, \ldots, \alpha_0 + n_K)$.

However, the LLM's generation probabilities provide additional constraints on the feasible probability space. For each semantic category j, the true probability p_j satisfy:

$$p_j \geq \sum_{r_i \in \mathcal{D}:\, f_x(r_i)=j} P_\theta(r_i|x) \triangleq b_j \tag{6}$$

This constraint arises because we have directly observed specific sequences belonging to category j with their generation probabilities. The constraint set is thus defined as:

$$\mathcal{C} = \{\mathbf{p} \in \Delta^{K-1} : p_j \geq b_j \text{ for all } j = 1, \ldots, K\} \tag{7}$$

where Δ^{K-1} denotes the $K-1$ dimensional probability simplex.

The posterior distribution becomes a truncated Dirichlet distribution. Let $\pi(\mathbf{p})$ denote the density of $\text{Dir}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_0 + n_1, \ldots, \alpha_0 + n_K)$. The truncated Dirichlet distribution over $\mathcal{C}$ has density:

$$\pi_{\mathcal{C}}(\mathbf{p}) = \begin{cases} \frac{\pi(\mathbf{p})}{Z_{\mathcal{C}}} & \text{if } \mathbf{p} \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $Z_{\mathcal{C}} = \int_{\mathcal{C}} \pi(\mathbf{p}) d\mathbf{p}$ is the normalization constant. The expected semantic entropy given K and $\mathcal{D}$ is:

$$\mathbb{E}[\mathbf{h}|K, \mathcal{D}] = \int H(\mathbf{p}) \cdot \pi_{\mathcal{C}}(\mathbf{p}) d\mathbf{p} \tag{9}$$

where $H(\mathbf{p}) = -\sum_{j=1}^{K} p_j \log p_j$ is the Shannon entropy. Similarly, the variance is:

$$\text{Var}[\mathbf{h}|K, \mathcal{D}] = \int H^2(\mathbf{p}) \cdot \pi_{\mathcal{C}}(\mathbf{p}) d\mathbf{p} - \mathbb{E}^2[\mathbf{h}|K, \mathcal{D}] \tag{10}$$

In practice, these integrals are computed via self-normalized importance sampling (Swaminathan and Joachims 2015). The detailed implementation is discussed in Appendix A.

Posterior Inference of $p(K|\mathcal{D})$ To compute the posterior distribution $p(K|\mathcal{D})$, we apply Bayes' theorem:

$$p(K|\mathcal{D}) = \frac{p(\mathcal{D}|K) \cdot p(K)}{\sum_{K'=1}^{\infty} p(\mathcal{D}|K') \cdot p(K')} \tag{11}$$

Calculation of marginal likelihood $p(\mathcal{D}|K)$ Now we compute the marginal likelihood $p(\mathcal{D}|K)$, which represents the probability of observing the dataset $\mathcal{D}$ given K semantic categories. This requires marginalizing over all possible probability distributions $\mathbf{p}$ consistent with K categories:

$$p(\mathcal{D}|K) = \int p(\mathcal{D}|\mathbf{p}, K) \cdot \pi_{\mathcal{C}}(\mathbf{p}) d\mathbf{p} \tag{12}$$

Given the observed category counts $\mathbf{n} = (n_1, \ldots, n_K)$ from the sampled responses, the likelihood follows a multinomial distribution:

$$p(\mathcal{D}|\mathbf{p}, K) = \frac{N!}{\prod_{j=1}^{K} n_j!} \prod_{j=1}^{K} p_j^{n_j} \tag{13}$$

The integral in Equation (12) can be efficiently evaluated using importance sampling employed for Equation (9).

Calculation of prior $p(K)$ via weighted perplexity We model the unknown number of semantic classes K using a Poisson prior with parameter λ:

$$p(K) = \frac{\lambda^K e^{-\lambda}}{K!} \tag{14}$$

This choice reflects the view that distinct semantic meanings emerge as rare events in the LLM's vast semantic space. The parameter λ controls the expected number of coherent semantic interpretations for a given prompt.

we propose an adaptive elicitation method that calibrates λ based on the LLM's inherent uncertainty for the specific prompt, which begins by generating a small initial sample of N_0 responses $\{r_1, \ldots, r_{N_0}\}$ from $P_\theta(\cdot|x)$. For each response r_i, we quantify the semantic importance of individual tokens to capture their contribution to the overall meaning. Let $r_i = (t_{i,1}, \ldots, t_{i,L_i})$ denote the token sequence of

length L_i. We compute the importance weight for token $t_{i,j}$ as:

$$w_{i,j} = 1 - \text{sim}(r_i, r_i \setminus \{t_{i,j}\}) \tag{15}$$

where sim measures the similarity between two sentences on a scale of 0 to 1 and $r_i^{(0)} \setminus \{t_{i,j}\}$ denotes the response with token $t_{i,j}$ removed.

Using these importance weights, we compute a weighted perplexity for each response that emphasizes semantically critical tokens:

$$\text{WPL}_i = \exp\left(-\frac{\sum_{j=1}^{L_i} w_{i,j} \log P_\theta(t_{i,j}|t_{i,<j}, x)}{\sum_{j=1}^{L_i} w_{i,j}}\right) \tag{16}$$

where $P_\theta(t_{i,j}|t_{i,<j}, x)$ is the conditional probability of token $t_{i,j}$ given the preceding tokens and prompt.

The empirical estimate of λ is then obtained as:

$$\hat{\lambda} = \frac{1}{N_0} \sum_{i=1}^{N_0} \text{WPL}_i \tag{17}$$

This weighted perplexity serves as a principled proxy for semantic diversity: higher values indicate greater uncertainty in the model's semantic space, suggesting more potential semantic categories.

To handle the infinite summation in Equation (11), we employ a truncation strategy. Since $p(K)$ decays exponentially for large K under the Poisson prior, we truncate the summation at $K_{max} = \max(K_{obs}, 3\lambda)$, where K_{obs} denotes the number of distinct semantic meanings observed in $\mathcal{D}$.

4.2 Guided Semantic Exploration

To improve the efficiency of semantic entropy estimation, we develop a guided exploration strategy that leverages importance sampling to systematically explore the LLM's semantic space. By strategically perturbing tokens at semantically critical positions and continuing generation from these points, we can discover diverse semantic interpretations while maintaining computational efficiency.

Guided Language Generation We employ a perturbation-based approach to construct sequences that explore alternative semantic branches. For a given response $r = (t_1, \ldots, t_L) \sim P_\theta(\cdot|x)$, we first identify semantically critical positions using the token importance weights defined in Equation (15). Let $\mathcal{I} = \{i_1, i_2, \ldots, i_L\}$ denote the indices of all tokens ordered by their importance weights in descending order, where $w_{i_1} \geq w_{i_2} \geq \ldots \geq w_{i_L}$.

At each critical position $i_j \in \mathcal{I}$, we examine the conditional token distribution $P_\theta(\cdot|t_{<i_j}, x)$ and identify the top-k alternative tokens excluding the original choice:

$$\mathcal{A}_{i_j} = \text{top-}k\{t \in \mathcal{V} \setminus \{t_{i_j}\} : P_\theta(t|t_{<i_j}, x)\} \tag{18}$$

where $\mathcal{V}$ denotes the vocabulary. For each alternative token $t' \in \mathcal{A}_{i_j}$, we generate a new response by replacing t_{i_j} with t' and continuing the generation:

$$r' = (t_1, \ldots, t_{i_j-1}, t', t'_{i_j+1}, \ldots, t'_{L'}) \tag{19}$$

where $t'_{i_j+1}, \ldots, t'_{L'}$ are sampled from $P_\theta(\cdot|t_1, \ldots, t_{i_j-1}, t', x)$.

Importance Sampling The guided semantic exploration process described above forcibly modifies certain tokens in the generated sequences, deviating from the LLM's original distribution $P_\theta(\cdot|x)$. Since we are no longer sampling directly from $P_\theta(\cdot|x)$, we must correct for this bias through importance sampling with a properly defined proposal distribution $q(\mathbf{r}|x)$ that accurately captures our modified sampling procedure.

The proposal distribution is defined as:

$$q(\mathbf{r}|x) = \sum_{r' \in \mathcal{R}_x} p(\mathbf{r}|r', x) \cdot P_\theta(r'|x) \tag{20}$$

where $p(\mathbf{r}|r', x)$ represents the probability of transforming an initial response r' into $\mathbf{r}$ through our guided generation process. Under specific assumptions about token selection (Aichberger et al. 2025), the proposal distribution takes the form:

$$q(\mathbf{r}|x) = \frac{P_\theta(\mathbf{r}|x)}{P_\theta(t_j|\mathbf{t}_{<j}, x)} \tag{21}$$

where j is the index of the perturbed token in sequence $\mathbf{r} = (t_1, \ldots, t_L)$.

The importance weight for a sample $\mathbf{r}$ drawn from $q(\cdot|x)$ is:

$$w(\mathbf{r}) = \frac{P_\theta(\mathbf{r}|x)}{q(\mathbf{r}|x)} = P_\theta(t_j|\mathbf{t}_{<j}, x) \tag{22}$$

This weight ensures unbiased estimation while promoting exploration of lower-probability but potentially semantically distinct sequences (See Appendix B for proof).

Bayesian Update with Weighted Samples When incorporating samples obtained through importance sampling into our hierarchical Bayesian framework, we must properly account for their importance weights. Let $\{(r^{(1)}, w^{(1)}), \ldots, (r^{(N-1)}, w^{(N-1)})\}$ denote the sequence of weighted samples, where $r^{(t)}$ is drawn from $q(\cdot|x)$ with importance weight $w^{(t)}$.

We modify the effective counts to reflect the importance weights. If sample $r^{(N)}$ is assigned to semantic category j, the effective count update becomes:

$$n_j^{(N)} = n_j^{(N-1)} + w^{(N)} \tag{23}$$

The Dirichlet posterior parameters are then updated as:

$$\alpha_j^{(N)} = \alpha_0 + n_j^{(N)} \tag{24}$$

To maintain proper normalization, we scale the posterior parameters:

$$\tilde{\alpha}_j^{(N)} = \alpha_j^{(N)} \cdot \frac{\sum_{k=1}^K \alpha_k^{(0)} + N}{\sum_{k=1}^K \alpha_k^{(t)}} \tag{25}$$

This scaling ensures that the effective sample size grows linearly with T while properly weighting each sample's contribution according to its importance in exploring the semantic space.

The modified likelihood for computing $p(\mathcal{D}|\mathbf{p}, K)$ in Equation (13) becomes:

$$p(\mathcal{D}|\mathbf{p}, K) = \frac{\Gamma(N)}{\prod_{j=1}^K \Gamma(n_j{}^{(T)})} \prod_{j=1}^K p_j^{n_j{}^{(T)}} \tag{26}$$

LLM	Dataset	P(True)	N=2				N=5			
			SAR	SE	SE_{SDLG}	**OURS**	SAR	SE	SE_{SDLG}	**OURS**
Llama-2-7B	CoQA	.468	.591	.609	.618	**.695**	.627	.683	.688	**.748**
	TriviaQA	.488	.708	.710	.724	**.835**	.741	.795	.827	**.897**
	TruthfulQA	.509	.598	.621	.635	**.732**	.632	.713	.724	**.795**
	SimpleQA	.521	.721	.796	.891	**.895**	.768	.930	.956	**.959**
Llama-3.1-8B	CoQA	.568	.655	.699	.648	**.738**	.648	.756	.769	**.799**
	TriviaQA	.714	.728	.759	.741	**.855**	.563	.866	.850	**.913**
	TruthfulQA	.633	.661	.642	.655	**.753**	.603	.741	.750	**.818**
	SimpleQA	.655	.652	.825	.830	**.913**	.614	.930	.934	**.942**
Mistral-Small-24B	CoQA	.618	.643	.658	.669	**.762**	.660	.758	.765	**.825**
	TriviaQA	.624	.641	.689	.703	**.872**	.778	.790	.817	**.928**
	TruthfulQA	.597	.622	.661	.673	**.773**	.663	.765	.772	**.839**
	SimpleQA	.668	.645	.725	.737	**.772**	.681	.868	**.873**	.871

Table 1: AUROC for hallucination detection on open-form QA datasets across three representative LLMs. N denotes the sampling budget, representing the average number of response samples generated per query for uncertainty estimation. The best results are in **bold** and the second best is marked with underline.

5 Experimental Setup

Datasets and models We evaluate on four free-form QA datasets: CoQA (Reddy, Chen, and Manning 2019), TriviaQA (Joshi et al. 2017), TruthfulQA(Lin, Hilton, and Evans 2022), and SimpleQA (Wei et al. 2024), covering both open-book and closed-book scenarios. All experiments use zero-shot settings. We test three LLMs: Llama-2-7B (Touvron et al. 2023), Llama-3.1-8B (Dubey et al. 2024), and Mistral-Small-24B (Rastogi et al. 2025), ensuring generalization across different architectures and scales.

Evaluation We measure hallucination detection performance using AUROC (Bradley 1997), treating estimator outputs as binary classification scores. Following (Li et al. 2024), we use GPT-4.1 to judge response correctness with a Pass-All@3 method: sampling three responses per question and marking as hallucination if any response contradicts ground truth. For efficiency comparison, we evaluate AUROC at sampling budgets N=1 to 10. As our method uses adaptive sampling, we calibrate variance thresholds to match baseline sample counts.

Baselines We compare against four methods: **P(True)** (Kadavath et al. 2022) uses LLM self-assessment; **SAR** (Duan et al. 2024) aggregates token-level prediction entropy; **SE** (Farquhar et al. 2024) clusters semantic equivalents and computes entropy; **SE_{SDLG}** (Aichberger et al. 2025) enhances SE with targeted perturbations for diverse outputs.

Implementation. We conduct our experiments on a single A800 80GB GPU. Semantic clustering uses DeBERTa-v3-large (He, Gao, and Chen 2023) for NLI-based equivalence detection. All generation uses temperature 1.0. N_0 is set to 1 to reduce the initialization overhead, and top-k is set to 3. Adaptive sampling variance thresholds are calibrated to achieve average sample counts comparable to baselines.

6 Results and Analyses

6.1 Main Results

Since the computational overhead beyond LLM sampling costs is negligible for all methods (detailed analysis in Appendix C), we set equivalent sampling budgets across different approaches to ensure fair comparison. Table 1 presents the performance comparison across multiple models and datasets. **First**, our method achieves the highest AUROC in 23 out of 24 settings, with up to 16.9% improvement over the strongest baseline SE_{SDLG} (TriviaQA, Mistral-Small-24B, N=2). **Second**, the advantage is most pronounced in low-sample regimes: 12.6% average improvement at N=2 versus 6.3% at N=5, confirming that our adaptive sampling efficiently explores semantic space under computational constraints. **Third**, consistent improvements across different model architectures and QA domains demonstrate that our method captures fundamental semantic uncertainty properties rather than dataset-specific patterns, ensuring robust practical deployment.

6.2 Ablation Studies

We conduct ablation studies on TriviaQA using Llama-3.1-8B to analyze each component's contribution (Table 2).

Prior Estimation Replacing adaptive prior with fixed $K = K_{obs} + 1$ causes 4.3% (N=2) and 3.5% (N=5) performance drops, confirming that weighted perplexity effectively captures prompt-specific semantic diversity.

Adaptive Sampling Fixed sampling without Bayesian framework degrades performance by 9.6% (N=2) and 4.7% (N=5). Adding Bayesian framework to fixed sampling partially recovers performance but still underperforms by 5.7% and 2.8%. This shows: (1) Bayesian uncertainty quantification benefits all sampling strategies, and (2) variance-based

Method	N=2	N=5
Full Model	**.855**	**.913**
Prior Estimation		
w/o adaptive prior ($K = K_{obs} + 1$)	.812	.878
Adaptive Sampling		
Fixed sampling w/o Bayesian	.759	.866
Fixed sampling w/ Bayesian	.798	.885
Exploration Strategy		
w/o guided exploration	.823	.892

Table 2: Ablation study on TriviaQA using Llama-3.1-8B, showing AUROC performance when removing key components of our method.

Model	CoQA	TriviaQA	TruthfulQA	SimpleQA
Llama-2-7B	27.4%	37.5%	58.8%	98.5%
Llama-3.1-8B	15.9%	34.7%	52.2%	76.2%
Mistral-Small-24B	13.7%	21.2%	46.9%	71.5%

Table 3: Hallucination rates across different models and QA datasets using the Pass-All@3 evaluation method.

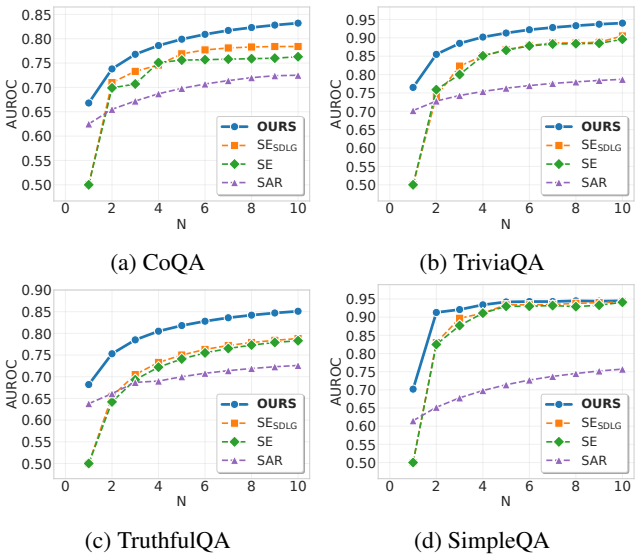
adaptive sampling is crucial for efficiency, especially in low-sample regimes.

Exploration Strategy Removing guided exploration decreases performance by 3.2% (N=2) and 2.1% (N=5), with larger impact at smaller budgets, confirming that importance sampling explores the semantic space and accelerates semantic discovery under constrained resources.

6.3 Further Analyses

Bayesian Framework as a Superior Estimator Our method maintains advantages even at higher sampling budgets (Figure 2) due to hierarchical uncertainty modeling: semantic space cardinality via $p(K|\mathcal{D})$ and probability distributions via truncated Dirichlet posteriors. Unlike traditional methods assuming fixed semantic categories, we account for unobserved meanings through posterior inference. Generation probability constraints (Equation 6) further tighten the feasible space, improving accuracy even when extensive sampling would reveal most variations.

Adaptive Resource Allocation Figure 3 shows our adaptive strategy allocating resources by query complexity. Simple datasets (CoQA) exhibit high variance—many queries need only 1-2 samples due to rapid convergence. Complex datasets (SimpleQA) concentrate near the budget limit, indicating persistent uncertainty. This aligns with intuition: high-probability, consistent responses trigger early termination through decreased posterior variance, ensuring efficiency without sacrificing accuracy for real-world queries.



(a) CoQA (b) TriviaQA

(c) TruthfulQA (d) SimpleQA

Figure 2: AUROC performance comparison of hallucination detection methods on Llama-3.1-8B across varying sampling budgets N.

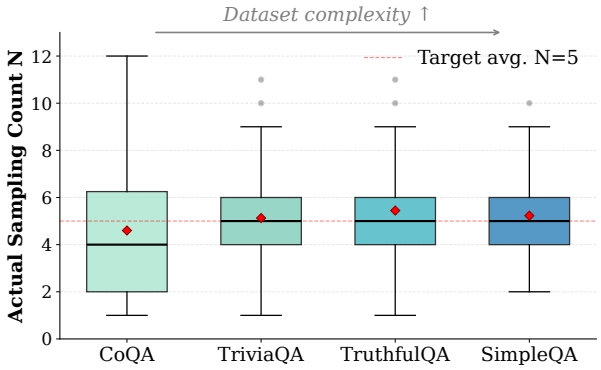


Figure 3: Distribution of actual sampling counts under a fixed average budget of N=5 across four QA datasets using Llama-3.1-8B.

7 Conclusions

In this paper, we propose an adaptive Bayesian estimation framework for semantic entropy that addresses the computational inefficiency of existing hallucination detection methods. Our hierarchical approach models semantic distributions through a Dirichlet prior, while guided exploration with importance sampling discovers diverse interpretations. Variance-based adaptive sampling dynamically allocates resources according to query complexity. Experiments demonstrate consistent superiority across multiple models and datasets, with significantly fewer samples required, particularly in low-budget scenarios. The framework's efficient resource allocation makes it practical for real-world deployment. Future directions include extensions to multimodal settings and broader uncertainty quantification tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62306331 and CAAI Youth Talent Lifting Project under Grant CAAI2023-2025QNRC001.

References

Aichberger, L.; Schweighofer, K.; Ielanskyi, M.; and Hochreiter, S. 2025. Improving Uncertainty Estimation through Semantically Diverse Language Generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Bakman, Y. F.; Yaldiz, D. N.; Buyukates, B.; Tao, C.; Dimitriadis, D.; and Avestimehr, S. 2024. MARS: Meaning-Aware Response Scoring for Uncertainty Estimation in Generative LLMs. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 7752–7767. Association for Computational Linguistics.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, 30(7): 1145–1159.

Chen, C.; Liu, K.; Chen, Z.; Gu, Y.; Wu, Y.; Tao, M.; Fu, Z.; and Ye, J. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Chen, L.; Deng, Y.; Bian, Y.; Qin, Z.; Wu, B.; Chua, T.; and Wong, K. 2023. Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 6325–6341. Association for Computational Linguistics.

Choi, S.; Fang, T.; Wang, Z.; and Song, Y. 2023. KCTS: Knowledge-Constrained Tree Search Decoding with Token-Level Hallucination Detection. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 14035–14053. Association for Computational Linguistics.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 3563–3578. Association for Computational Linguistics.

Duan, J.; Cheng, H.; Wang, S.; Zavalny, A.; Wang, C.; Xu, R.; Kailkhura, B.; and Xu, K. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 5050–5063. Association for Computational Linguistics.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

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

He, P.; Gao, J.; and Chen, W. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2): 42:1–42:55.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1601–1611. Association for Computational Linguistics.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; Showk, S. E.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. *CoRR*, abs/2207.05221.

Kossen, J.; Han, J.; Razzak, M.; Schut, L.; Malik, S. A.; and Gal, Y. 2024. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. *CoRR*, abs/2406.15927.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Lai, H.; and Nissim, M. 2024. A Survey on Automatic Generation of Figurative Language: From Rule-based Systems to Large Language Models. *ACM Comput. Surv.*, 56(10): 244.

Lee, J.; Stevens, N.; Han, S. C.; and Song, M. 2024. A Survey of Large Language Models in Finance (FinLLMs). *CoRR*, abs/2402.02315.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2024. From Generation to Judgment:

Opportunities and Challenges of LLM-as-a-judge. *CoRR*, abs/2411.16594.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 3214–3252. Association for Computational Linguistics.

Lin, Z.; Liu, J. Z.; and Shang, J. 2022. Towards Collaborative Neural-Symbolic Graph Semantic Parsing via Uncertainty. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, 4160–4173. Association for Computational Linguistics.

Luo, Z.; Yang, Z.; Xu, Z.; Yang, W.; and Du, X. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research. *CoRR*, abs/2501.04306.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 12076–12100. Association for Computational Linguistics.

Nikitin, A.; Kossen, J.; Gal, Y.; and Marttinen, P. 2024. Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Rastogi, A.; Jiang, A. Q.; Lo, A.; Berrada, G.; Lample, G.; Rute, J.; Barmentlo, J.; Yadav, K.; Khandelwal, K.; Chandu, K. R.; Blier, L.; Saulnier, L.; et al. 2025. Magistral. *CoRR*, abs/2506.10910.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. CoQA: A Conversational Question Answering Challenge. *Trans. Assoc. Comput. Linguistics*, 7: 249–266.

Swaminathan, A.; and Joachims, T. 2015. The Self-Normalized Estimator for Counterfactual Learning. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 3231–3239.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Wang, C.; Liu, X.; Yue, Y.; Tang, X.; Zhang, T.; Jiayang, C.; Yao, Y.; Gao, W.; Hu, X.; Qi, Z.; Wang, Y.; Yang, L.; Wang, J.; Xie, X.; Zhang, Z.; and Zhang, Y. 2023a. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. *CoRR*, abs/2310.07521.

Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P. S.; and Wen, Q. 2024. Large Language Models for Education: A Survey and Outlook. *CoRR*, abs/2403.18105.

Wang, X.; Yan, Y.; Huang, L.; Zheng, X.; and Huang, X. 2023b. Hallucination Detection for Generative Large Language Models by Bayesian Sequential Estimation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 15361–15371. Association for Computational Linguistics.

Wei, J.; Karina, N.; Chung, H. W.; Jiao, Y. J.; Papay, S.; Glaese, A.; Schulman, J.; and Fedus, W. 2024. Measuring short-form factuality in large language models. *CoRR*, abs/2411.04368.

Zhang, T.; Qiu, L.; Guo, Q.; Deng, C.; Zhang, Y.; Zhang, Z.; Zhou, C.; Wang, X.; and Fu, L. 2023. Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 915–932. Association for Computational Linguistics.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2025. A Survey of Large Language Models. arXiv:2303.18223.