FLOW-ABSTAIN: Modeling the Answerable Manifold for Query-Aware Abstention in Large Language Models

Jixin Zhang and Bensheng Qiu*
University of Science and Technology of China
zjx24006113@mail.ustc.edu.cn, bqiu@ustc.edu.cn

Abstract

Large language models (LLMs) frequently hallucinate when confronted with queries they cannot answer reliably. The root causes are multifaceted, including entity rarity, semantic ambiguity, temporal invalidity, false presuppositions, and other forms of epistemic uncertainty. Existing abstention mechanisms overwhelmingly rely on binary probes trained with manually curated "unanswerable" labels or calibrated confidence scores, both of which presuppose access to negative examples that are expensive to collect and fundamentally incomplete in coverage. We propose FLOW-ABSTAIN, an unsupervised framework that reframes query-aware abstention as density estimation in the model's own latent space. Our method operates in three stages: (1) extracting intermediate hidden-state representations from a curated set of *confidently answerable* queries, (2) applying PCA-based dimensionality reduction to obtain a compact manifold, and (3) fitting a Masked Autoregressive Flow (MAF) to the resulting representations, yielding an exact, tractable log-likelihood for arbitrary new inputs. At inference time, queries whose latent projections receive low log-likelihood under the learned density are flagged for abstention—without requiring any negative or out-of-distribution training samples. Experiments across three instruction-tuned LLMs (Qwen3-8B, Qwen3-14B, Llama3-8B-Instruct) and five diverse evaluation benchmarks demonstrate that FLOW-ABSTAIN achieves highly competitive abstention performance among methods with practical inference cost while providing a continuous, interpretable likelihood signal that naturally stratifies query difficulty along cognitive dimensions of unanswerability.[1]

1 Introduction

The deployment of large language models (LLMs) in knowledge-intensive applications—from open-domain question answering to medical consultation and legal analysis—has amplified the urgency of the *selective prediction* problem: determining when a model should abstain from answering rather than risk generating a plausible but factually incorrect response (Kadavath et al., 2022; Varshney et al., 2023). Despite remarkable advances in instruction tuning and reinforcement learning from human feedback, contemporary LLMs remain prone to *hallucination*—the confident generation of statements unsupported by their training data or the provided context (Ji et al., 2023; Zhang et al., 2025).

A critical yet underexplored observation is that the queries on which LLMs fail do not constitute a monolithic "unanswerable" category (Cohen et al., 2025). Rather, they span a *continuous spectrum of cognitive out-of-distribution (OOD) phenomena*, including entity rarity, where the model lacks sufficient parametric knowledge about obscure entities (Mallen et al., 2023); inherent ambiguity, where a query admits multiple valid interpretations and no single definitive answer exists (Min et al., 2020); false presuppositions, where the question embeds factually incorrect premises that render any direct answer misleading (Kim et al., 2023); and temporal invalidity, where the correct answer has changed since the model's knowledge cutoff (Liska et al., 2022). We argue that these diverse failure modes share a common signature in the model's internal representation space: queries the model *cannot* reliably answer induce latent activations that deviate systematically from the manifold of confidently processed inputs. This perspective motivates a fundamentally different approach to abstention—one grounded in *density estimation* rather than discriminative classification.

Current abstention methods can be broadly divided into two categories. Probe-based approaches train linear or shallow classifiers on internal representations to distinguish answerable from unanswerable queries (Azaria and Mitchell, 2023; Burns

* Corresponding author.
[1] https://github.com/Mike-ihr/FlowAbstain.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 13556–13566
October 24-29, 2026 ©2026 Association for Computational Linguistics

et al., 2023; Lavi et al., 2026). However, these methods rely on curated negative examples, which are inevitably incomplete, as no finite collection of "unanswerable" queries can exhaust the combinatorial diversity of possible failure modes. Output-based approaches estimate uncertainty using signals such as token probabilities, semantic entropy, and self-consistency (Kuhn et al., 2023; Manakul et al., 2023). While effective in some settings, they operate on the generated outputs rather than the latent representations where the model's epistemic state is most directly encoded.

To address these limitations, we propose FLOWABSTAIN, a fully unsupervised framework for query-aware abstention. Our method first extracts intermediate hidden-state representations from a reference corpus of queries that the model answers correctly and confidently. These representations are then projected into a compact and denoised subspace using Principal Component Analysis (PCA), which serves as an estimate of the manifold of answerable queries. Finally, we fit a Masked Autoregressive Flow (MAF) (Papamakarios et al., 2017) to model the exact probability density of this manifold. At inference time, the log-likelihood assigned to a new query's latent representation provides a continuous abstention score: queries lying in low-density regions are identified as cognitively out-of-distribution and are therefore candidates for abstention, with lower likelihoods indicating greater epistemic uncertainty. Our contributions are summarized as follows:

1. We formalize query unanswerability as a *continuous density phenomenon* in latent space, unifying disparate failure modes (e.g., rarity, ambiguity, false premise and outdated knowledge) under a single probabilistic framework.
2. We demonstrate that a Masked Autoregressive Flow trained exclusively on positive (answerable) examples offers the best trade-off between abstention accuracy and inference efficiency across five benchmarks and three LLM architectures, *without requiring any negative training samples*.
3. We show that the learned log-likelihood provides an interpretable, continuous signal that naturally stratifies queries along dimensions of difficulty, enabling fine-grained abstention policies beyond binary accept/reject decisions.

2 Related Work

2.1 LLM Abstention and Selective Prediction

Selective prediction—the capacity of a model to withhold answers when uncertain—has a rich history in machine learning (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017). In the context of LLMs, abstention mechanisms broadly partition into *internal* and *external* approaches.

Internal methods leverage the model's own outputs or intermediate computations. Kadavath et al. (2022) demonstrate that LLMs can be prompted to self-evaluate their confidence, though this metacognitive signal is often poorly calibrated. Kuhn et al. (2023) propose semantic entropy, aggregating uncertainty across semantically equivalent generations. Manakul et al. (2023) introduce self-consistency checks via multiple sampled responses. Chen et al. (2024) propose EigenScore to measure semantic consistency and diversity in dense embedding spaces. While effective in certain regimes, these output-level methods are computationally expensive at inference time and cannot distinguish *why* a model is uncertain. Others instead fine-tune the model to emit a dedicated refusal token, making abstention part of the decoding rule (Cohen et al., 2024; Huang et al., 2025).

External methods train auxiliary classifiers on model internals. Azaria and Mitchell (2023) and Burns et al. (2023) probe hidden states with linear classifiers to detect truthfulness, but require balanced positive/negative supervision. Slobodkin et al. (2023) explore attention patterns for answerability detection. Our work departs from all of the above by eschewing discriminative classification entirely, instead framing abstention as one-class density estimation that requires only positive (answerable) examples.

2.2 Latent Space Probing in LLMs

The practice of probing neural network hidden states to extract linguistic or factual information has become a standard analytical tool (Belinkov, 2022). Geva et al. (2021) show that feed-forward layers in transformers encode factual associations, while Meng et al. (2022) demonstrate that specific layers store retrievable knowledge. Li et al. (2023) investigate how representation geometry correlates with model confidence, finding that correctly answered queries cluster distinctly from incorrect ones.

These findings motivate our choice to operate on intermediate hidden states: if the model's epistemic

state is geometrically encoded, then density estimation over this geometry provides a principled abstention criterion. Unlike probing classifiers, which require labeled failure cases, our approach learns the support of the "confident" distribution and flags deviations from it.

2.3 Normalizing Flows for Out-of-Distribution Detection

Normalizing flows (Rezende and Mohamed, 2015; Dinh et al., 2017) provide exact likelihood computation through invertible transformations, making them natural candidates for OOD detection. Nalisnick et al. (2019) first highlighted that flow-based likelihoods can paradoxically assign higher density to OOD inputs in pixel space, but subsequent work (Kirichenko et al., 2020; Zhang et al., 2021) demonstrated that this pathology is largely resolved when operating on learned feature representations rather than raw inputs.

In the vision domain, Kirichenko et al. (2020) successfully apply normalizing flows to penultimate-layer features for OOD detection. Zisselman and Tamar (2020) extend this to residual flows with typicality scoring. In NLP, flow-based OOD detection remains underexplored. Our work bridges this gap by applying Masked Autoregressive Flows (Papamakarios et al., 2017)—chosen for their expressiveness and exact density computation—to LLM hidden states, establishing the first flow-based abstention framework for generative language models.

3 Methodology

We present FLOW-ABSTAIN, a three-stage pipeline for unsupervised query-aware abstention. The core insight is that confidently answerable queries occupy a structured, learnable manifold in the LLM's latent space, and deviations from this manifold—quantified via exact log-likelihood under a normalizing flow—signal cognitive OOD inputs warranting abstention. Figure 1 provides a schematic overview.

3.1 Stage 1: Hidden State Extraction

Let $\mathcal{M}$ denote a pre-trained instruction-tuned LLM with L transformer layers, and let $\mathcal{Q}_{\text{ref}} = \{q_1, q_2, \ldots, q_N\}$ denote a reference corpus of queries for which the model produces verifiably correct answers with high confidence.

For each query $q_i \in \mathcal{Q}_{\text{ref}}$, we extract the hidden-state vector $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^D$ at a designated intermediate layer ℓ, where D is the model's hidden dimension. Specifically, we take the representation corresponding to the *last token position* of the input query, as this position aggregates contextual information across the full input via causal attention:

$$\mathbf{h}_i^{(\ell)} = \text{Layer}_\ell(\mathcal{M}, q_i)[\texttt{last_token}] \quad (1)$$

The choice of layer ℓ is motivated by prior findings that intermediate layers (typically layers $\frac{L}{3}$ to $\frac{2L}{3}$) encode the richest factual and epistemic information (Geva et al., 2021; Meng et al., 2022). We treat ℓ as a hyperparameter and select it via validation performance.

Reference corpus construction. The reference corpus $\mathcal{Q}_{\text{ref}}$ is constructed by filtering a general-purpose QA dataset (TriviaQA; Joshi et al., 2017) to retain only queries where the model's top-1 generated answer exactly matches the gold answer. This ensures that the extracted representations correspond to states where the model demonstrably possesses and retrieves the relevant knowledge, providing a relatively clean characterization of the "answerable manifold."

3.2 Stage 2: PCA Dimensionality Reduction

The raw hidden-state vectors $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^D$ reside in a high-dimensional space ($D = 4096$ for Qwen3-8B and Llama3-8B-Instruct, $D = 5120$ for Qwen3-14B), much of which encodes information orthogonal to the answerability signal. Direct density estimation in such high-dimensional spaces is both statistically and computationally challenging due to the curse of dimensionality.

We apply Principal Component Analysis (PCA) to project the hidden states onto a lower-dimensional subspace that captures the dominant modes of variation among confidently answered queries:

$$\mathbf{z}_i = W_{\text{PCA}}^\top (\mathbf{h}_i^{(\ell)} - \boldsymbol{\mu}) \in \mathbb{R}^d \quad (2)$$

where $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i^{(\ell)}$ is the empirical mean, $W_{\text{PCA}} \in \mathbb{R}^{D \times d}$ contains the top-d principal components, and $d \ll D$ is chosen to retain at least 95% of the cumulative explained variance.

This reduction serves a dual purpose: (a) it removes noise dimensions that would degrade density estimation, and (b) it concentrates the representation on the subspace most relevant to distinguishing answerable from non-answerable queries. In

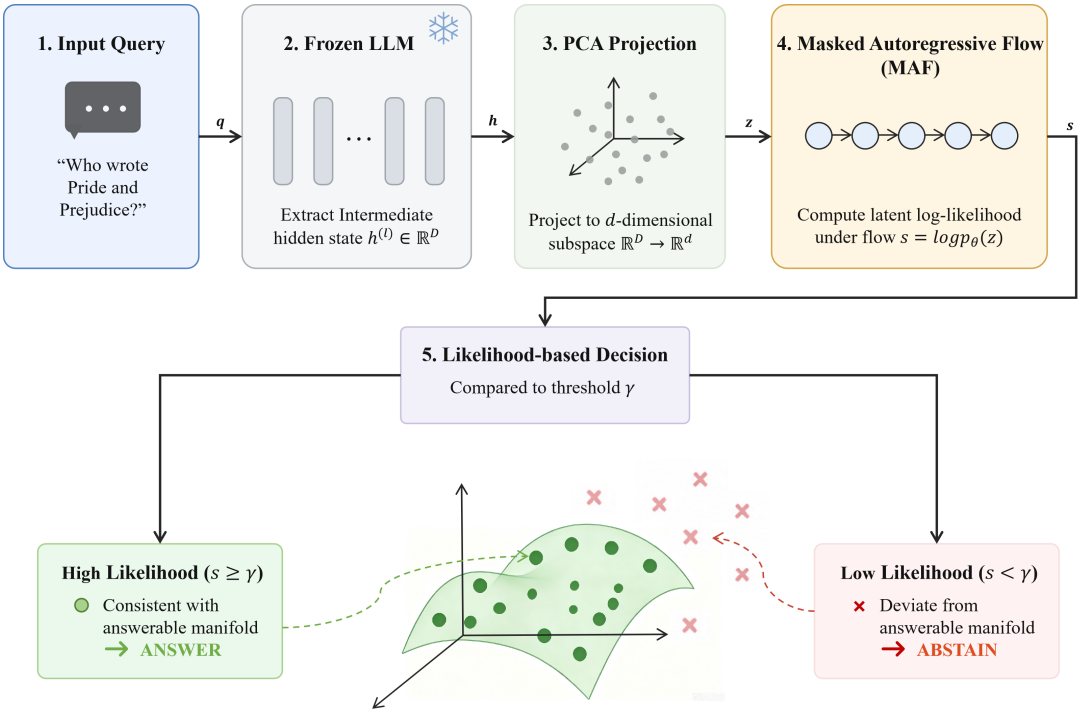


Figure 1: Overview of FLOW-ABSTAIN. Given a query q, we extract the hidden state from layer ℓ of a frozen LLM, project it into a compact latent space using PCA, and estimate its log-likelihood with a Masked Autoregressive Flow. Queries with likelihood below threshold τ are flagged for abstention.

practice, we find $d \in [64, 256]$ sufficient across all models tested.

3.3 Stage 3: Masked Autoregressive Flow

Given the low-dimensional PCA-reduced representations $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\} \subset \mathbb{R}^d$, we fit a Masked Autoregressive Flow (MAF; Papamakarios et al., 2017) to model their density $p_\theta(\mathbf{z})$.

Architecture. A MAF consists of a sequence of K autoregressive transformations, each implemented via a Masked Autoregressive Dense Estimator (MADE; Germain et al., 2015). The k-th transformation maps $\mathbf{z}^{(k-1)}$ to $\mathbf{z}^{(k)}$ via:

$$z_j^{(k)} = z_j^{(k-1)} \cdot \exp(\alpha_j^{(k)}) + \mu_j^{(k)} \tag{3}$$

where $\mu_j^{(k)} = f_\mu^{(k)}(\mathbf{z}_{<j}^{(k-1)})$ and $\alpha_j^{(k)} = f_\alpha^{(k)}(\mathbf{z}_{<j}^{(k-1)})$ are outputs of masked neural networks that depend only on preceding dimensions $\mathbf{z}_{<j}^{(k-1)}$, enforcing the autoregressive property.

After K transformations, the final variable $\mathbf{z}^{(K)}$ is assumed to follow a standard multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The exact log-likelihood of an input $\mathbf{z}$ is computed via the change-of-variables formula:

$$\log p_\theta(\mathbf{z}) = \log \mathcal{N}(\mathbf{z}^{(K)}; \mathbf{0}, \mathbf{I}) + \sum_{k=1}^{K} \sum_{j=1}^{d} \alpha_j^{(k)} \tag{4}$$

where the second term is the log-determinant of the Jacobian, which decomposes into a simple sum of the scale parameters due to the triangular structure of autoregressive transformations.

Training. The MAF is trained by maximizing the log-likelihood over the reference representations:

$$\theta^* = \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathbf{z}_i). \tag{5}$$

Unless otherwise specified, we use an 8-layer MAF with MADE-based autoregressive blocks and optimize the objective using AdamW. Detailed hyperparameters are provided in Section 4.

3.4 Inference: Log-Likelihood Scoring

At inference time, given a new query q_{new}, the abstention decision proceeds as follows:

1. Extract the hidden state: $\mathbf{h}_{\text{new}}^{(\ell)} = \text{Layer}_\ell(\mathcal{M}, q_{\text{new}})[\texttt{last_token}]$
2. Project via PCA: $\mathbf{z}_{\text{new}} = W_{\text{PCA}}^\top(\mathbf{h}_{\text{new}}^{(\ell)} - \boldsymbol{\mu})$
3. Compute log-likelihood: $s(q_{\text{new}}) = \log p_{\theta^*}(\mathbf{z}_{\text{new}})$
4. Abstain if $s(q_{\text{new}}) < \tau$

where τ is a threshold calibrated on a small validation set to achieve a desired false-positive rate.

Computational cost. The inference overhead of FLOW-ABSTAIN is minimal. Hidden-state extraction reuses the standard forward pass of the LLM, requiring no additional model evaluation. The PCA projection is a single matrix multiplication, and density evaluation in the MAF operates in a low-dimensional latent space (typically $d \leq 256$). In practice, this additional computation is negligible relative to the cost of autoregressive generation.

4 Experiments

We evaluate FLOW-ABSTAIN across multiple LLM architectures and diverse OOD query benchmarks, addressing three research questions:

- **RQ1**: How effectively does FLOW-ABSTAIN detect queries warranting abstention compared to existing methods?
- **RQ2**: Does the log-likelihood score provide a meaningful continuous signal that stratifies queries along dimensions of cognitive OOD-ness?
- **RQ3**: What is the computational overhead of FLOW-ABSTAIN at inference time?

4.1 Models

We evaluate on three instruction-tuned LLMs spanning different scales and architectural families:

- **Qwen3-8B** (Yang et al., 2025): An 8-billion parameter model from the Qwen3 family with 36 transformer layers and hidden dimension $D = 4096$. We extract representations from the intermediate layer $\ell = 18$.
- **Qwen3-14B** (Yang et al., 2025): A 14-billion parameter model from the Qwen3 family with 40 transformer layers and hidden dimension $D = 5120$. We extract representations from the intermediate layer $\ell = 20$.
- **Llama3-8B-Instruct** (Grattafiori et al., 2024): A decoder-only transformer with 8 billion parameters, 32 layers, and hidden dimension $D = 4096$. We extract hidden states from layer $\ell = 16$.

For all models, we use greedy decoding (temperature $= 0$) to generate answers. Correctness is determined against gold references using either normalized exact match or an LLM-based judge.

4.2 Datasets

In-distribution (reference corpus). We use **TriviaQA** (Joshi et al., 2017) as the source for constructing $\mathcal{Q}_{\text{ref}}$. From the validation split (17,944 questions), we retain queries where the model's greedy-decoded answer exactly matches any of the gold aliases, yielding approximately 8,000–11,000 reference queries per model (exact counts vary by model capability). A held-out subset of the correctly answered queries is reserved for evaluation.

Out-of-distribution evaluation benchmarks. We evaluate abstention performance on four benchmarks, each targeting a distinct dimension of unanswerability:

- **SQuAD 2.0 (Unanswerable)** (Rajpurkar et al., 2018): 5,945 questions deliberately constructed to be unanswerable given their associated context, testing the model's ability to recognize insufficient evidence.
- **PopQA** (Mallen et al., 2023): 14,267 entity-centric questions stratified by Wikipedia page popularity. We focus on the long-tail subset (popularity rank > 1000), where entity rarity induces knowledge gaps.
- **TruthfulQA** (Lin et al., 2022): 817 questions designed to elicit common misconceptions and false beliefs, testing resistance to systematic biases in training data.
- **AmbigQA** (Min et al., 2020): 14,042 questions identified as ambiguous, admitting multiple valid answers depending on interpretation. These test whether the model recognizes inherent query underspecification.

Evaluation metric. For evaluation, we formulate a binary detection task in which correctly answered TriviaQA queries (the held-out subset of $\mathcal{Q}_{\text{ref}}$) serve as the positive class, while incorrectly answered TriviaQA queries together with queries from OOD benchmarks form the negative class. We report AUROC and FPR@95%TPR as the primary metrics. Here FPR@95%TPR is the false positive rate at the threshold that retains 95% of answerable queries, i.e., the fraction of queries warranting abstention that are still answered.

4.3 Baselines

We compare against eight baselines that span every abstention paradigm currently used in the literature, all applied to the same frozen model as FLOW-ABSTAIN. At the *token-probability* level, Perplexity scores the average token-level negative log-likelihood of the generated answer. At the *output level*, three methods aggregate uncertainty over n=10 sampled generations: Semantic Entropy (Kuhn et al., 2023) over semantic clusters, Eigen-Score (Chen et al., 2024) over the covariance of

sentence embeddings, and SAR (Duan et al., 2024) over relevance-weighted token and sentence scores. P(True) (Kadavath et al., 2022) instead elicits *self-verbalized* confidence, reading the probability of "True" when the model judges its own answer. The remaining baselines act on the same hidden states as FLOW-ABSTAIN: Linear Probe (Azaria and Mitchell, 2023) is a *supervised* logistic regression that additionally requires labeled negative examples, whereas Mahalanobis (Lee et al., 2018) and GMM-16 estimate *unsupervised* density—a class-conditional Gaussian and a 16-component mixture over the PCA-reduced representations, the latter ablating the flow while keeping the rest of our pipeline. Training-time methods that modify model weights, such as refusal-token tuning (Cohen et al., 2024; Huang et al., 2025), are not comparable under our frozen-model protocol.

4.4 Implementation Details

The MAF is implemented in PyTorch using the `nflows` library. We use $K = 8$ MADE blocks, each with 2 hidden layers of 512 units. PCA dimensionality is set to $d = 128$ for both 8B- and 14B-scale models. Training uses AdamW with learning rate 10^{-3} and batch size 8. All experiments are conducted on NVIDIA A40 GPUs with 48GB memory. Hidden state extraction for the full TriviaQA validation set requires approximately 0.5 hours for the 8B models and 1 hour for the 14B model. MAF training converges in under 5 minutes on a single GPU.

5 Results and Discussion

5.1 RQ1: Detection Effectiveness

Table 1 presents abstention detection performance across three instruction-tuned LLMs and five evaluation benchmarks. Overall, FLOW-ABSTAIN consistently achieves competitive performance across most settings without requiring labeled negative examples or multiple sampled generations.

The largest gains are observed on benchmarks involving fundamentally unanswerable or cognitively out-of-distribution queries. On SQuAD 2.0, FLOW-ABSTAIN achieves the best overall performance across all three models, reaching AUROC values of 95.9, 95.2, and 94.2 while simultaneously obtaining the lowest FPR@95 values. Similarly, on TruthfulQA, FLOW-ABSTAIN consistently attains the highest AUROC across all architectures, indicating strong robustness to false presuppositions and misconception-style queries.

Compared with the supervised Linear Probe baseline, FLOW-ABSTAIN exhibits substantially stronger cross-benchmark robustness despite operating in a fully unsupervised setting. While Linear Probe occasionally achieves competitive AUROC on TriviaQA and PopQA, its performance collapses on SQuAD 2.0, suggesting limited generalization to unseen forms of unanswerability. In contrast, FLOW-ABSTAIN maintains stable performance across all benchmarks, indicating that density estimation better captures the geometry of the answerable manifold than discriminative classifiers trained on finite negative examples. Although benchmark-specific stylistic differences may partially contribute to separability, the consistent likelihood ordering across semantically distinct failure modes (e.g., ambiguity, adversarial misconceptions, false presupposition, and entity rarity) suggests that the learned density captures a broader notion of answerability-related latent regularity rather than merely exploiting superficial dataset artifacts.

Among unsupervised baselines, Mahalanobis distance performs inconsistently, likely due to its restrictive unimodal Gaussian assumption. GMM-16 achieves competitive AUROC on several benchmarks but consistently exhibits higher FPR@95, suggesting that Gaussian mixtures tend to overassign probability mass around in-distribution regions. Semantic Entropy also performs competitively, but requires 10 sampled generations per query and therefore incurs substantially higher inference latency.

Overall, the performance advantage of FLOW-ABSTAIN remains consistent across all three LLM architectures, indicating that the learned density structure generalizes across both model families and parameter scales.

5.2 RQ2: Continuous Likelihood Spectrum

A key advantage of FLOW-ABSTAIN over binary classification approaches is that the log-likelihood score provides a *continuous* measure of how far a query deviates from the answerable manifold. Figure 2 visualizes the distribution of log-likelihood scores across query categories for Qwen3-8B and reveals three notable patterns. As FLOW-ABSTAIN models a continuous latent density via normalizing flows, positive log-likelihood values may naturally arise.

First, the likelihood distributions exhibit clear separation between in-distribution and out-of-

Table 1: Abstention detection performance across models and evaluation benchmarks. We report AUROC (↑) and FPR@95%TPR (↓), both in percentage (%). Best results in **bold**, second-best underlined.

Model	Method	TriviaQA		SQuAD 2.0		PopQA		TruthfulQA		AmbigQA	
		AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95
Qwen3-8B	Perplexity	44.9	96.2	36.6	96.1	24.1	99.6	36.6	96.1	28.5	99.5
	Semantic Entropy	77.3	71.7	56.7	89.2	75.6	60.1	76.7	58.6	74.9	58.3
	EigenScore	77.9	70.4	60.2	85.7	76.8	61.9	78.4	55.2	76.2	58.1
	SAR	78.3	70.6	62.5	82.1	78.7	**59.8**	80.1	51.7	78.3	**57.9**
	P(True)	78.4	73.9	91.2	36.3	78.6	89.9	90.2	**36.3**	80.7	63.0
	Linear Probe	**80.2**	**67.3**	43.1	86.9	**82.9**	64.6	74.1	81.0	71.9	71.6
	Mahalanobis	64.0	85.4	92.3	38.5	57.5	89.8	86.9	61.2	73.1	81.7
	GMM-16	77.1	71.1	94.4	23.2	74.3	71.3	90.6	53.6	82.4	71.4
	FLOW-ABSTAIN (Ours)	78.5	69.8	**95.9**	**19.1**	80.2	67.6	**91.1**	43.5	**83.8**	65.1
Qwen3-14B	Perplexity	37.5	98.6	32.9	97.4	47.3	98.1	35.0	98.4	47.1	96.4
	Semantic Entropy	70.4	**66.5**	58.5	82.9	75.5	62.1	71.0	67.0	75.5	63.3
	EigenScore	72.6	68.8	62.4	78.5	76.9	**61.4**	73.5	63.1	77.1	63.7
	SAR	74.1	69.7	63.2	74.8	78.6	62.5	75.8	62.4	79.2	63.1
	P(True)	**78.1**	74.9	85.3	39.2	**81.8**	72.5	86.0	**50.7**	**84.7**	64.5
	Linear Probe	77.3	68.9	47.9	88.2	77.2	69.7	76.9	66.4	81.3	63.9
	Mahalanobis	67.7	84.9	92.5	19.1	69.4	81.5	76.8	72.2	64.6	84.0
	GMM-16	74.0	69.0	93.8	24.6	78.9	70.2	87.2	63.0	80.8	65.6
	FLOW-ABSTAIN (Ours)	77.6	68.2	**95.2**	**10.2**	80.3	67.9	**91.9**	54.2	83.6	**62.9**
Llama3-8B	Perplexity	43.8	97.5	45.5	96.9	49.6	93.9	46.4	96.0	47.3	96.4
	Semantic Entropy	72.2	64.5	64.2	78.1	74.9	59.0	73.8	59.2	79.5	56.4
	EigenScore	74.0	65.2	67.9	73.6	76.1	57.2	75.6	55.8	80.7	**54.1**
	SAR	75.2	64.4	70.4	71.5	78.3	57.1	78.2	52.3	**81.4**	55.5
	P(True)	74.8	69.3	69.9	52.3	**85.0**	56.6	80.4	49.9	71.8	80.0
	Linear Probe	**79.0**	67.8	66.7	84.8	76.6	66.9	85.4	57.3	81.0	68.5
	Mahalanobis	63.4	84.5	92.2	29.8	63.1	89.7	70.5	69.4	64.5	86.8
	GMM-16	78.1	68.5	93.9	20.1	81.8	**51.5**	83.5	51.6	81.3	61.1
	FLOW-ABSTAIN (Ours)	77.7	**64.2**	**94.2**	**12.3**	79.1	55.6	**89.4**	**48.3**	80.6	58.5

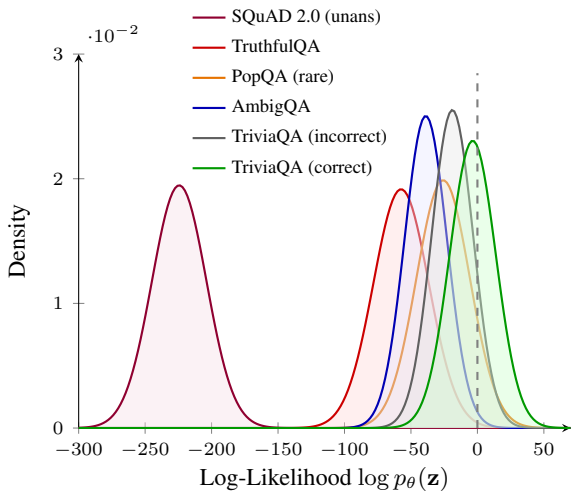


Figure 2: Gaussian fits to empirical log-likelihood scores under FLOW-ABSTAIN for different query categories (Qwen3-8B).

distribution queries. Correctly answered TriviaQA examples are concentrated in a high-likelihood region, while all OOD categories are systematically shifted toward lower values. Second, the OOD categories are naturally stratified along the likelihood axis in a manner that aligns with their intuitive cognitive difficulty. Context-dependent unanswerable questions from SQuAD 2.0 remain farthest from the answerable manifold, followed by adversarial misconception-driven examples from TruthfulQA, ambiguous queries from AmbigQA, and rare-entity questions from PopQA, which occupy the highest-likelihood region among OOD categories. Third, each category exhibits substantial internal variance, indicating that OODness is not binary but exists on a spectrum: even within the same benchmark, some queries lie only marginally outside the answerable manifold, whereas others represent more severe departures.

This continuous uncertainty signal enables more flexible abstention strategies than a single accept-or-reject decision threshold. In practice, systems can route extremely low-likelihood queries to human experts, request clarification for moderately ambiguous inputs, or calibrate the degree of response hedging according to the magnitude of the likelihood deficit.

5.3 RQ3: Inference Efficiency

Table 3 compares the inference-time computational overhead of each method. FLOW-ABSTAIN requires only a single matrix multiplication (PCA projection) and a forward pass through the MAF. This is comparable to the trivial cost of the Mahalanobis baseline, GMM-k, and the linear probe, and

Table 2: **Mean log-likelihood ($\pm$ std) by query category.** Lower values indicate greater deviation from the answerable manifold.

Category	Qwen3-8B	Qwen3-14B	Llama3-8B
TriviaQA (correct)	-3.5 ± 17.3	-2.1 ± 17.8	-1.1 ± 20.0
TriviaQA (incorrect)	-18.8 ± 15.6	-23.1 ± 21.4	-11.3 ± 26.7
PopQA (rare)	-25.8 ± 20.0	-36.0 ± 22.2	-34.7 ± 20.1
AmbigQA	-38.8 ± 15.9	-42.7 ± 16.9	-43.1 ± 27.8
TruthfulQA	-57.4 ± 20.8	-62.4 ± 33.2	-59.6 ± 31.5
SQuAD 2.0 (unans.)	-224.3 ± 20.5	-235.7 ± 91.9	-211.6 ± 40.3

faster than output-based methods such as Semantic Entropy (which requires 10 additional full generations) or P(True) (which requires one additional prompted generation).

Crucially, the hidden-state extraction itself incurs *zero* additional cost, as it is a byproduct of the standard forward pass already required for answer generation. This makes FLOW-ABSTAIN suitable for latency-sensitive production deployments where abstention decisions must be made in real time without incurring additional GPU compute.

Table 3: **Generation and compute savings.** We report the average number of generated tokens and the added latency per query.

Method	Avg. Tokens	Latency (ms)
Perplexity	26.5	0.0
Semantic Entropy	249.7	255.1
EigenScore	249.7	232.6
SAR	249.7	278.4
P(True)	27.5	42.5
Linear Probe	26.5	33.4
Mahalanobis	26.5	32.0
GMM-16	26.5	32.5
FLOW-ABSTAIN (Ours)	**26.5**	**33.7**

5.4 Ablation Studies

We conduct ablations along three axes: PCA dimensionality d, number of MAF blocks K, and extraction layer ℓ. Results (Table 4) confirm that: (a) PCA is essential—without it, density estimation in the full D-dimensional space degrades substantially; (b) $d = 128$ provides the best trade-off between expressiveness and statistical efficiency; (c) $K = 8$ blocks suffice, with diminishing returns beyond this; (d) intermediate layers ($\ell \approx L/2$) consistently outperform early or late layers, consistent with prior findings on factual knowledge localization.

Table 4: **Ablation study on PCA dimensionality, MAF depth, and extraction layer** using Qwen3-8B. We report average AUROC (%) and FPR@95%TPR (%) across QA benchmarks. Default configuration in **bold**.

Component	Setting	Avg. AUROC	Avg. FPR95
PCA (dim d)	64	83.8	57.2
	128	**85.9**	**53.0**
	256	81.3	59.4
	None (4096-D)	77.9	65.2
MAF (blocks K)	2	82.9	58.5
	4	83.7	55.6
	8	**85.9**	**53.0**
	16	86.3	51.1
Layer (ℓ)	12 (early)	82.1	64.8
	18 (mid.)	**85.9**	**53.0**
	24 (late)	85.6	55.1
	36 (final)	72.4	76.1

6 Conclusion

We have presented FLOW-ABSTAIN, an unsupervised framework for query-aware abstention in large language models that reframes the problem as density estimation in the model's latent representation space. By training a Masked Autoregressive Flow exclusively on hidden states from confidently answered queries, our method provides log-likelihood scores that serve as a continuous, interpretable abstention signal—without requiring any manually labeled negative examples or output-level sampling.

Our experiments across three instruction-tuned LLMs and five diverse evaluation benchmarks demonstrate that FLOW-ABSTAIN achieves highly competitive abstention detection performance while adding negligible inference latency. The learned likelihood naturally stratifies queries along a continuous answerability spectrum, unifying disparate failure modes (e.g., entity rarity, ambiguity, false premises, adversarial misconceptions) under a single probabilistic framework.

Beyond binary abstention, the continuous nature of the FLOW-ABSTAIN score opens avenues for graduated response strategies—routing uncertain queries to human review, calibrating hedging language, or triggering retrieval augmentation for borderline cases. Future work will explore extending the framework to multi-turn dialogue settings, improving robustness through layer-wise ensembles, and integrating latent density signals into retrieval-augmented or uncertainty-aware generation pipelines.

Limitations

Our work has several limitations. First, the reference corpus construction assumes access to a QA dataset with gold answers for filtering; in purely open-ended generation settings, alternative confidence criteria (e.g., self-consistency) would be needed. Second, our evaluation focuses primarily on factoid QA benchmarks; whether similar latent density structures emerge in broader generative tasks such as summarization, multi-step reasoning, or creative writing remains unclear. Third, although latent density is strongly correlated with query answerability, low likelihood does not necessarily imply factual unanswerability. Queries involving rare yet answerable knowledge may also lie outside the learned manifold, potentially leading to over-abstention. Finally, since the latent density is learned from correctly answered TriviaQA queries, part of the separability may still reflect dataset-specific stylistic regularities rather than a fully task-agnostic notion of answerability.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms'internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, pages 3056–3076.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [idk] token. *Advances in Neural Information Processing Systems*, 37:10935–10958.

Roi Cohen, Omri Fahn, and Gerard de Melo. 2025. Pretrained llms learn multiple types of uncertainty. *arXiv preprint arXiv:2505.21218*.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using real-NVP. In *International Conference on Learning Representations*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. 2015. MADE: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. 2025. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24564–24579, Vienna, Austria. Association for Computational Linguistics.

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, Thomas Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna

Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2020. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, volume 33.

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*.

Maor Juliet Lavi, Tova Milo, and Mor Geva. 2026. Detecting (un) answerability in large language models with linear directions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 682–699.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hans-Rüdiger Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *ArXiv*, abs/2306.03341.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark 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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5783–5797.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. 2019. Do deep generative models know what they don't know? In *International Conference on Learning Representations*.

George Papamakarios, Theo Pavlakou, and Iain Murray. 2017. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. 2021. Understanding failures in out-of-distribution detection with deep generative models. *International Conference on Machine Learning*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

Ev Zisselman and Aviv Tamar. 2020. Deep residual flow for out of distribution detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13994–14003.