Energy-Guided Ghost-Cluster Bayesian Semantic Entropy for Hallucination Detection

Anonymous ACL submission

Abstract

Semantic Entropy (SE) is a robust metric for uncertainty quantification but suffers from high computational costs and an inability to detect "consistent hallucinations", where a model confidently repeats the same incorrect answer. This failure stems from SE's closed-set assumption, which neglects plausible semantic alternatives outside the sampled outputs. To address these limitations, we propose **GLIB-SE** (Ghost-Logit Integrated Bayesian Semantic Entropy). This Bayesian framework models the semantic distribution with an explicit "Ghost Cluster" for unobserved semantics. Crucially, we utilize the model's raw logit energy as a dynamic prior to calibrate this cluster, allowing GLIB-SE to expose hidden uncertainty when a model is internally confused despite generating consistent outputs. Furthermore, we derive an adaptive sampling strategy based on posterior entropy variance to optimize the inference budget. Experiments across six benchmarks demonstrate that GLIB-SE significantly outperforms baselines in hallucination detection (AUROC) while reducing sampling costs by over 30% compared to fixed-sample strategies.

1 Introduction

Large Language Models (LLMs) have become foundational to modern AI, excelling in tasks from open-domain question answering to complex reasoning. Despite these capabilities, their reliable deployment in high-stakes domains is severely hindered by hallucinations producing fluent but factually incorrect content (Ji et al., 2023; Zhang et al., 2023). To mitigate this risk, Uncertainty Quantification (UQ) has emerged as a critical defense mechanism, enabling systems to detect errors, trigger abstention, or solicit human intervention (Huang et al., 2023). Ideally, a robust UQ metric should serve as a faithful proxy for the model's knowledge boundary: exhibiting high uncertainty precisely when the model is "confidently wrong" or lacks the necessary information (Kadavath et al., 2022).

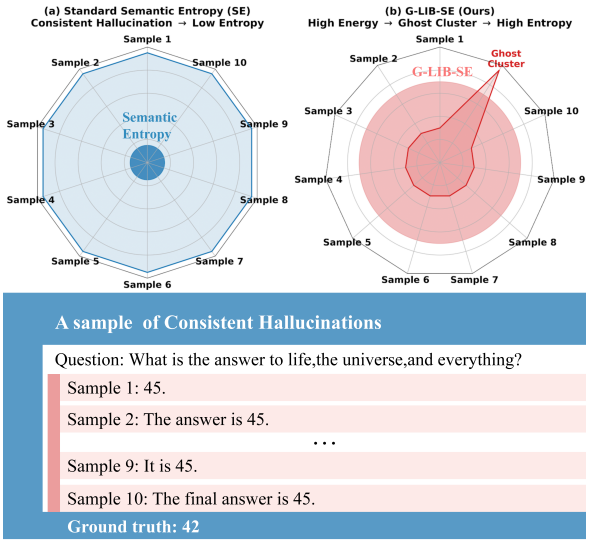

Figure 1: For the same query, the model does not know the correct answer yet produces several semantically consistent wrong responses. Standard Semantic Entropy cannot distinguish this regime, and GLIB-SE uses the **Ghost Cluster** to explicitly flag these hallucinations.

Among existing UQ methodologies, Semantic Entropy (SE) has established itself as a prevailing baseline due to its ability to handle semantic equivalence in natural language generation (Farquhar et al., 2024; Kuhn et al., 2023). Unlike traditional token-level entropy, SE computes uncertainty based on the distribution of semantic classes by clustering multiple sampled responses, thereby avoiding inflated entropy scores caused by lexical variations of identical meanings. Yet, as illustrated in Figure 1 this closed-set view collapses under "consistent hallucinations": the language model can return the same wrong answer with low entropy but high risk. Although SE excels at detecting divergent errors, its reliance on a "closed-set assumption" presents a fundamental flaw when confronting "consistent hallucinations"(Xu et al.,

2025). Consistency-based detectors such as SelfCheckGPT (Manakul et al., 2023) share this weakness: when the model repeatedly produces the same incorrect response, semantic divergence is absent and the detector remains silent. This failure stems from estimating probabilities solely based on limited sampled observations, ignoring the true confidence information embedded in the model's internal latent states. Furthermore, achieving stable entropy estimates typically requires a fixed and substantial number of samples, imposing high computational overhead during inference and limiting real-time applicability.

To address these challenges, we propose **GLIB-SE** (Ghost-Logit Integrated Bayesian Semantic Entropy). Fundamentally, standard SE estimates semantic probabilities solely based on observed cluster counts; under consistent hallucinations, this closed-set view falsely assigns zero probability to unobserved alternatives, leading to severe overconfidence. GLIB-SE corrects this bias by adopting a Bayesian framework that explicitly allocates probability mass to a Ghost Cluster which construct representing the unobserved semantic space (the "unknown unknowns"). Crucially, we utilize the model's raw logit energy as a dynamic prior to calibrate this Ghost mass: when the model exhibits internal confusion, we inflate the weight of the Ghost Cluster, effectively uncovering the hidden uncertainty behind consistent outputs.

A key innovation of GLIB-SE lies in utilizing the model's raw logits as a dynamic prior to guide the weight estimation of the Ghost Cluster. When the model exhibits internal confusion (manifested as a uniform logit distribution), our mechanism assigns higher probability mass to the Ghost Cluster, even if the output semantics appear highly consistent. This effectively calibrates the final entropy to warn of potential risks. This approach synergizes intrinsic logit information with extrinsic semantic distributions, bridging the gap left by single-perspective methods. Moreover, leveraging the Bayesian framework, we derive the variance of the posterior entropy to design an adaptive sampling strategy. This strategy enables automatic termination of generation upon the convergence of uncertainty estimates, significantly reducing computational costs without compromising detection performance.

Our main contributions are threefold:

- **Ghost-Cluster Bayesian Framework:** We identify the "closed-set limitation" of Semantic Entropy as the root cause of its failure on consistent hallucinations and propose a Bayesian solution with an explicit **Ghost Cluster** to model the unknown.

- **Energy-Guided Prior Modulation:** We propose an energy-guided dynamic prior that utilizes raw logits to detect internal model confusion, enabling GLIB-SE to distinguish between "justified confidence" (true knowledge) and "blind confidence" (hallucination).

- **Pareto-Efficient Inference:** We implement a variance-based adaptive sampling algorithm, achieving a superior Pareto frontier between detection accuracy and computational cost, reducing sampling overhead by $\sim$30-50% on complex reasoning tasks.

2 Related Work

2.1 Semantic Uncertainty and Its Limits.

Uncertainty quantification is pivotal for trustworthy LLMs. While early metrics like perplexity (Min et al., 2020) and recent supervised token-probability approaches (Quevedo Caballero et al., 2024) are sensitive to lexical variations, **Semantic Entropy (SE)** (Farquhar et al., 2024; Kuhn et al., 2023) mitigates this by clustering semantically equivalent responses. However, SE relies on a *closed-set assumption*, equating observed cluster frequency with true probability. Consequently, SE fails under **consistent hallucinations**, where a model confidently repeats the same incorrect answer (i.e., a single cluster), causing entropy to collapse to zero despite factual error. GLIB-SE addresses this by explicitly modeling unobserved semantics to break the closed-set constraint.

2.2 Consistency Methods and Sampling Cost.

Consistency-based detectors such as SelfCheck-GPT (Manakul et al., 2023) flag divergence across generated answers but fail when hallucinations are semantically consistent. High-quality uncertainty methods like Deep Ensembles (Lakshminarayanan et al., 2017) improve calibration but require training multiple models, making them costly. GLIB-SE instead couples semantic structure with internal energy and uses adaptive sampling to reach a Pareto-efficient frontier without training ensembles (Li et al., 2025).

2.3 Bayesian Hallucination Detection.

Prior works like Wang et al. (2023); Ciosek et al. (2025) have utilized Bayesian sequential frameworks to optimize query budgets.However, these works lack a mechanism to distinguish between *justified confidence* (knowledge) and *blind confidence* (hallucination). GLIB-SE extends this framework by making the ghost prior *dynamic*, modulated by the model's internal confusion.

2.4 Logit-Based Energy Signals.

Internal states often reveal uncertainty even when outputs are consistent. Energy-based OOD detection (Liu et al., 2020) establishes the free-energy link for detecting out-of-distribution or erroneous samples via logits. **Semantic Energy** (Ma et al., 2025) leverages the logit landscape as a truthfulness proxy. *Distinction:* While Ma et al. demonstrate the utility of energy, they primarily operate at the sample level or simple aggregation, without integrating energy into the semantic clustering structure. In contrast, GLIB-SE fuses semantic energy directly into the Bayesian framework as a prior. This allows high internal energy to actively expand the probability mass of the Ghost Cluster, effectively detecting consistent hallucinations that output-based clustering alone would miss.

2.5 Kernel-Based Uncertainty.

Moving beyond hard clustering, **Kernel Language Entropy (KLE)** (Nikitin et al., 2024) uses continuous semantic kernels to capture fine-grained similarities. While rigorous, KLE requires computing an $N \times N$ similarity matrix, leading to quadratic complexity $O(N^2)$ that limits scalability. GLIB-SE maintains the linear efficiency $O(N)$ of hard clustering while achieving robust calibration through the energy-guided Ghost Cluster, offering a more Pareto-efficient solution for real-time applications.

3 Methodology

This section formalizes the uncertainty estimation problem, diagnoses the closed-set failure of standard Semantic Entropy (SE), and presents **GLIB-SE**. The framework combines: (1) **Bayesian Semantic Modeling** with an explicit Ghost Cluster for unobserved semantics; (2) a **Logit-Informed Dynamic Prior** driven by Semantic Energy; and (3) a **Variance-Based Adaptive Sampling** strategy. These components are designed jointly for robustness to consistent errors and for **computational efficiency**, decoupling uncertainty estimation from fixed sampling budgets.

3.1 Problem Formulation & Preliminaries

Given an input context $\mathbf{x}$, an LLM generates a response sequence $\mathbf{y}$ according to $p(\mathbf{y} \mid \mathbf{x})$. Since multiple sequences can convey the same meaning, we define a semantic clustering function $\mathcal{C}(\cdot)$ such that $\mathcal{C}(\mathbf{y}_i) = c_k$ if $\mathbf{y}_i$ belongs to semantic class c_k.

Semantic Entropy (SE) quantifies uncertainty over meanings. For N sampled sequences $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ grouped into K clusters with counts $\mathbf{n} = [n_1, \ldots, n_K]$, the standard estimator is:

$$\hat{H}_{\text{SE}}(\mathbf{x}) = -\sum_{k=1}^{K} \frac{n_k}{N} \log \frac{n_k}{N}. \tag{1}$$

Limitation: This frequentist approach suffers from the *closed-set assumption*, assuming the N samples cover the entire semantic space. Under *consistent hallucination*, where the model confidently generates the same incorrect answer N times (i.e., $K = 1$), $\hat{H}_{\text{SE}}$ collapses to 0. This fails to capture the *epistemic uncertainty* arising from the finite sample size and ignores the model's internal confusion.

Consistent Hallucination. We define a consistent hallucination as a scenario where the model generates a set of responses Y that are semantically equivalent ($K = 1, n_1 = N$) and confident (low predictive entropy), yet factually incorrect. In this case, standard SE yields $\hat{H}_{\text{SE}}(\mathbf{x}) = 0$, failing to signal risk. Our goal is to leverage internal states so that low entropy reflects true knowledge (low energy) rather than blind confidence (high energy).

3.2 GLIB-SE Framework

We propose to estimate the semantic distribution $\boldsymbol{\pi}$ in a Bayesian setting, dynamically guided by the model's internal state. Figure 2 presents an overview of the proposed GLIB-SE framework.

3.2.1 Bayesian Modeling of Semantic Distribution

We introduce an unobserved **Ghost Cluster** c_{g} to represent semantics not seen in the sampled responses, extending $\boldsymbol{\pi}$ to $[\pi_1, \ldots, \pi_K, \pi_{\text{g}}]$ following Bayesian SE (Ciosek et al., 2025). With Dirichlet modeling,

$$\boldsymbol{\pi} \mid Y \sim \text{Dir}(\alpha_{\text{base}}+n_1, \ldots, \alpha_{\text{base}}+n_K, \alpha_{\text{g}}), \tag{2}$$

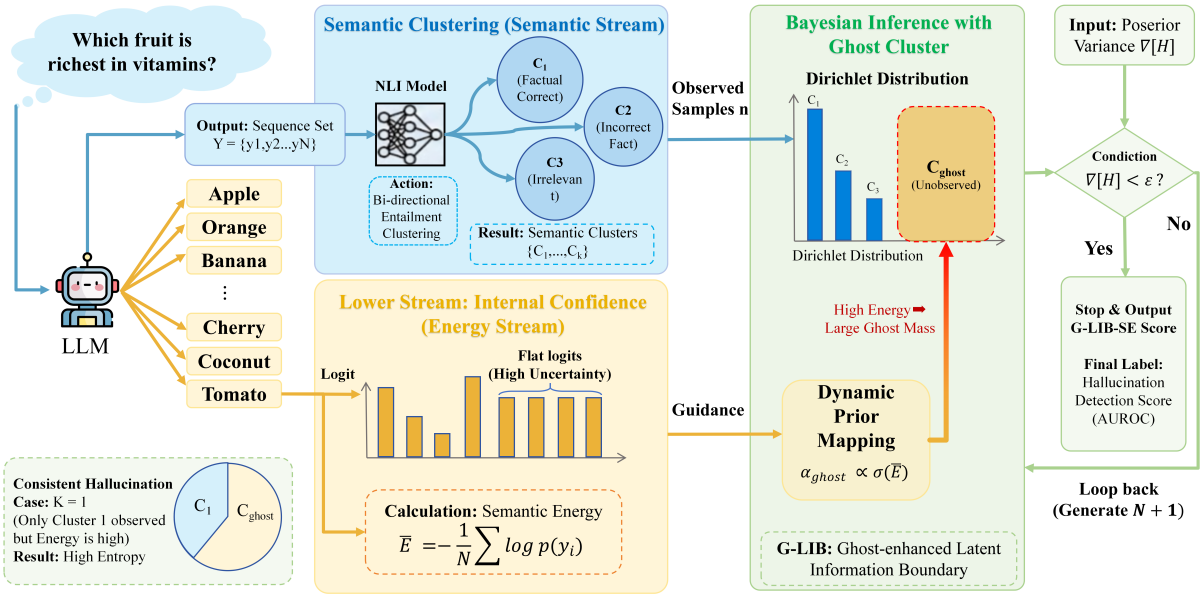

Figure 2: Overview of the GLIB-SE framework. Given an input query, the LLM generates multiple responses. We cluster responses based on semantic equivalence to obtain observed semantic distributions. Simultaneously, we compute Semantic Energy from raw logits to capture the model's internal confusion. The core GLIB mechanism uses the Semantic Energy to dynamically calibrate the prior weight (α_g) of a Ghost Cluster, representing unobserved semantics. This Bayesian integration allows the framework to detect consistent hallucinations. An adaptive sampling strategy terminates generation when the posterior entropy variance converges, ensuring efficiency.

where α_{base} smooths observed clusters and α_{g} is a dynamic prior for the ghost mass. The ghost count is always zero ($n_{\text{g}} = 0$), so the posterior parameters are $(\alpha_{\text{base}} + n_1, \ldots, \alpha_{\text{base}} + n_K, \alpha_{\text{g}})$. Posterior expectations smooth the empirical frequencies:

$$\mathbb{E}[\pi_k \mid Y] = \frac{n_k + \alpha_{\text{base}}}{N + K\alpha_{\text{base}} + \alpha_{\text{g}}}. \tag{3}$$

The ghost count is always zero, so π_{g} is entirely prior-driven. Uncertainty is measured by the posterior expected entropy:

$$H_{\text{GLIB}}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\pi}\sim P(\cdot|Y)}\Big[-\sum_{c\in\{c_1,\ldots,c_K,c_{\text{g}}\}} \pi_c \log \pi_c\Big], \tag{4}$$

approximated via Monte Carlo samples from the Dirichlet posterior.

3.2.2 Energy-Guided Prior

A static ghost cluster prior α_{g} implies a constant belief about the unknown, ignoring the model's varying internal confidence. We propose to dynamically modulate this prior by interpreting **Semantic Energy** (Ma et al., 2025) as a proxy for *epistemic uncertainty* in the latent space.

Theoretical Justification. Standard Semantic Entropy operates in the discrete output space $\mathcal{Y}$, which is often sparse. In contrast, the model's logits $\mathbf{z}$ lie in a continuous latent space. High Semantic Energy, characterized by a flat logit distribution (high entropy in $\mathbf{z}$), indicates that the model's internal state is close to the decision boundary or Out-of-Distribution (OOD) regions. From a Bayesian perspective, when the model's internal state is ambiguous (High Energy), our prior belief in the observed semantic distribution should decrease, and the probability mass assigned to the "unknown" (Ghost Cluster) should increase. Thus, we formulate the ghost prior α_{g} not as a fixed hyperparameter, but as a function of the internal energy state $E(\mathbf{x})$.

Energy Calculation. For a generated sequence $\mathbf{y}$ of length L, we compute the *sequence-level energy* $E(\mathbf{y})$ as the length-normalized token-level negative log-likelihood under p_θ (approximating the free energy while normalizing for length) (Liu et al., 2020):

$$E(\mathbf{y}) = -\frac{1}{L}\sum_{t=1}^{L} \log p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x}). \tag{5}$$

We then average across the N sampled sequences to obtain the *query-level energy* $\bar{E}(\mathbf{x})$, which serves as a proxy for global internal confusion. All experiments and the Appendix A.2 pseudocode use Eq. (5), implemented from token log-probabilities; a raw negative-mean-logit proxy is not used (it differs by the log-sum-exp softmax shift and is not a monotone transform of Eq. (5)).

Dynamic Prior Modulation. To map the continuous energy $\bar{E}$ to the Dirichlet prior α_{g}, we use a **Soft-Gating Mechanism**:

$$\alpha_{\text{g}}(\bar{E}) = \alpha_{\text{base}} \cdot \left(1 + S_{\max} \cdot \sigma\left(\frac{\bar{E} - \tau}{T}\right)\right), \quad (6)$$

where $\sigma(\cdot)$ is the sigmoid function.

- **Physical Interpretation:** This function models a phase transition. When $\bar{E} \ll \tau$ (High Confidence), the term $\sigma(\cdot) \to 0$, and α_{g} collapses to α_{base}, reducing our method to standard Bayesian SE. When $\bar{E} \gg \tau$ (High Uncertainty/Hallucination), $\sigma(\cdot) \to 1$, and α_{g} is amplified by $S_{\max}$, effectively forcing the posterior distribution to flatten and shifting mass to the ghost cluster.
- **Role of Parameters:** $S_{\max}$ defines the maximum penalty for internal confusion, while T (temperature) controls the sensitivity of the transition.

Unsupervised Self-Calibration via GMM. To set the energy threshold τ without labels, we fit a two-component Gaussian Mixture Model (GMM) on historical energy scores:

$$p(E) = w_1\mathcal{N}(\mu_1, \sigma_1^2) + w_2\mathcal{N}(\mu_2, \sigma_2^2). \quad (7)$$

We select τ at the intersection of the two Gaussians (where $P(C_1 \mid E) = P(C_2 \mid E)$), enabling model-specific, label-free calibration of what constitutes "High Energy". A full-step pseudocode description of GLIB-SE is provided in Appendix A.2.

3.3 Variance-Based Adaptive Sampling

Computing SE typically requires a fixed, high number of samples N. We propose an adaptive strategy that stops generation when the uncertainty estimate stabilizes, measured by the variance of the posterior entropy:

$$\mathbb{V}_{\text{H}} = \text{Var}_{\boldsymbol{\pi} \sim P(\boldsymbol{\pi} \mid Y_n)}\left[H(\boldsymbol{\pi})\right]. \quad (8)$$

In practice we draw S posterior samples $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{S}$ from the Dirichlet posterior to estimate

$$\hat{\mathbb{V}}_{\text{H}}(n) = \frac{1}{S-1}\sum_{s=1}^{S}\left(H(\boldsymbol{\pi}^{(s)}) - \frac{1}{S}\sum_{j=1}^{S}H(\boldsymbol{\pi}^{(j)})\right)^2, \quad (9)$$

and iteratively generate samples until $\hat{\mathbb{V}}_{\text{H}}(n) < \epsilon$ or a budget $N_{\max}$ is reached. This allocates more compute to ambiguous queries while saving resources on simple ones.

4 Experiments

To comprehensively evaluate the effectiveness of the GLIB-SE framework, we conducted extensive experiments across six benchmark datasets covering diverse task types, including short-form question answering, long-form text generation, and mathematical reasoning. The experiments aim to answer three core research questions: **(1)** Does GLIB-SE outperform existing state-of-the-art methods in detecting LLM hallucinations, particularly in scenarios prone to consistency errors? **(2)** Can the introduced adaptive sampling strategy significantly reduce inference costs while maintaining detection performance? **(3)** Does the logit-based energy guidance mechanism effectively address the failure of semantic entropy in consistent hallucination scenarios?

4.1 Experimental Setup

Datasets and Models. Our evaluation encompasses diverse task scenarios to ensure generalizability. For short-form QA, we used **TriviaQA** (Joshi et al., 2017) and **NQ-Open** (Kwiatkowski et al., 2019); for long-form generation, we selected **BioASQ** (Tsatsaronis et al., 2015) and **CoQA** (Reddy et al., 2019); for mathematical reasoning, we employed **GSM8K** (Cobbe et al., 2021) and **SVAMP** (Patel et al., 2021). We selected mainstream LLMs covering different parameter scales and architectures, including lightweight models (Qwen3-4B-Instruct, Llama-3-8B-Instruct), mid-scale models (Llama-2-13B-Chat), and high-performance models (Llama-3.1-70B-Instruct, Qwen3-32B, Qwen3-30B-A3B, Phi-4) (Touvron et al., 2023; Dubey et al., 2024; Yang et al., 2024; Abdin et al., 2024).

Baselines and Metrics. We compared GLIB-SE against a range of mainstream uncertainty estimation methods: (1) **Token-level**: Naive Entropy and P(True); (2) **Semantic-level**: Semantic Entropy

Models	P(True)		Semantic Entropy		Semantic Energy		KLE		GLIB-SE (Ours)		
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	Avg. Samples
TriviaQA											
Llama-3-8B	0.791	0.702	0.864	0.810	0.832	0.755	**0.887**	**0.830**	0.865	0.815	6.07
Llama-2-13B	0.594	0.452	0.825	0.731	0.809	0.715	0.825	0.728	**0.832**	**0.740**	6.78
Qwen3-8B	**0.872**	**0.798**	0.837	0.765	0.800	0.712	0.848	0.774	0.856	0.782	7.87
Qwen3-4B	**0.883**	**0.812**	0.851	0.795	0.851	0.788	0.868	0.801	0.863	0.805	6.92
NQ-Open											
Llama-3-8B	0.761	0.642	0.713	0.605	0.759	0.655	**0.777**	**0.670**	0.762	0.658	7.69
Llama-2-13B	0.627	0.510	0.739	0.615	0.701	0.588	**0.758**	**0.625**	0.724	0.608	6.11
Qwen3-8B	**0.781**	**0.665**	0.779	0.660	0.670	0.535	0.500	0.415	0.761	0.652	6.57
Qwen3-4B	0.769	0.654	**0.780**	**0.672**	0.657	0.512	0.500	0.415	0.754	0.648	6.79
BioASQ											
Llama-3.1-70B	0.758	0.685	0.816	0.752	0.827	0.760	0.500	0.450	**0.844**	**0.775**	5.62
Seed-Coder-8B	0.672	0.590	0.824	0.765	0.802	0.730	0.500	0.450	**0.847**	**0.788**	8.48
Qwen3-30B-A3B	0.640	0.550	0.743	0.680	0.757	0.695	0.500	0.450	**0.774**	**0.712**	4.38
Qwen3-32B	0.500	0.450	0.773	0.710	0.794	0.725	0.500	0.450	**0.810**	**0.745**	4.57
CoQA											
Llama-3.1-70B	0.798	0.912	0.806	0.909	0.825	**0.923**	0.500	0.752	**0.836**	0.922	8.32
Qwen3-30B-A3B	0.739	0.889	**0.835**	0.932	0.790	0.907	0.500	0.762	0.828	**0.938**	6.56
Qwen3-32B	0.500	0.420	0.815	0.920	0.809	0.903	0.500	0.756	**0.831**	**0.936**	7.33
GSM8K											
Llama-3.1-70B	0.614	0.530	**0.831**	**0.765**	0.761	0.690	0.500	0.420	0.819	0.758	7.38
Phi-4	0.496	0.420	0.689	0.610	0.554	0.480	0.500	0.420	**0.692**	**0.615**	9.41
Qwen3-32B	0.500	0.420	**0.781**	**0.720**	0.656	0.580	0.500	0.420	0.724	0.655	5.25
SVAMP											
Llama-3.1-70B	0.583	0.505	0.840	0.775	0.878	0.815	0.856	0.795	**0.901**	**0.845**	8.25
Llama-2-70B	0.480	0.410	0.849	0.785	0.844	0.780	**0.850**	**0.790**	0.801	0.740	5.48
Phi-4	0.563	0.485	0.870	0.810	0.891	0.835	0.891	0.835	**0.909**	**0.855**	6.73

Table 1: **Hallucination detection: AUROC/AUPR vs. baselines.** *(textbfBold means per-row best; Underline means per-row second-best. All the Non-GLIB methods use a fixed 10 samples; GLIB-SE extraly reports its adaptive average.)*

(SE) (Farquhar et al., 2024) and Distributed Semantic Entropy (DSE); (3) **Logit/Energy-based**: Semantic Energy (SEn) (Ma et al., 2025); (4) **Kernel-based**: Kernel Language Entropy (KLE) (Nikitin et al., 2024). To verify the unique contribution of the energy guidance mechanism, we also included **BayesSE** (Bayesian Semantic Entropy without energy guidance) as an ablation baseline. All experiments use two core metrics for hallucination detection: **Area Under the Receiver Operating Characteristic (AUROC)**, which measures threshold-agnostic ranking quality, and **Area Under the Precision–Recall Curve (AUPR)**, which captures detector precision under class imbalance typical of hallucination vs. truthful responses.

Implementation Details. For semantic equivalence judgment, we used the DeBERTa-large-mnli model (He et al., 2021) for bi-directional entailment checking in short-form tasks. In long-form and reasoning tasks, we employed GPT-4o as a judge model to ensure accurate semantic clustering. In GLIB-SE, the base prior α_{base} was set to 0.1. The energy threshold τ was automatically determined using the GMM-based self-calibration in Section 3.2.2, fitted on 50 unlabelled validation samples; other hyperparameters followed the same global setting across tasks.

4.2 Main Results: Hallucination Detection

Table 1 summarizes AUROC/AUPR across baselines and GLIB-SE with their sampling costs.

Solving Consistent Hallucinations (SVAMP & GSM8K). The most significant improvements are observed in mathematical reasoning tasks, which are notoriously prone to consistent hallucinations (where models confidently repeat the same wrong calculation). On **SVAMP**, GLIB-SE outperforms the standard Semantic Entropy (SE) by a large margin on several models. For instance, on **Llama-3.1-70B-Instruct**, GLIB-SE achieves an AUROC of **0.901**, compared to 0.840 for SE (+6.1%). Similarly, on **Phi-4**, it improves from 0.870 (SE) to **0.909**. This validates our core hypothesis: when

Dataset	Model	BayesSE	GLIB-SE
TriviaQA	Qwen3-4B-Instruct	0.844 / 6.92	**0.863 / 6.92**
	Qwen3-8B	0.846 / 7.87	**0.856 / 7.87**
	Llama-2-13B-chat-hf	0.817 / 6.78	**0.832 / 6.78**
	Llama-3-8B-Instruct	0.853 / 6.07	**0.865 / 6.07**
NQ-Open	Qwen3-4B-Instruct	0.749 / 6.79	**0.754 / 6.79**
	Qwen3-8B	0.766 / 6.57	**0.761 / 6.57**
	Llama-2-13B-chat-hf	0.709 / 6.11	**0.724 / 6.11**
	Llama-3-8B-Instruct	0.747 / 7.69	**0.762 / 7.69**
BioASQ	Llama-3.1-70B-Instruct	0.807 / 5.62	**0.844 / 5.62**
	Seed-Coder-8B-Instruct	0.845 / 8.48	**0.847 / 8.48**
	Qwen3-32B	0.745 / 4.57	**0.810 / 4.57**
	Qwen3-30B-A3B	0.700 / 4.38	**0.774 / 4.38**
CoQA	Qwen3-30B-A3B	0.794 / 6.56	**0.828 / 6.56**
	Qwen3-32B	0.809 / 7.33	**0.831 / 7.33**
	Llama-3.1-70B-Instruct	0.834 / 8.32	**0.836 / 8.32**
GSM8K	Phi-4	0.682 / 9.41	**0.692 / 9.41**
	Qwen3-32B	0.715 / 5.25	**0.724 / 5.25**
	Llama-3-70B	0.801 / 7.38	**0.819 / 7.38**
SVAMP	Phi-4	0.882 / 6.73	**0.909 / 6.73**
	Llama-2-70B-chat-hf	0.788 / 5.48	**0.801 / 5.48**
	Llama-3.1-70B-Instruct	0.864 / 8.25	**0.901 / 8.25**

Table 2: **Ablation: BayesSE vs. GLIB-SE** (AUROC / Avg. Samples).

models are "confidently wrong" (producing consistent outputs but with flat logits), the high Semantic Energy successfully triggers the Ghost Cluster to inflate uncertainty, whereas standard SE collapses to zero.

Robustness on Long-Form Generation (BioASQ). On long-form QA, GLIB-SE consistently surpasses SE. For **Qwen3-32B** on BioASQ, GLIB-SE improves AUROC from 0.773 (SE) to **0.810** (+3.7%).

Short-form factoid tasks (TriviaQA). On TriviaQA the gains are modest because errors are typically either clearly correct or clearly wrong with divergent semantics, where standard SE already issues high entropy. GLIB-SE matches strong baselines here, indicating the Ghost Cluster primarily activates in regimes with consistent-but-wrong generations rather than single-shot factoid mistakes.

Failure Mode Analysis of Baselines. We observe distinct failure modes in existing methods, highlighting the robustness of our approach:

- **KLE Collapse:** On complex tasks like BioASQ and GSM8K, KLE performance degenerates to near-random guessing (AUROC ≈ 0.50). This is attributed to the use of discrete judge models (e.g., GPT-4o) which output hard entailment labels. This collapses KLE's continuous semantic kernel into a binary matrix, stripping the method of its ability to measure fine-grained semantic distance (Nikitin et al., 2024). GLIB-SE avoids this by operating on the semantic distribution level rather than pairwise kernels.
- **P(True) Miscalibration:** Simple probability-based metrics fail on reasoning tasks with smaller models (e.g., Phi-4 on GSM8K, AUROC 0.49). This reflects severe *miscalibration*, where the model places high confidence on incorrect reasoning paths. GLIB-SE mitigates this by introducing the Ghost Cluster, which effectively acts as a "calibration buffer" triggered by the internal energy state.

4.3 Efficiency Analysis: Pareto Efficiency

Inference cost is a critical bottleneck for semantic uncertainty methods, which typically require sampling $N = 5$ to 10 sequences. GLIB-SE achieves a superior Pareto frontier by dynamically allocating the sample budget.

Cost-Performance Ratio. As shown in the last column of Table 1 and Figure 3, GLIB-SE maintains SOTA detection performance with significantly fewer sampled sequences. For instance, on **BioASQ** (Qwen3-32B), it requires only **4.57 sampled sequences** on average to surpass the performance of fixed-sample SE ($N = 10$). This represents a **54.3% reduction** in computational overhead. Similarly, on **TriviaQA**, GLIB-SE converges in $\sim$6 sampled sequences.

Adaptive Mechanism. This efficiency stems from our variance-based stopping criterion. The model exits early (in 2–3 samples) for unambiguous queries where the posterior $\mathbb{V}[H]$ vanishes quickly, while reserving the budget for high-entropy, ambiguous queries. Figure 3 illustrates this dominance, where GLIB-SE's curve lies strictly to the upper-left (better utility, lower cost) of the fixed-N baselines.

4.4 Mechanism Analysis

Ablation Study. Table 2 details the BayesSE vs. GLIB-SE ablation. On **SVAMP (Llama-3.1-70B)**, adding energy guidance improves AUROC from 0.864 (BayesSE) to 0.901 (GLIB-SE). This confirms that the performance gain is not solely due to the Bayesian formulation but specifically stems from the *Logit-Informed Dynamic Prior*, which correctly up-weights the Ghost Cluster when internal confidence is low.

Visualizing the Correction. As visualized in our Density-Entropy analysis (Figure 4), standard SE fails for samples in the "low entropy, high energy" region (red dots, representing consistent hallucinations). GLIB-SE utilizes the vertical axis (energy) to distinguish these from true correct an-

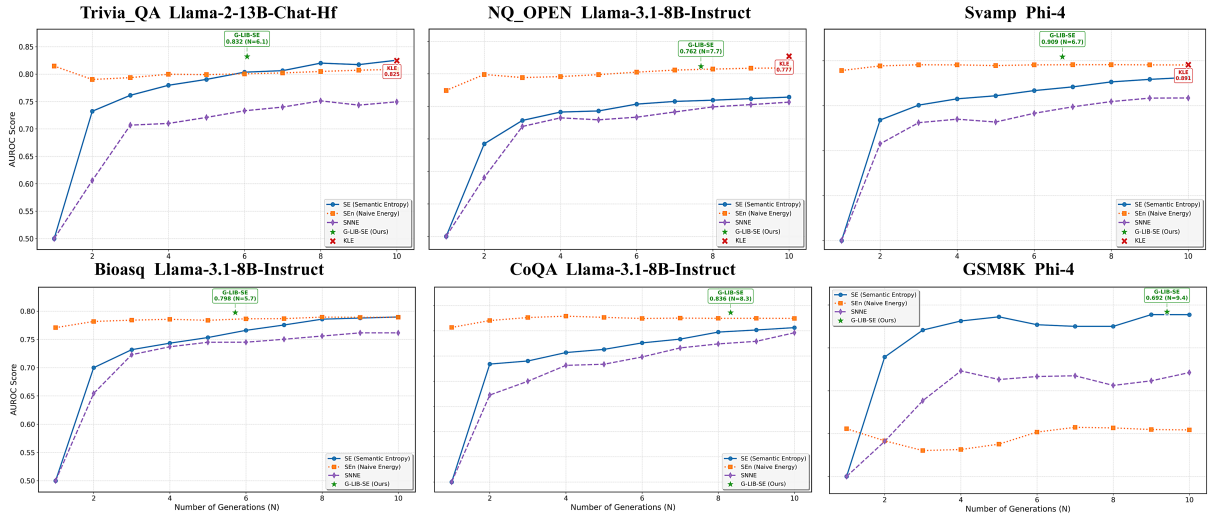

Figure 3: AUROC versus sampled sequences for SE, SEn, SNNE, KLE, and GLIB-SE across all datasets. GLIB-SE curves consistently sit toward the upper-left region of the Pareto frontier, indicating that our method dynamically balances cost and effectiveness to achieve stronger performance with fewer samples.

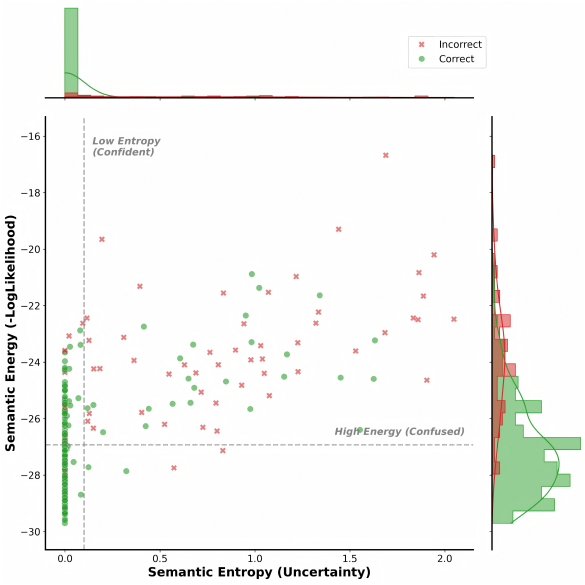

Figure 4: Semantic entropy (x-axis) versus semantic energy (y-axis) for correct (green) and incorrect (red) answers. Points concentrated in the lower-left region (low entropy, low energy) correspond to correct, non-hallucinatory responses, while incorrect answers exhibit higher entropy or energy.

swers (green dots), effectively "pulling" them into the high-uncertainty regime.

5 Conclusion

In this paper, we presented **GLIB-SE**, a probabilistic framework that synergizes internal model confidence (Logit Energy) with external semantic consistency (Semantic Entropy) for robust hallucination detection. The core innovation lies in modeling the unobserved semantic space via a dynamically weighted "Ghost Cluster", which effectively resolves the critical failure mode of standard semantic entropy under "consistent hallucinations"—scenarios where models are confident yet factually incorrect. Furthermore, our variance-based adaptive sampling strategy successfully decouples uncertainty estimation from fixed computational budgets, establishing a superior Pareto frontier between detection accuracy and inference efficiency on reasoning and generation tasks.

6 Future Work

Despite its state-of-the-art performance, GLIB-SE currently relies on white-box access to output logits, which limits its deployment in certain environments. Future research will prioritize two avenues to broaden its applicability:

- **Black-box Adaptation**: We aim to explore proxy metrics for semantic energy, such as verbalized confidence or perplexity approximations, to extend the framework to closed-source API models.
- **Multimodal Expansion**: We plan to investigate the correlation between "internal confusion" and "semantic consistency" in Vision-Language Models (VLMs), thereby adapting the Ghost Cluster mechanism to address hallucinations in image captioning and visual reasoning domains.

Limitations

While GLIB-SE demonstrates robust performance, it has three primary limitations:

1. White-box dependency. The approach relies on access to output logits to compute Semantic Energy. Closed-source API-only models that hide log-probabilities cannot be directly covered unless they expose token-level scores.

2. NLI bottleneck. Semantic clustering depends on an NLI model (e.g., DeBERTa or GPT-4o). For extremely high-throughput deployments, the latency from pairwise NLI checks can dominate total runtime, motivating lighter-weight semantic judges.

Ethics Statement

This work focuses on improving the reliability and safety of Large Language Models by detecting hallucinations.

Data Usage and Privacy. All datasets used in this study (TriviaQA, NQ-Open, BioASQ, CoQA, GSM8K, SVAMP) are publicly available benchmarks widely used in the NLP community. Our experiments do not involve the collection of new personal data or crowdsourcing that would require IRB approval. We have adhered to the intended use licenses of these datasets and the open-weights models (Llama, Qwen, Phi) employed.

Potential Risks. While GLIB-SE significantly improves hallucination detection, it is not infallible. We caution against deploying this method as the sole arbiter of truth in high-stakes domains (e.g., medical diagnosis or legal advice) without human-in-the-loop verification. False negatives remain a possibility, aligning with theoretical findings on the impossibility of perfect automated detection without external ground truth.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awad, Nguyen Awad, Nguyen Bach, Amit Bahree, Arash Bakhtiari, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. 2025. Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy. *Transactions on Machine Learning Research*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letak, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Yuxin Huang, Rui Li, Liu Zhang, S Yan, and 1 others. 2023. A survey on uncertainty quantification in large language models. *arXiv preprint arXiv:2311.12324*.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30.

Jiuck Li, A. Magesh, and V. V. Veeravalli. 2025. Principled detection of hallucinations in large language models via multiple testing. *arXiv preprint arXiv:2508.18473*.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. 2025. Semantic energy: Detecting LLM hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambitious and rubbish! Revisiting the evaluation of open domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6015–6024.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. *arXiv preprint arXiv:2405.20003*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2080–2094.

E. Quevedo Caballero, A. S. Abdelfattah, A. Rodriguez, and T. Černý. 2024. Detecting hallucinations in large language model generation: A token probability approach. *arXiv preprint arXiv:2405.19648*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:1–28.

Xinyu Wang, Yifan Yan, Lifu Huang, Xipeng Zheng, and Xuanjing Huang. 2023. Hallucination detection for generative large language models by Bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15361–15371.

Fangyuan Xu, Xiting Hu, Ziyi Yu, L. Lin, X. Zhang, Y. Zhang, W. Zhou, J. Gu, and Xiaojun Wan. 2025. HAD: HAllucination detection language models based on a comprehensive hallucination taxonomy. *arXiv preprint arXiv:2510.19318*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Appendix

A.1 Implementation Details

Hyperparameters. We set the base Dirichlet prior $\alpha_{\text{base}} = 0.1$, scaling factor $S_{\max} = 100$, and temperature $T = 0.1$. The energy threshold τ is determined via the GMM self-calibration in Section 3.2.2 on 50 unlabeled queries per model.

Adaptive sampling. The posterior variance threshold is $\epsilon = 1\text{e-}4$ with a maximum sample budget $N_{\max} = 10$.

Compute resources. Experiments use NVIDIA H200 (144GB) GPUs. Semantic equivalence uses `microsoft/deberta-large-mnli` for short-form tasks and GPT-4o for long-form or reasoning tasks.

A.2 GLIB-SE Pseudocode

Algorithm 1: GLIB-SE: Energy-Guided Ghost-Cluster Semantic Entropy

Input: Query $\mathbf{x}$, budget $N_{\max}$, variance threshold γ, base prior α_{base}, mapping params Θ
Output: Uncertainty scores H_{glib}, H_{base}; sample count n

Initialize $n \leftarrow 0$, semantic clusters $\mathcal{C} \leftarrow \emptyset$, energy history $\mathcal{E} \leftarrow \emptyset$.
while $n < N_{\max}$ **do**
  $n \leftarrow n + 1$.
  Sample response $\mathbf{y}_n \sim p(\mathbf{y}|\mathbf{x})$; compute energy $E_n \leftarrow -\frac{1}{L}\sum_{t=1}^{L} \log p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x})$ (Eq. (5)).
  Update clusters $\mathcal{C}$ and counts $\mathbf{n}$ with $\mathbf{y}_n$ via bi-directional entailment.
  Update query energy $\bar{E} \leftarrow \text{mean}(\mathcal{E} \cup \{E_n\})$.
  Compute $\alpha_g \leftarrow \alpha_{\text{base}}\big(1 + S_{\max} \cdot \sigma(\frac{\bar{E} - \text{offset}}{\text{temp}})\big)$ via the chosen mapping $f(\cdot)$.
  Form posterior $\boldsymbol{\alpha} \leftarrow [\mathbf{n} + \alpha_{\text{base}}, \alpha_g]$; estimate H_{glib} and variance $\hat{\mathbb{V}}_H$ with MC samples from $\text{Dir}(\boldsymbol{\alpha})$.
  if $\hat{\mathbb{V}}_H < \gamma$ **then**
    break
Compute baseline H_{base} using posterior $[\mathbf{n} + \alpha_{\text{base}}]$ (no ghost).
return H_{glib}, H_{base}, n.

A.3 Adaptive Sampling Distribution

Figure 5 visualizes the empirical distribution of sample counts that GLIB-SE uses across datasets and models.

A.4 Sensitivity to Max Scale and Prior Slope

Interpretation. Figures 6 and 7 summarize a grid search over the ghost prior mapping, varying the maximum scale $S_{\max}$ and the sigmoid sensitivity σ (slope). Brighter cells indicate higher AUROC; performance is stable across a broad region, with mild gains when $S_{\max} \in [100, 500]$ and $\sigma \in [0.1, 5.0]$. This suggests the method is robust to moderate mis-specification of these hyperparameters.

Figure 5: Adaptive sampling histogram for GLIB-SE across datasets and models. Bars indicate the fraction of queries that stop at each sample count N; the red dashed line denotes the average samples used. The distribution shows aggressive early stopping on easy cases and budget allocation on harder queries.

Figure 6: Heatmap analysis (part 1) of AUROC under different ghost prior settings.

Figure 7: Heatmap analysis (part 2) of AUROC under different ghost prior settings. Rows sweep the sigmoid sensitivity σ (how sharply the prior reacts to logit energy), columns sweep the maximum scale $S_{\max}$ (upper bound on ghost mass inflation). The large yellow plateau shows GLIB-SE maintains strong performance over a wide hyperparameter range.