Semantic Self-Distillation for Language Model Uncertainty

Edward Phillips[1] **Sean Wu**[1] **Fredrik K. Gustafsson**[1] **Boyan Gao**[1] **David A. Clifton**[1, 2]

[1]Department of Engineering Science, University of Oxford
[2]Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou

Abstract

Large language models present challenges for principled uncertainty quantification, in part due to their complexity and the diversity of their outputs. Semantic dispersion, or the variance in the meaning of sampled answers, has been proposed as a useful proxy for model uncertainty, but the associated computational cost prohibits its use in latency-critical applications. We show that sampled semantic distributions can be distilled into lightweight student models which estimate a prompt-conditioned density before the language model generates an answer token. The student model predicts a semantic distribution over possible answers; the entropy of this distribution provides a prompt-level uncertainty signal, and the probability density allows answer-level reliability evaluation. Across experiments on TriviaQA and MMLU, we find our student models perform competitively relative to the teacher's sampled semantic dispersion on a hallucination prediction task, whilst offering additional uncertainty primitives for out-of-domain detection and multiple-choice answer selection. We term this technique Semantic Self-Distillation (SSD), which can serve as a general framework for distilling predictive uncertainty in complex output spaces beyond language.

1 INTRODUCTION

Uncertainty quantification (UQ) can be used to assess the reliability of large language model (LLM) outputs [Xiong et al., 2024, Liu et al., 2025]. As LLMs are rapidly deployed in high-stakes domains such as healthcare and law, it becomes increasingly important to have meaningful measures for the trustworthiness of model-generated content. Well-designed UQ methods can provide signals for detecting errors like *hallucinations* - outputs that are misaligned with fact or user intention - and enable mitigation strategies such as abstention, information retrieval, or answer selection [Wen et al., 2025].

LLMs are now frequently embedded in *agentic frameworks*, which chain multiple model invocations to solve complex, multi-step tasks [Yao et al., 2023, Xi et al., 2025]. In these settings a single unreliable output can introduce an error that compounds destructively over long-horizon tasks, potentially resulting in complete failure of the system to achieve its goal [Kwa et al., 2025]. Given the computational burden of such agentic frameworks, practical uncertainty estimators must be low latency compared to the base language models.

Semantic dispersion has emerged as a useful proxy for uncertainty: when a language model generates semantically diverse answers to the same prompt, the response is unlikely to be factual and correct. Sampling-based measures of semantic dispersion are effective for hallucination detection [Farquhar et al., 2024], but require generating multiple stochastic completions, rendering them prohibitively expensive in latency-sensitive applications.

Probing techniques can address this latency issue by training lightweight models to predict semantic dispersion or hallucination probability directly from internal model representations [Kadavath et al., 2022, Obeso et al., 2025, Han et al., 2025, Kossen et al., 2024]. However, these methods typically collapse the complex notion of language model uncertainty into a single scalar metric, ignoring the topology of the underlying output distribution. This information reduction limits their application to simple risk classification, preventing the extraction of richer reliability signals.

In this work, we propose **Semantic Self-Distillation** (SSD), a general UQ framework that bridges lightweight probing techniques and expensive sampling-based methods. Instead of regressing a scalar uncertainty statistic, SSD trains a small student model to predict a prompt-conditioned distribution over semantic answer embeddings. We model the student as a mixture density network (MDN) [Bishop, 1994], and ap-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5427–5447.

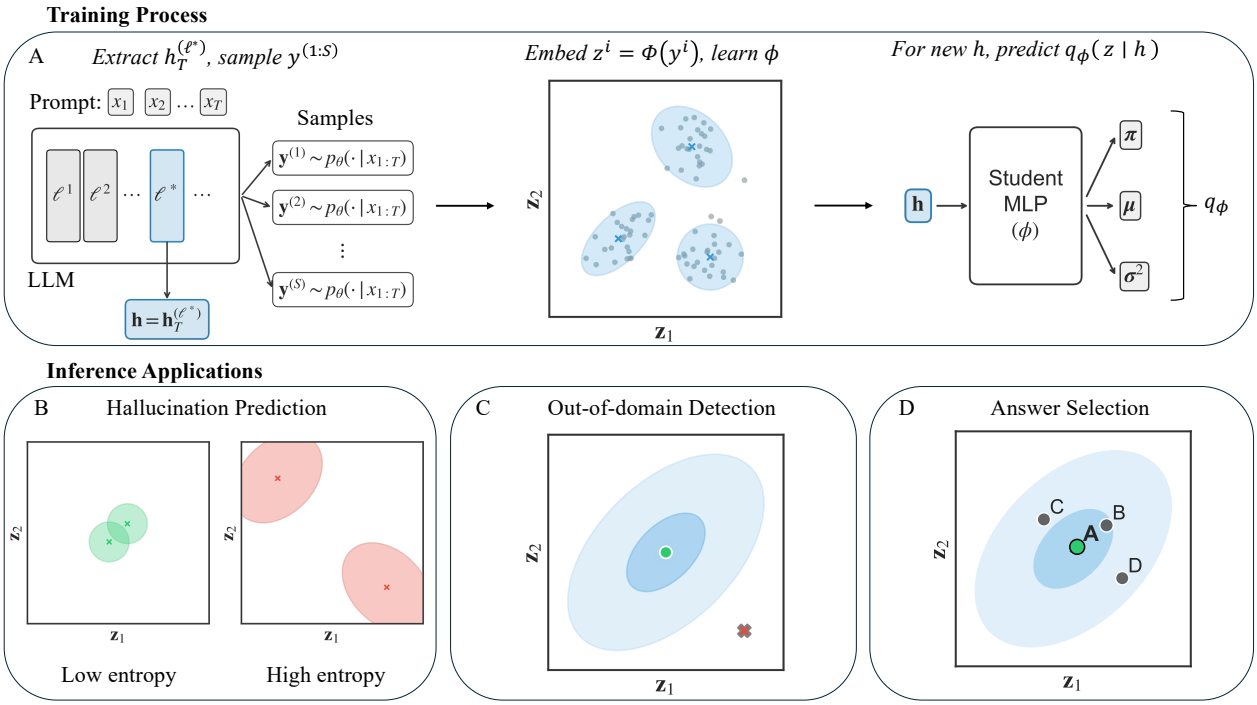

Figure 1: Overview of Semantic Self-Distillation (SSD). **(A) Training:** For each prompt, a hidden representation $\mathbf{h}$ conditions a student that distils sampled answer embeddings into a semantic mixture distribution. **(B-D) Inference:** The predicted density supports pre-generation hallucination prediction via entropy, post-generation context verification via likelihood, and multiple-choice answer selection. Plots illustrate the semantic embedding space, where points denote sampled answers.

proximate the teacher distribution using stochastic samples from the base language model. The predicted density serves as a *general-purpose uncertainty object*: its analytic entropy provides a pre-generation risk forecast, whilst estimated likelihoods enable post-generation answer verification.

Figure 1 provides an overview of the SSD training procedure and the inference-time applications investigated in this work. We evaluate SSD on a hallucination prediction task with TriviaQA and MMLU across multiple LLM families, demonstrating competitive, and in some cases even superior, performance relative to the teacher dispersion baseline.

While specialized scalar probes can be strong pre-generation classifiers, SSD uniquely retains distributional structure in a single forward pass. This enables additional post-generation capabilities unavailable to standard single-pass methods, which we demonstrate by evaluating SSD on likelihood-based out-of-domain answer detection and multiple-choice answer selection.

We summarize our contributions as the following:

- We introduce **Semantic Self-Distillation** (SSD), a method for distilling a language model's sampled semantic answer distribution into a lightweight, prompt-conditioned density estimator.
- We show that SSD provides competitive pre-generation hallucination prediction performance across multiple model families, while matching the inference latency of single-pass probing methods.
- We demonstrate that modelling the full semantic distribution unlocks new reliability primitives beyond scalar uncertainty scoring, including post-generation answer verification and distribution-based answer selection.

2 METHODS

In this section, we describe how SSD distils a language model's sampled semantic output distribution into a prompt-conditioned density estimator, rather than a scalar uncertainty proxy. We first define the representations used for prompts and answers, specify the MDN student model, and detail the distillation objective used for training. We then show how the resulting predictive density can be used to extract both pre-generation, prompt-level dispersion estimates, and post-generation, answer-level likelihood scores, all in a single forward pass at inference time.

2.1 PROBLEM SETUP

Let $x_{1:T}$ be an input prompt and $y_{1:L}$ a completion generated by a decoding model. Let $\mathbf{h} = f_\theta(x_{1:T}) \in \mathbb{R}^{d_h}$ denote

a vector representation of the prompt extracted from the model's internal states. We define a semantic representation $\mathbf{z} \in \mathbb{R}^{d_z}$ of the completed sequence via a mapping Φ, which in practice is an off-the-shelf sequence embedding model:

$$\mathbf{z} = \Phi(y_{1:L}).$$

We train a student model $q_\phi(\mathbf{z} \mid \mathbf{h})$ to predict the distribution over semantic embeddings $\mathbf{z}$ given only $\mathbf{h}$.

2.1.1 Prompt and Answer Representations

Given a user prompt $x_{1:T}$ of length T, we extract hidden states directly from the base language model. Let $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$ denote the hidden state at layer ℓ for token position t. We utilize the representation of the *final token* of the prompt extracted from a single specific layer ℓ^*:

$$\mathbf{h} \;=\; \mathbf{h}_T^{(\ell^*)}.$$

The layer index ℓ^* is treated as a hyperparameter. In practice, we estimate the optimal ℓ^* by identifying the layer most predictive of the model's semantic uncertainty, as explained in Appendix D.1.

For the answer, we apply a specialist embedding model Φ to the generated sequence $y_{1:L}$ to obtain the semantic representation, $\mathbf{z} = \Phi(y_{1:L})$, with full model details provided in Section 4.1. To improve training efficiency and focus on the principal axes of semantic variation, we also apply Principal Component Analysis (PCA) to reduce the dimensionality of the target embeddings $\mathbf{z}$.

2.1.2 Student Model

We define the student model $q_\phi(\mathbf{z} \mid \mathbf{h})$ using an MDN [Bishop, 1994]. Conditioned on the prompt representation $\mathbf{h} \in \mathbb{R}^{d_h}$, the student predicts a K-component Gaussian mixture over the semantic embedding $\mathbf{z} \in \mathbb{R}^{d_z}$:

$$q_\phi(\mathbf{z} \mid \mathbf{h}) = \sum_{k=1}^{K} \pi_k(\mathbf{h})\, \mathcal{N}\big(\mathbf{z};\, \boldsymbol{\mu}_k(\mathbf{h}),\, \mathrm{diag}(\boldsymbol{\sigma}_k^2(\mathbf{h}))\big),$$

$$\pi_k(\mathbf{h}) = \mathrm{softmax}_k\big(\boldsymbol{\alpha}(\mathbf{h})\big),$$

where an MLP with parameters ϕ maps $\mathbf{h}$ to mixture logits $\boldsymbol{\alpha}(\mathbf{h}) \in \mathbb{R}^K$, component means $\boldsymbol{\mu}_k(\mathbf{h}) \in \mathbb{R}^{d_z}$, and scales $\boldsymbol{\sigma}_k(\mathbf{h}) \in \mathbb{R}^{d_z}$. For data efficiency, we use diagonal covariance matrices.

For each training prompt x, we extract the representation $\mathbf{h}$ (using the selected layer ℓ^*) and generate S stochastic completions $y^{(1)}, \ldots, y^{(S)}$ using the base language model. We compute semantic embeddings $\mathbf{z}^{(s)} = \Phi(y^{(s)})$ for $s{=}1, \ldots, S$, and fit the student parameters ϕ by conditional maximum likelihood:

$$\max_{\boldsymbol{\phi}} \sum_x \frac{1}{S} \sum_{s=1}^{S} \log q_{\boldsymbol{\phi}}\big(\mathbf{z}^{(s)} \mid \mathbf{h}(x)\big).$$

2.2 EXTRACTING UNCERTAINTY SIGNALS

At inference time, SSD produces a prompt-conditioned semantic density $q_\phi(\mathbf{z} \mid \mathbf{h})$ over answer embeddings, which we treat as a general-purpose uncertainty object. In this work, we focus on two primary operations derived from this object: prompt-level dispersion and answer-level likelihood. We also highlight that the explicit density enables additional distributional operations such as sampling.

(i) Prompt evaluation. We use the entropy of the predicted semantic distribution $q_\phi(\cdot \mid \mathbf{h})$ to provide a prompt-level uncertainty estimate in a single forward pass.

(ii) Answer evaluation. Given any candidate response $\mathbf{y}'$, we embed it as $\mathbf{z}' = \Phi(\mathbf{y}')$ and score its compatibility with the prompt via the posterior log-likelihood $\log q_\phi(\mathbf{z}' \mid \mathbf{h})$. We may treat answers with low likelihood as out-of-domain relative to the prompt context.

(iii) Latent sampling. Because $q_\phi(\mathbf{z} \mid \mathbf{h})$ is an explicit mixture distribution, we can compute summary statistics such as the mixture mean $\mathbb{E}_{q_\phi}[\mathbf{z}] = \sum_k \pi_k(\mathbf{h})\, \boldsymbol{\mu}_k(\mathbf{h})$ and, when desired, draw latent samples $\tilde{\mathbf{z}} \sim q_\phi(\cdot \mid \mathbf{h})$.

We investigate operations (i) and (ii) in Sections 4.2 and 4.3, respectively. Operation (iii) illustrates a structural consequence of modelling the full predictive density and is discussed in Appendix B.2.

2.2.1 Quantifying Dispersion

We measure dispersion using the order-2 Rényi entropy [Rényi, 1961] of the predictive density $q_\phi(\cdot \mid \mathbf{h})$:

$$\mathsf{H}_2\big(q_\phi(\cdot \mid \mathbf{h})\big) = -\log \int q_\phi(\mathbf{z} \mid \mathbf{h})^2\, d\mathbf{z}.$$

Intuitively, the integral $\int q(\mathbf{z})^2 d\mathbf{z}$ is the *collision probability* of the density: it is larger when mass is concentrated, and smaller when the distribution is spread out [Principe, 2010]. Higher H_2 corresponds to greater semantic dispersion.

We use Rényi-2 because it yields an exact closed form for Gaussian mixtures [Wang et al., 2009], enabling a sampling-free, analytic dispersion score at inference time. By contrast, the Shannon differential entropy [Shannon, 1948] does not, in general, admit a closed-form expression for Gaussian mixture models; it is typically approximated using Monte Carlo estimation or via analytic bounds [Huber et al., 2008, Dahlke and Pacheco, 2023]. Using Rényi-2 therefore preserves SSD's goal of avoiding additional sampling or expensive numerical estimation at deployment. We provide the closed form expression and derivation in Appendix A.

2.3 EXTENSION TO GENERIC SEQUENCE DOMAINS

Although we focus on autoregressive language models in this work, the SSD construction is agnostic to modality. In language tasks, $\mathbf{z} = \Phi(y_{1:L})$ captures the semantic content of a generated answer. More generally, $\mathbf{z}$ may represent any task-relevant latent characterization of a predicted sequence $y_{1:L}$, rather than its individual tokens or events.

Under stochastic generation, a sequence model induces a predictive distribution over such latent representations. SSD distils this sampled distribution into a lightweight conditional density $q_\phi(\mathbf{z} \mid \mathbf{h})$, where $\mathbf{h}$ encodes the input sequence. The resulting density provides a *sampling-free approximation to the model's belief over high-level outcomes*. We provide further discussion of potential application domains in Section 5.

3 RELATED WORK

Uncertainty estimation in language models. Xiong et al. [2024] categorize UQ methods for LLMs into three broad categories: single-sample, multi-sample, and probing-based. Single-sample methods are training-free and include output perplexity, negative sequence probability, and mean token entropy [Malinin and Gales, 2021, Aichberger et al., 2026].

Probing-based methods similarly use a single sample at inference time but train a predictive model on internal LLM activations to predict an uncertainty-relevant quantity, such as the probability of correctness [Kadavath et al., 2022, Liu et al., 2024], semantic dispersion [Kossen et al., 2024], or token-level density [Vazhentsev et al., 2025]. Related supervised approaches fine-tune the model itself [Kapoor et al., 2024] or learn a scoring function over token probabilities [Yaldiz et al., 2025].

Multi-sample methods often rely on notions of semantic dispersion, either explicitly, such as *Semantic Entropy* [Farquhar et al., 2024], *Number of Clusters* [Lin et al., 2024], and *EigenScore* [Chen et al., 2024a], or implicitly, as with *P(true)* [Kadavath et al., 2022]. Further consistency-based variants aggregate agreement across sampled generations or perturbed inputs [Gao et al., 2024, Bhattacharjya et al., 2025], a premise formalized by the consistency hypothesis [Xiao et al., 2025]. We emphasize that these approaches typically model a *proxy* for language model uncertainty, geared for application in a particular task. There is limited consensus on how to decompose classical notions of aleatoric and epistemic uncertainty in LLMs, and the resulting uncertainty signals are often only modestly correlated [Hou et al., 2024, Xiong et al., 2024].

Distillation in language models. Distillation typically refers to training smaller language models on the *logit* or *token* outputs of more capable models, with notable success in reasoning domains [Guo et al., 2025]. Such approaches transfer predictive behaviour but do not explicitly target calibrated uncertainty or structured output distributions beyond the token level. In contrast, we focus on *sequence-level* distributions of model outputs and investigate whether these can be elicited without autoregressive sampling. Piskorz et al. [2026] study a related but narrower phenomenon concerning language model predicted distributions over numerical targets, whereas we consider broader semantic targets.

Knowledge distillation for uncertainty estimation. A related line of work seeks to distil ensemble-based uncertainty estimates into a single model [Malinin et al., 2020, Fathullah and Gales, 2022, Landgraf et al., 2024]. Ensembles are widely regarded as producing high-quality uncertainty estimates, but incur substantial computational cost at inference time. Recent approaches therefore train a student model to approximate ensemble predictive distributions or uncertainty signals, aiming to recover ensemble-quality uncertainty in a single forward pass, typically by matching logits, predictive variances, or other distributional summaries [Fathullah et al., 2023, Park et al., 2025].

In contrast, our setting does not assume access to an external ensemble and does not aim to replicate variance estimates or token-level predictive distributions. Instead, we distil a language model's *semantic sequence-level output distribution* induced by stochastic sampling, learning a prompt-conditioned density over answer embeddings. The most closely related approaches in the LLM literature are probing-based uncertainty estimators [Kadavath et al., 2022, Kossen et al., 2024], which however compress uncertainty into a single scalar statistic. By contrast, we predict a *full sequence-level distribution* in a single forward pass, enabling downstream operations beyond scalar risk scoring, including likelihood-based post-generation verification and multiple-choice answer selection. Rather than estimating a task-specific uncertainty heuristic, we approximate the underlying semantic output distribution itself.

4 EXPERIMENTS

We evaluate SSD across multiple LLM families in order to answer two core questions. First, does SSD faithfully distil sampling-based semantic dispersion into a single-pass model? Second, what additional capabilities arise from modelling a full sequence-level density rather than a scalar uncertainty proxy? To address the first question, we study *prompt-level* hallucination prediction, treating the student's entropy as a *pre-generation* risk score and comparing it against sampling-based and probing baselines. To address the second, we evaluate *answer-level* likelihood scores, demonstrating how the predicted density supports *post-generation* verification, including out-of-domain answer

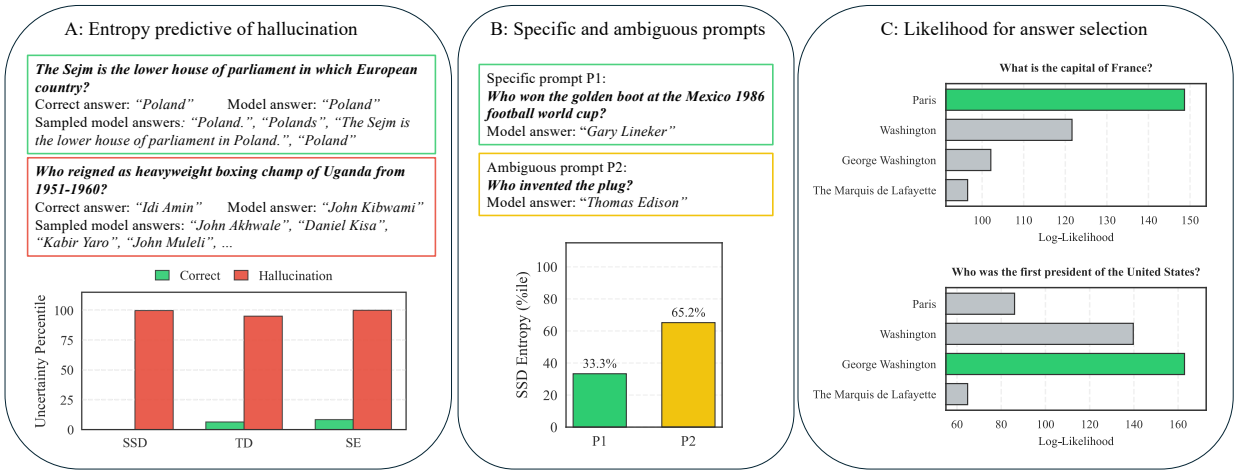

Figure 2: **Qualitative analysis of the SSD density. (A)** A correct answer yields minimal SSD entropy (0.1^{st} percentile), while a hallucination yields high uncertainty (99.7^{th}), aligning with sampling-based Teacher Dispersion (TD) and Semantic Entropy (SE). Percentiles denote rank within the TriviaQA test set (N=1,000). **(B)** A precise prompt (P1) yields significantly lower entropy than an underspecified one (P2). **(C)** The SSD likelihood enables context-aware scoring: given identical candidate sets, the student assigns the highest likelihood to the semantically appropriate answer for each prompt.

detection and multiple-choice answer selection. Figure 2 illustrates examples of each operation type.

4.1 MODELS AND DATASET GENERATION

We evaluate SSD in the context of short-form question answering, using the TriviaQA [Joshi et al., 2017] and MMLU [Hendrycks et al., 2021] datasets. We construct training sets of $N_{train} = 4{,}000$ prompts and report results on held-out test sets of $N_{test} = 1{,}000$ prompts.

MMLU is formatted as a multiple-choice benchmark, as opposed to the open-ended TriviaQA. To properly assess semantic dispersion and enable comparison across datasets, we use a filtered subset of MMLU containing only open-ended questions [Myrzakhan et al., 2024].

We select these datasets because they represent two distinct task types: **knowledge-seeking** (TriviaQA) and **reasoning-heavy** (MMLU). Prior work has found semantic dispersion to be a useful proxy for language model uncertainty in the former task type and less so in the latter [Xiong et al., 2024]; as such MMLU provides a more challenging setting for evaluating SSD, although in practice it reflects a mixture of both task types.

We evaluate across five model families: **Mistral** (Ministral-8B-Instruct), **Llama-3** (3.1-8B-Instruct, 3.2-3B-Instruct), **Qwen** (Qwen3-4B-Instruct-2507, Qwen3-8B), **Gemma** (Gemma-3-4B) and **SmolLM** (SmolLM3-3B) [Grattafiori et al., 2024, Mistral, 2024, Yang et al., 2025, Team et al., 2025, Bakouch et al., 2025]. These models span multiple architectural families, parameter scales (3B-8B), and training pipelines, allowing us to assess the robustness and generality of SSD across diverse backbone designs rather than tailoring it to a single model.

For all experiments, we use Qwen3-8B as the oracle 'judge' model to determine the correctness of generated answers given a reference answer [Zheng et al., 2023]. For each prompt and model, we generate $S = 32$ stochastic samples at temperature $T = 1$. We additionally sample a single 'default' answer at $T = 0.1$, which is provided to the judge model to obtain the hallucination label.

For answer embeddings, we use EmbeddingGemma [Vera et al., 2025], with the embedding truncated to a dimension of $d_z = 128$. For hallucination prediction, we further reduce dimensionality using PCA to $d'_z = 16$, following ablation studies detailed in Appendix C. For likelihood-based OOD verification, we set $d'_z = 32$.

The student MDNs themselves are lightweight: across the configurations selected in Appendix D, total parameter counts range from approximately 600K to 10M depending on hidden width, depth, and the base model's embedding dimension - orders of magnitude smaller than the 3B-8B parameter base LLMs.

4.2 PROMPT-LEVEL HALLUCINATION PREDICTION

We first evaluate whether the distilled semantic density provides a reliable *pre-generation* hallucination signal. Specifically, we treat the entropy of the predicted distribution as a prompt-level risk score: higher entropy corresponds to greater semantic dispersion and increased hallucination risk. We report AUROC and AUPRC against correctness labels

Table 1: **Hallucination prediction trade-offs.** AUROC is reported per dataset as a macro-average (mean ± std) across models. Latency summarizes inference-time requirements (single-pass vs. S sampled generations, with entailment-based SE additionally requiring S^2 natural language inference (NLI) comparisons). Target cost denotes the computation required to construct supervision targets during training; C_{CL} is the cost of acquiring correctness labels. Only SSD produces an explicit predictive density, enabling post-generation likelihood scoring and latent sampling in addition to pre-generation uncertainty.

Method	TriviaQA ↑	MMLU ↑	Latency ↓	Target cost ↓	Predictive density
SE	**0.804 ± 0.035**	0.663 ± 0.040	S samples + S^2 NLI	–	×
SEP	0.759 ± 0.026	0.628 ± 0.029	**Single-pass**	S samples + S^2 NLI	×
PCP	0.771 ± 0.020	**0.679 ± 0.019**	**Single-pass**	C_{CL}	×
TD	0.699 ± 0.027	0.647 ± 0.023	S samples	–	×
SSD	0.708 ± 0.024	0.623 ± 0.025	**Single-pass**	S samples + S Φ	✓

Table 2: **Hallucination prediction, comparison with teacher dispersion.** TD computes dispersion from $S = 32$ sampled semantic embeddings at inference time, while SSD predicts a prompt-conditioned semantic density and scores dispersion analytically. We report hallucination prediction AUROC/AUPRC in % as the mean over 1,000 bootstrap resamples of the test set, with standard deviation as subscript; bold indicates the better of TD vs. SSD per metric and dataset.

	AUROC		AUPRC	
Model	TD	SSD	TD	SSD
TriviaQA				
Qwen3 8B	**$68.3_{1.6}$**	$67.7_{1.8}$	$52.4_{2.6}$	**$56.9_{2.8}$**
Qwen3 4B	**$73.2_{1.6}$**	$69.9_{1.6}$	$59.7_{2.4}$	**$61.9_{2.5}$**
Llama 3.1 8B	**$69.6_{1.7}$**	$68.7_{1.9}$	$44.4_{2.7}$	**$51.9_{2.9}$**
Llama 3.2 3B	$66.6_{1.7}$	**$73.7_{1.6}$**	$55.8_{2.6}$	**$65.5_{2.5}$**
Ministral 8B	$74.3_{1.7}$	**$74.9_{1.6}$**	$57.9_{2.8}$	**$59.9_{2.8}$**
SmolLM3 3B	$69.9_{1.6}$	**$71.1_{1.6}$**	$57.1_{2.5}$	**$62.4_{2.5}$**
Gemma 3 4B	$67.6_{1.6}$	**$70.0_{1.7}$**	$63.1_{2.3}$	**$66.1_{2.5}$**
MMLU				
Qwen3 8B	**$63.5_{1.7}$**	$57.1_{1.8}$	**$61.9_{2.3}$**	$55.9_{2.3}$
Qwen3 4B	$63.1_{1.8}$	**$64.9_{1.7}$**	$58.5_{2.4}$	**$59.4_{2.5}$**
Llama 3.1 8B	**$69.0_{1.6}$**	$63.8_{1.7}$	**$70.9_{2.1}$**	$62.7_{2.4}$
Llama 3.2 3B	**$67.1_{1.8}$**	$63.8_{1.7}$	$74.0_{2.0}$	**$74.5_{1.9}$**
Ministral 8B	**$65.1_{1.8}$**	$63.6_{1.8}$	$71.4_{2.1}$	**$71.7_{2.1}$**
SmolLM3 3B	**$62.3_{1.8}$**	$62.0_{1.8}$	**$64.6_{2.2}$**	$63.0_{2.3}$
Gemma 3 4B	**$62.8_{1.7}$**	$60.7_{1.8}$	**$71.1_{2.0}$**	$69.2_{2.1}$

determined using an LLM-as-judge framework.

SSD compares favourably to teacher dispersion. We compare the student's dispersion score to the standard deviation of the teacher's $S = 32$ semantic embeddings, which we term *teacher dispersion* (TD). Table 2 reports results across models and datasets. On the TriviaQA dataset, despite requiring no inference-time sampling, SSD outperforms TD on four of the seven models investigated in terms of AUROC and on all seven in AUPRC. We attribute this in part to a smoothing effect: TD is a finite-sample statistic and assigns zero uncertainty to prompts where the model consistently produces the same answer, even when that answer is incorrect. By contrast, SSD learns a continuous mapping from prompt representations to semantic uncertainty, enabling differentiated risk estimates in such degenerate cases, which likely contributes to its stronger AUPRC.

On the MMLU dataset, SSD is a relatively weaker performer compared to the sampled teacher signal, which we attribute to the task being less suited for the semantic dispersion uncertainty proxy. SSD still surpasses TD in three out of seven models from an AUPRC perspective.

Overall, these results demonstrate that SSD successfully distils sampling-based semantic dispersion into a single-pass density estimator. Despite requiring no inference-time sampling, the student recovers much of the teacher's uncertainty signal, and in some cases improves upon the finite-sample baseline. The distilled signal is also not narrowly tied to its training distribution: students evaluated on the dataset they were not trained on lose only 0.04-0.05 mean AUROC relative to in-distribution students (Appendix B.3).

Comparison to additional baselines. We compare SSD against three widely used hallucination prediction methods: (1) **Probability of Correctness Probe (PCP)**, a logistic regression classifier trained on the prompt representation to predict correctness. Originally proposed as 'P(I know)' by Kadavath et al. [2022], this method requires externally provided correctness labels. (2) **Semantic Entropy (SE)**, a sampling-based approach that measures semantic dispersion by clustering S completions using a natural language inference (NLI) model to assess mutual entailment [Farquhar et al., 2024]. (3) **Semantic Entropy Probe (SEP)**, a lightweight probe with the same architecture as PCP, trained to regress a binarized SE target, and inheriting SE's sampling and NLI-based target construction cost [Kossen et al., 2024].

To provide a fuller comparison of methods, we report hallucination prediction performance alongside method characteristics: inference latency, training supervision target construction cost, and whether a method supports post-generation signals in addition to pre-generation uncertainty. Table 1 summarizes these trade-offs and reports macro-

average AUROC across model families; full per-model AUROC/AUPRC tables are provided in Appendix F. We provide further analysis of the computational complexity of each method in Appendix E.

Overall, SE achieves the highest AUROC on TriviaQA, whilst PCP is the strongest performer on MMLU, consistent with prior work [Xiong et al., 2024]. While SE and PCP achieve the best performance in their respective regimes, these methods either incur significant sampling cost (SE) or reduce uncertainty to a supervised scalar classification problem (PCP). SSD occupies a distinct regime: it matches probe-level inference latency, avoids entailment-based target construction during training, and preserves an explicit predictive density. Although its pre-generation hallucination prediction performance is slightly weaker than specialized probes, it uniquely provides post-generation likelihood signals derived from the same model.

We further observe that the SSD performance is model-dependent. On TriviaQA, for example, it achieves an AUROC of 74.9 for Ministral 8B but only 67.7 for Qwen3 8B (Table A7). Analysis in Appendix B.1 indicates that this variation is driven primarily by the *fidelity* of the distillation process: for some models, it is particularly difficult to predict the semantic distribution from the prompt representation. We discuss these findings further in Section 5.

4.3 ANSWER-LEVEL RESPONSE EVALUATION

Having established that SSD recovers sampling-based dispersion signals, we next evaluate the additional capabilities enabled by modelling an explicit predictive density. Unlike scalar uncertainty scores, the density allows us to assess the quality of *specific* candidate answers by measuring their likelihood under the model's predicted semantic distribution. We examine this utility in two settings: out-of-domain answer detection (binary classification) and multiple-choice question answering (multi-class selection).

4.3.1 Out-of-domain Answer Detection

We first evaluate the student's ability to detect mismatched prompt-answer pairs. We frame this as an out-of-domain detection task: for each test prompt x_i, we pair its representation $\mathbf{h}_i$ with (i) the model's default answer y_i and (ii) a random answer y_j drawn from a different prompt. Each candidate answer is embedded as $\mathbf{z} = \Phi(y)$ and scored using the student's log-likelihood $\log q_\phi(\mathbf{z} \mid \mathbf{h})$. We report AUROC for distinguishing matched from mismatched pairs.

As shown in Table 3, the SSD log-likelihood is a strong context-verification signal, achieving AUROCs above 0.95 for most base models on both TriviaQA and MMLU. Practically, this could enable SSD to act as a content 'guard' [Inan et al., 2023], for example detecting cache poisoning [Chen et al., 2024b] or filtering retrieved content in retrieval-augmented generation [Lewis et al., 2020]. This mismatch-detection capability also generalizes across distribution shift: students evaluated on the dataset they were not trained on retain AUROCs of 0.75–0.84 (Appendix B.3).

Table 3: **Out-of-domain answer detection.** We report AUROC for two binary classification tasks using the student's log-likelihood score. **Mismatch** measures the ability to distinguish the model's answer to the prompt from a random answer from the dataset. **Hallucination** measures the ability to distinguish correct model-generated answers from incorrect ones. Values report mean AUROC with bootstrap standard deviation over 1,000 resamples.

Model	Mismatch	Hallucination
TriviaQA		
Qwen3 8B	0.984 ± 0.002	0.571 ± 0.019
Qwen3 4B	0.984 ± 0.002	0.557 ± 0.018
Llama 3.1 8B	0.989 ± 0.002	0.579 ± 0.019
Llama 3.2 3B	0.980 ± 0.003	0.613 ± 0.017
Ministral 8B	0.945 ± 0.005	0.616 ± 0.017
SmolLM3 3B	0.988 ± 0.002	0.562 ± 0.018
Gemma 3 4B	0.895 ± 0.007	0.578 ± 0.018
MMLU		
Qwen3 8B	0.975 ± 0.003	0.526 ± 0.018
Qwen3 4B	0.960 ± 0.004	0.566 ± 0.019
Llama 3.1 8B	0.981 ± 0.003	0.576 ± 0.018
Llama 3.2 3B	0.942 ± 0.005	0.519 ± 0.019
Ministral 8B	0.966 ± 0.004	0.529 ± 0.019
SmolLM3 3B	0.985 ± 0.003	0.528 ± 0.018
Gemma 3 4B	0.923 ± 0.006	0.548 ± 0.019

We also evaluate the student likelihood as a *post-generation* hallucination signal, motivated by prior work showing that probability density in semantic space is predictive of correctness [Qiu and Miikkulainen, 2024]. While informative, our likelihood signal is weaker than the pre-generation entropy, with AUROC scores of approximately 0.6. Unlike sampling-based density methods, which construct the semantic distribution from multiple generations at inference time, SSD predicts the density from the prompt representation alone. The weaker performance reflects the limits of this single-pass approximation to a sampling-derived density.

We therefore do not advocate the likelihood as a stand-alone hallucination predictor. We report these results to establish that the signal is informative (AUROC above 0.5) and complementary in principle to the pre-generation entropy; the primary post-generation use case remains the mismatch detection task above.

4.3.2 Multiple-choice Answer Selection

We next evaluate whether the student's predicted density can be used to identify the correct answer in a multiple-choice setting. We utilize the open-ended subset of MMLU provided by Myrzakhan et al. [2024], which filters for questions that are self-contained and do not require the context of multiple-choice options to answer.

The student is trained using only open-ended completions to prompts in the training set. For the test set, we take the four provided options (A, B, C, D) for each question, embed them using Φ, and rank them according to their log-likelihood $\log q_\phi(\mathbf{z}' \mid \mathbf{h})$ under the student mixture.

Whereas the mismatched answers in Section 4.3.1 are semantically distant from the prompt by construction, the four MMLU options are deliberately close in semantic space, so this experiment poses a more demanding test: can the density discriminate fine-grained semantic structure within the plausible answer region?

Table 4 reports the accuracy of selecting the highest-likelihood option (**Top-1**) alongside the base model's open-ended answer accuracy as assessed by an LLM-as-judge. SSD consistently exceeds the 0.25 random baseline, and for our best performing models (Llama 3.2 and Ministral) approaches the base model's open-ended accuracy.

The open-ended accuracy of the base model provides a rough estimate of its dataset-specific knowledge. Despite being trained only on sampled open-ended completions and without multiple-choice supervision, the student approaches this level of performance when provided with the options, suggesting that the distilled semantic density captures meaningful structure in the model's answer distribution. We emphasize that SSD is not proposed as a substitute for directly querying the model; the purpose of the experiment is to validate the predicted density as a usable belief state.

Summary of experimental findings. Taken together, the prompt-level and answer-level evaluations demonstrate that SSD's value lies not in a particular scalar uncertainty score, but in exposing an explicit semantic belief state from which multiple reliability signals can be derived. By modelling a prompt-conditioned distribution rather than a scalar proxy, SSD unifies pre-generation risk estimation and post-generation verification within a single, low-latency framework. The performance variation observed across base LLMs points to differences in how semantic uncertainty is encoded in internal activations.

We therefore position SSD as an appropriate UQ tool when inference latency precludes sampling-based methods, when the labelled correctness data required by supervised probes is unavailable or expensive, or when post-generation answer-level operations are needed alongside pre-generation risk estimates. Practically, we would recommend computing the distillation fidelity (discussed in Section 5 and Appendix B.1) on a held-out validation set before using SSD in deployment. A student that fails to track teacher dispersion at training time should not be relied upon for risk estimation at inference.

Table 4: **Multiple-choice answer selection.** We evaluate the ability of the SSD student - trained solely on open-ended generations - to identify the correct answer in a multiple-choice setting using an open-ended subset of MMLU. We report **Top-1** accuracy, where the option with the highest predicted likelihood is selected (random baseline $= 0.25$). **Open-Ended** accuracy denotes the base model's accuracy when generating answers directly, as determined by an LLM-as-judge.

Model	SSD Top-1	Open-Ended
Qwen3 8B	0.327 ± 0.015	0.515 ± 0.015
Qwen3 4B	0.312 ± 0.014	0.545 ± 0.016
Llama 3.1 8B	0.339 ± 0.015	0.483 ± 0.015
Llama 3.2 3B	0.346 ± 0.015	0.358 ± 0.015
Ministral 8B	0.363 ± 0.015	0.396 ± 0.015
SmolLM3 3B	0.332 ± 0.015	0.466 ± 0.016
Gemma 3 4B	0.292 ± 0.014	0.399 ± 0.015

5 DISCUSSION & FUTURE WORK

The experimental results highlight both the strengths and current limitations of SSD. While modelling a full semantic density enables richer post-generation capabilities than scalar probes, its effectiveness depends on how faithfully the student can reconstruct the teacher's sampled distribution from prompt representations alone. We now discuss these limitations and outline directions for strengthening the distributional framework.

Improving distillation fidelity. Our results in Appendix B.1 show a strong correlation between distillation fidelity and hallucination prediction performance, indicating that SSD's effectiveness depends on how accurately the student reconstructs the teacher's prompt-induced uncertainty ordering. Promising directions to improve this fidelity include scaling the distillation dataset size, and distilling from stronger teachers (for example sampling more responses per question, or using higher-quality semantic targets).

We believe the most significant gains could be found by enriching the student input beyond a single-token/single-layer representation, something we explore in Appendix B.4. Using only simple enrichment approaches, we find modest and model-dependent gains in SSD performance, motivating more principled investigation of optimal probe design [Kramár et al., 2026]. Subsequent findings may clarify whether observed performance gaps between models

stem from limited semantic information in the prompt representation or insufficient supervision for learning high-dimensional densities.

Combining pre- and post-generation signals. SSD exposes complementary reliability signals from a single predicted density: pre-generation dispersion forecasts risk, while post-generation likelihood verifies prompt–answer compatibility. A natural next step is to integrate these signals into unified decision policies for LLM-based systems, for example abstaining or retrieving evidence under high dispersion, and accepting or rejecting candidate actions based on likelihood thresholds.

Diffusion models. Although our experiments focus on autoregressive models, SSD may be particularly well suited to diffusion-based language models [Sahoo et al., 2024]. In autoregressive decoding, the model's belief over the final answer is implicit in a long sequence of conditional token distributions. In masked diffusion models, by contrast, the model produces predictions over the full output sequence at every denoising step, so sequence-level information is more concentrated in early internal states. The mapping from prompt representation to final output distribution that SSD must learn may therefore be more direct, potentially reducing the gap between the model's internal belief state and the semantic distribution induced by sampling.

Domain-specific uncertainty estimation. As described in Section 2.3, SSD extends beyond language to sequence models whose outputs admit meaningful latent representations. We see particular potential utility for this technique in electronic health record (EHR) foundation models, which tokenize longitudinal patient histories and generate future clinical event sequences [Renc et al., 2024]. These models typically predict discrete events, but clinical utility often depends on a higher-level characterization of the patient's future state, such as overall disease burden or risk profile [Steinberg et al., 2024, Zeng et al., 2025].

In this setting, the embedding model could correspond to a transformer-based EHR encoder such as BEHRT [Li et al., 2020], summarizing a predicted trajectory into a clinically informative latent health representation. Sampling from $q_\phi(\mathbf{z} \mid \mathbf{h})$ could then yield low-latency forecasts of plausible future patient states, while the entropy of the density would quantify uncertainty over long-term outcomes. Similar structure arises in time-series foundation models [Faw et al., 2025], where future trajectories may be summarized by latent representations capturing regime, trend, or seasonal structure [Wang et al., 2022].

LLM-as-judge evaluation. Our correctness labels derive from a single judge model (Qwen3-8B). LLM-as-judge has been found to be the least length-biased among common correctness functions [Santilli et al., 2025], supporting this protocol choice, but judge-induced bias cannot be excluded, and marginalizing over multiple judge variants could offer more reliable labels in future work [Ielanskyi et al., 2026].

Scaling to larger models. SSD requires S stochastic completions per training prompt, a one-time offline cost that grows with model scale. In practice, such datasets are routinely produced during RL-based post-training, where multiple rollouts per prompt are generated by algorithms such as PPO [Schulman et al., 2017] and GRPO [Shao et al., 2024]. Where such pipelines exist, the marginal cost of SSD implementation is small, even for very large models.

6 CONCLUSION

We introduced Semantic Self-Distillation (SSD), a framework for distilling a language model's sampled semantic answer distribution into a lightweight, prompt-conditioned predictive density. Rather than reducing uncertainty to a scalar score, SSD models the geometry of the semantic output space itself, enabling analytic dispersion estimates and likelihood-based evaluation in a single forward pass.

Empirically, SSD performs competitively with the teacher's sampled semantic dispersion across several model families, while avoiding the inference-time overhead of generating multiple completions. Although it does not dominate specialized scalar probes on any single task, SSD uniquely combines probe-level latency, freedom from correctness-label and entailment-based supervision, and an explicit predictive density within a single estimator. Crucially, modelling the full density enables capabilities unavailable to scalar methods, including answer-level reliability scoring and distribution-based selection.

More broadly, SSD demonstrates that uncertainty in complex generative models can be distilled as an explicit density over latent outcome representations. By shifting computational cost from inference to training, SSD provides a practical mechanism for integrating principled, distribution-level uncertainty into latency-critical language systems and other sequence domains.

Acknowledgements

DAC was funded by an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE); and was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and the Pandemic Sciences Institute at the University of Oxford. EP was funded by an NIHR Research Studentship. SW was supported by the Rhodes Scholarship.

References

Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. Rethinking uncertainty estimation in LLMs: A principled single-sequence measure. In *The Fourteenth International Conference on Learning Representations*, 2026.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. `https://huggingface.co/blog/smollm3`, 2025.

Debarun Bhattacharjya, Balaji Ganesan, Junkyu Lee, Radu Marinescu, Katya Mirylenka, Michael Glass, and Xiao Shou. SIMBA UQ: Similarity-based aggregation for uncertainty quantification in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15880–15894, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.859.

Christopher M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, Birmingham, UK, 1994. URL `https://publications.aston.ac.uk/id/eprint/373/`.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

Caleb Dahlke and Jason Pacheco. On convergence of polynomial approximations to the Gaussian mixture entropy. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. Publisher: Nature Publishing Group.

Yassir Fathullah and Mark JF Gales. Self-distribution distillation: efficient uncertainty estimation. In *Uncertainty in Artificial Intelligence (UAI)*, pages 663–673. PMLR, 2022.

Yassir Fathullah, Guoxuan Xia, and Mark JF Gales. Logit-based ensemble distribution distillation for robust autoregressive sequence uncertainties. In *Uncertainty in Artificial Intelligence (UAI)*, pages 582–591. PMLR, 2023.

Matthew Faw, Rajat Sen, Yichen Zhou, and Abhimanyu Das. In-context fine-tuning for time-series foundation models. In *Forty-second International Conference on Machine Learning*, 2025.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.143. URL `https://aclanthology.org/2024.eacl-long.143/`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Jiatong Han, Neil Band, Muhammed Razzak, Jannik Kossen, Tim G. J. Rudner, and Yarin Gal. Simple factuality probes detect hallucinations in long-form natural language generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16209–16226, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.880.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for

large language models through input clarification ensembling. In *Forty-first International Conference on Machine Learning*, 2024.

Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188. IEEE, 2008.

Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. Addressing pitfalls in the evaluation of uncertainty estimation methods for natural language generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: LLM-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-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, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for gemini, 2026. URL https://arxiv.org/abs/2601.11516.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Roa Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long software tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Steven Landgraf, Kira Wursthorn, Markus Hillemann, and Markus Ulrich. DUDES: Deep uncertainty distillation using ensembles for semantic segmentation. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(2):101–114, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. Uncertainty estimation and quantification for LLMs: A simple supervised approach, 2024. URL https://arxiv.org/abs/2404.15993.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6107–6117, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736569.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

Mistral. Un ministral, des ministraux. Mistral AI research announcement, Oct 2024. URL https://mistral.ai/news/ministraux.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-LLM-Leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena, 2024. URL https://arxiv.org/abs/2406.07545.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-time detection of hallucinated entities in long-form generation, 2025. URL https://arxiv.org/abs/2509.03531.

Sehyun Park, Jongjin Lee, Yunseop Shin, Ilsang Ohn, and Yongdai Kim. Knowledge distillation of uncertainty using deep latent factor model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Edward Phillips, Sean Wu, Soheila Molaei, Danielle Belgrave, Anshul Thakur, and David Clifton. Geometric uncertainty for detecting and correcting hallucinations in LLMs, 2025. URL https://arxiv.org/abs/2509.13813.

Julianna Piskorz, Kasia Kobalczyk, and Mihaela van der Schaar. Eliciting numerical predictive distributions of LLMs without auto-regression. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jose C Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7(1):256, 2024.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 743–759, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.60.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie

Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.113.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. Embeddinggemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

Fei Wang, Tanveer Syeda-Mahmood, Baba C Vemuri, David Beymer, and Anand Rangarajan. Closed-form Jensen-Renyi divergence for mixture of Gaussians and applications to group-wise shape registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–655. Springer, 2009.

Zhiyuan Wang, Xovee Xu, Goce Trajcevski, Weifeng Zhang, Ting Zhong, and Fan Zhou. Learning latent seasonal-trend representations for time series forecasting. In *Advances in Neural Information Processing Systems*, 2022.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katsiaryna Mirylenka, Nhan H Pham, Michael Glass, and Junkyu Lee. The consistency hypothesis

in uncertainty quantification for large language models. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Miao Xiong, Andrea Santilli, Michael Kirchhof, Adam Golinski, and Sinead Williamson. Efficient and effective uncertainty quantification for LLMs. In *Neurips Safe Generative AI Workshop*, 2024.

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. Do not design, learn: A trainable scoring function for uncertainty estimation in generative LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 691–713, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.41.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Sihang Zeng, Lucas Jing Liu, Jun Wen, Meliha Yetisgen, Ruth Etzioni, and Gang Luo. Trajsurv: Learning continuous latent trajectories from electronic health records for trustworthy survival prediction, 2025. URL https://arxiv.org/abs/2508.00657.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Semantic Self-Distillation for Language Model Uncertainty (Appendix)

Edward Phillips[1] **Sean Wu**[1] **Fredrik K. Gustafsson**[1] **Boyan Gao**[1] **David A. Clifton**[1, 2]

[1]Department of Engineering Science, University of Oxford
[2]Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou

A DERIVATIONS

Closed-form computation of Rényi–2 entropy. For a Gaussian mixture, the quadratic functional $\int q(\mathbf{z})^2 d\mathbf{z}$ decomposes into pairwise overlaps between mixture components:

$$\int q_\phi(\mathbf{z} \mid \mathbf{h})^2 \, d\mathbf{z} = \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i(\mathbf{h}) \, \pi_j(\mathbf{h}) \int \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \, \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \, d\mathbf{z}.$$

Using the standard Gaussian product identity,

$$\int \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \, \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \, d\mathbf{z} = \mathcal{N}\big(\boldsymbol{\mu}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\big),$$

we obtain:

$$\mathsf{H}_2\big(q_\phi(\cdot \mid \mathbf{h})\big) = -\log \left(\sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i(\mathbf{h}) \, \pi_j(\mathbf{h}) \, \mathcal{N}\big(\boldsymbol{\mu}_i(\mathbf{h}); \boldsymbol{\mu}_j(\mathbf{h}), \boldsymbol{\Sigma}_i(\mathbf{h}) + \boldsymbol{\Sigma}_j(\mathbf{h})\big) \right).$$

Equivalently, defining the collision matrix $\mathbf{K}(\mathbf{h}) \in \mathbb{R}^{K \times K}$ with entries

$$K_{ij}(\mathbf{h}) = \mathcal{N}\big(\boldsymbol{\mu}_i(\mathbf{h}); \boldsymbol{\mu}_j(\mathbf{h}), \boldsymbol{\Sigma}_i(\mathbf{h}) + \boldsymbol{\Sigma}_j(\mathbf{h})\big),$$

and $\boldsymbol{\pi}(\mathbf{h}) \in \mathbb{R}^K$ the mixture weights, we have

$$\mathsf{H}_2\big(q_\phi(\cdot \mid \mathbf{h})\big) = -\log\big(\boldsymbol{\pi}(\mathbf{h})^\top \mathbf{K}(\mathbf{h}) \, \boldsymbol{\pi}(\mathbf{h})\big).$$

This computation is $\mathcal{O}(K^2)$ in the number of mixture components and is negligible relative to the transformer backbone.

B FURTHER ANALYSES

In this section we provide additional analyses supporting observations made in the main text. Unless otherwise specified, results are reported on the TriviaQA dataset, where SSD exhibited the strongest performance variation across base models.

B.1 DISTILLATION FIDELITY VS. DETECTION PERFORMANCE

We examine whether the quality of distillation (how closely the student learns the teacher distribution) constrains hallucination detection performance. We define *distillation fidelity* (ρ_{fidelity}) as the Spearman rank correlation, over $N = 1000$ test prompts, between the student's predicted entropy and the teacher dispersion. A high ρ_{fidelity} indicates a more faithful recovery of the teacher's prompt-induced uncertainty ordering.

Figure A1 plots hallucination detection AUROC against distillation fidelity across base models and student hyperparameter settings. We observe a strong positive relationship, with a cross-model Spearman correlation of $\rho_{\text{meta}} = 0.65$, indicating that detection performance is limited by the student's ability to learn the prompt-to-dispersion mapping. The failure cases, notably Qwen3-4B, exhibit near-zero ρ_{fidelity}, suggesting that the relevant uncertainty signal is not accessible from the chosen prompt representation.

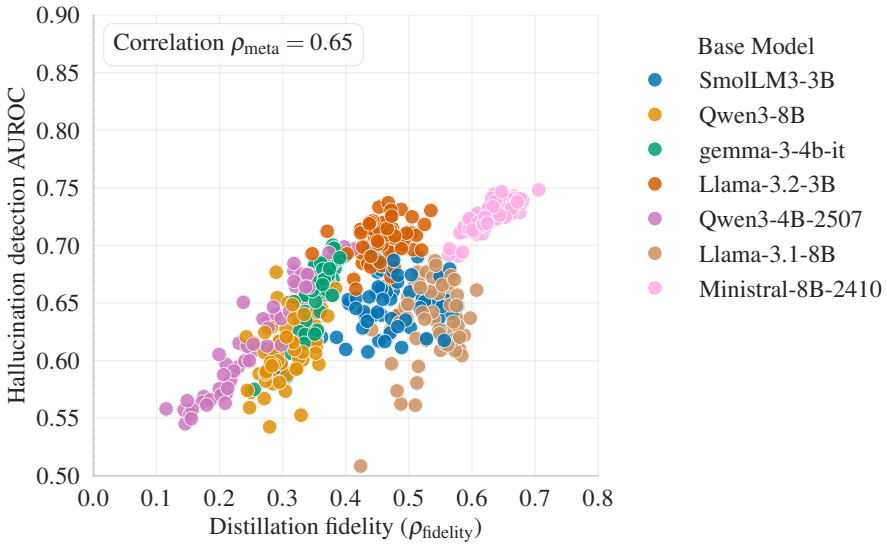

Figure A1: **Distillation fidelity drives detection performance.** The X-axis shows *distillation fidelity* (ρ_{fidelity}): the Spearman correlation between the student's predicted entropy and the teacher dispersion across TriviaQA test prompts. Multiple points per model correspond to different student hyperparameter settings. ρ_{meta} is the Spearman correlation between the X and Y variables; it shows that models where the student fails to learn the distribution (low ρ_{fidelity}) result in poor detection.

B.2 SEMANTIC CONSENSUS FROM THE PREDICTED MIXTURE

In Section 2.2, we noted that modelling an explicit predictive density enables latent-space sampling and computation of mixture statistics such as the mean. In language models, the mixture mean can serve as a proxy for semantic consensus. Whereas autoregressive decoding produces a single trajectory through output space, the mixture mean aggregates across the model's sampled distribution in embedding space. We evaluate whether this analytic mean more faithfully approximates the empirical semantic centroid of sampled answers than a single low-temperature decoding.

We treat the empirical mean of $S = 32$ stochastic samples in embedding space as a reference semantic centroid, then measure the mean squared distance (MSD) in Euclidean space between this centroid and both (i) the model's low-temperature output and (ii) the analytical mean of the student mixture, $\boldsymbol{\mu}_{ssd} = \sum_k \pi_k \boldsymbol{\mu}_k$. Lower MSD indicates better recovery of the model's aggregate semantic belief. MSD is computed as $\mathbb{E}_i \left[\|\hat{\mathbf{z}}_i - \boldsymbol{\mu}_i^{\text{true}}\|_2^2 \right]$ over examples i in the subset.

We report results for the best-performing student configuration (Ministral 8B), reflecting the strong architectural dependence of distillation fidelity. As shown in Table A1, the SSD mixture mean provides a closer approximation to the semantic centroid than a single default decoding pass. This effect is most pronounced for incorrect answers, where SSD improves centroid recovery in over 76% of test cases. This aligns with prior work showing that hallucinated outputs tend to deviate from semantic consensus [Lin et al., 2024, Phillips et al., 2025]. Together with the OOD results in Table 3, this result suggests that SSD exposes a coherent semantic belief state from which both prior (entropy) and posterior (likelihood and distance-to-consensus) reliability signals can be derived.

B.3 CROSS-DATASET GENERALIZATION

We test whether trained SSD students capture structure that transfers beyond their training distribution: a student trained on one dataset is evaluated directly on the other, with no retraining. The PCA basis fit on the source dataset is reused to project target answer embeddings, so all scores are computed in a consistent space. The teacher dispersion (TD) baseline is

Table A1: Semantic consensus approximation with Ministral 8B. We report the mean squared distance (MSD) of the default answer embedding and the SSD predicted mean to the true semantic centroid (average of $S = 32$ samples). Values denote mean MSD, with the bootstrap standard deviation expressed as a percentage of the mean shown as a subscript.

Subset	Default MSD ↓	SSD MSD ↓	Imp. (%) ↑	SSD Win Rate (%) ↑
All	$0.015_{3.2}$	$\mathbf{0.008_{1.9}}$	+45.5	58.3
Correct	$0.013_{4.4}$	$\mathbf{0.009_{2.3}}$	+32.3	48.8
Incorrect	$0.018_{4.2}$	$\mathbf{0.007_{3.4}}$	+63.0	76.5

Table A2: **Cross-dataset generalization.** Students trained on the source dataset are evaluated on the target dataset without retraining. Values report AUROC with bootstrap standard deviation over 1,000 resamples of the target test set.

Model	In-dist (target)	SSD	TD	Mismatch
MMLU → TriviaQA				
Qwen3 4B	0.699	0.672 ± 0.017	0.728 ± 0.016	0.815 ± 0.009
Llama 3.2 3B	0.737	0.663 ± 0.017	0.664 ± 0.017	0.751 ± 0.011
Ministral 8B	0.748	0.717 ± 0.017	0.732 ± 0.016	0.780 ± 0.010
TriviaQA → MMLU				
Qwen3 4B	0.649	0.595 ± 0.018	0.633 ± 0.017	0.777 ± 0.010
Llama 3.2 3B	0.638	0.574 ± 0.018	0.669 ± 0.018	0.793 ± 0.010
Ministral 8B	0.635	0.595 ± 0.019	0.641 ± 0.018	0.838 ± 0.009

computed on the target dataset in the same source-fit basis, ensuring a single coordinate system for the comparison. We report results for three models, selected to span the major architectural families in our evaluation, using the evaluation protocol of Tables 2 and 3: scores on the target test set (N=1,000) with bootstrap standard deviations over 1,000 resamples. *In-dist (target)* denotes the AUROC of a student trained on the target dataset itself; the gap between this column and *SSD* is the transfer cost. *Mismatch* repeats the out-of-domain answer detection task of Table 3 using the source-trained student.

Table A2 reports the results, from which we highlight three findings. First, hallucination prediction transfers with limited cost: the mean AUROC drop relative to an in-distribution student is 0.04 (MMLU → TriviaQA) and 0.05 (TriviaQA → MMLU), comparable in both directions. Second, mismatch detection generalizes strongly (AUROC 0.75–0.84): while reduced relative to the in-distribution setting (≈0.95, Table 3), the predicted density continues to correctly localize matched answers and reject mismatched ones under distribution shift. Third, the asymmetry between directions is mild, suggesting the result is not driven by one dataset being more 'complete' than the other. Together, these results indicate that the prompt-to-density mapping captures structure that is not narrowly tied to a single training distribution.

B.4 RICHER PROMPT REPRESENTATIONS

The main experiments condition the student on a single-token hidden state from a single model layer. Here we test whether this choice constitutes a fundamental performance ceiling by training additional students on two richer prompt representations: *concat3*, the concatenated final-token hidden states of the three deepest transformer layers, and *mean10*, the element-wise mean of final-token hidden states across the deepest ten layers. For each model and representation we sweep $H \in \{128, 256, 512, 1024\}$, $D \in \{3, 6\}$, and dropout $\in \{0.2, 0.4\}$ (16 configurations), with K=10 mixture components and PCA dimension 16 fixed. We report the mean test AUROC and mean distillation fidelity ρ_{fidelity} (Appendix B.1) across configurations, with standard deviations computed over the 16 runs.

Table A3 shows the results. Richer representations improve both AUROC and distillation fidelity for two of the three models (Llama 3.2 3B and Ministral 8B), indicating that the single-token/single-layer representation is a conservative engineering choice rather than a fundamental limit for these models. Qwen3 4B shows the opposite pattern: both rich variants underperform the single-layer baseline, with corresponding decreases in fidelity. This is consistent with overfitting at our training set size (N_{train}=4,000): increasing input dimensionality without scaling supervision can hurt generalization.

Notably, Qwen3 4B also exhibits the highest sensitivity to hyperparameter choice (std 0.031–0.035 vs. 0.015–0.027 for the other models), itself consistent with the distillation-fidelity gap identified in Appendix B.1. We conclude that richer prompt

Table A3: **Richer prompt representations.** Mean hallucination prediction AUROC and distillation fidelity over 16 student configurations per (model, representation) on TriviaQA. Δ denotes the change in mean AUROC relative to the single-layer baseline.

Model	Prompt rep.	AUROC	ρ_{fidelity}	Δ
Qwen3 4B	Single layer	0.660 ± 0.035	0.315	–
	Concat3	0.628 ± 0.031	0.196	-0.032
	Mean10	0.641 ± 0.033	0.251	-0.019
Llama 3.2 3B	Single layer	0.712 ± 0.015	0.418	–
	Concat3	0.727 ± 0.016	0.430	$+0.014$
	Mean10	0.722 ± 0.017	0.420	$+0.009$
Ministral 8B	Single layer	0.721 ± 0.027	0.597	–
	Concat3	0.739 ± 0.019	0.634	$+0.018$
	Mean10	0.724 ± 0.024	0.603	$+0.003$

representations are a non-trivial engineering axis whose benefit depends jointly on per-model regularization and training data quantity. A systematic exploration along the lines of Kramár et al. [2026] is a natural direction for future work.

C ABLATIONS

We analyze two design axes underlying SSD: the dimensionality of the semantic embedding space and the capacity of the student mixture model. All ablations are conducted using the TriviaQA dataset.

C.1 PCA DIMENSION

Standard sentence embeddings are high-dimensional (e.g., $d_z = 768$). Training an MDN to model a full covariance density in this space is data-intensive. We apply PCA to reduce the target space to $d_{pca} \in \{16, 32, 64, 128\}$.

This introduces a trade-off between semantic resolution and distillation fidelity: higher dimensions preserve more semantic detail (raising the teacher dispersion baseline), but make the prompt-to-density regression task harder for the student to learn with fixed data. As shown in Figure A2 and Table A4, while the detection performance of teacher dispersion improves monotonically with dimensionality, the student's performance (SSD) peaks at $d_{pca} = 16$. Beyond this point, the student fails to effectively distil the increasingly sparse density, leading to a drop in detection accuracy. For the OOD experiments we find PCA dimensions of 16 and 32 perform similarly, followed by a drop-off at higher dimensions. We fix $d_{pca} = 16$ for all main experiments bar the OOD task, where we set $d_{pca} = 32$.

We underline that the same training dataset size ($N_{train} = 4000$) was used in all PCA ablations; the learning task is therefore more difficult in higher dimensions, as models must learn more complex relationships using the same number of datapoints. For larger training datasets, we expect the SSD performance will continue to match or exceed the TD performance.

Table A4: Effect of PCA dimensionality on Hallucination prediction (SSD) and OOD Verification. Scores are averaged across the best student configuration for each model.

PCA Dim	Teacher Dispersion (TD)	Peak SSD AUROC	Peak OOD AUROC
16	0.699 ± 0.029	$\mathbf{0.708 \pm 0.026}$	0.965 ± 0.026
32	0.724 ± 0.027	0.706 ± 0.029	$\mathbf{0.966 \pm 0.035}$
64	0.747 ± 0.025	0.701 ± 0.037	0.960 ± 0.047
128	$\mathbf{0.757 \pm 0.025}$	0.703 ± 0.029	0.950 ± 0.055

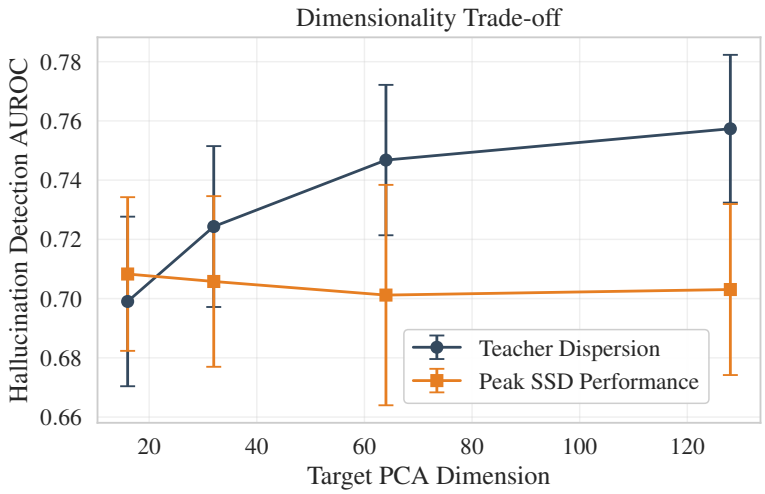

Figure A2: As PCA dimension increases, the semantic resolution improves and the teacher dispersion baseline rises. The distillation task however becomes harder, causing SSD performance to fall.

C.2 STUDENT CAPACITY

We investigate whether the performance gap on difficult models can be closed by increasing the student's capacity. We analyze two axes of complexity:

1. **Distributional Complexity:** We vary the number of mixture components $K \in \{1, 2, 5, 10\}$.

2. **Backbone Capacity:** We vary the MLP hidden dimension $H \in \{128, 256, 512, 1024\}$ and depth $D \in \{2, 3, 4, 6\}$.

Table A6 reveals that the optimal hyperparameters vary by base model. Aggregating results across all models (Table A5) reveals **capacity saturation**, where increasing the network width or depth yields negligible gains in mean AUROC. Figure A3 visualizes the full sweep across width, depth, and mixture components. This saturation indicates that performance is not limited by student expressivity, but by the information available in the prompt representation used for distillation. The student appears to extract most of the accessible uncertainty signal from this representation; further improvements would require richer inputs (e.g., attention-pooled representations) rather than larger students. Although we observe optimal performance for $K = 10$ mixture components, we opt not to increase this further given our set number of $S = 32$ samples; modelling more than ten unique semantic concepts in this setting is unrealistic and would likely damage the generalization ability of the trained student.

Table A5: Marginal impact of student capacity (Fixed PCA=16). Scores represent the mean AUROC across all runs with the specified parameter; standard deviations indicate variability across different base models and remaining hyperparameter configurations. We observe negligible performance gains from increasing complexity, indicating information saturation in the prompt representation.

(a) Components (K)

K	AUROC
1	0.648 ± 0.050
2	0.657 ± 0.052
5	0.660 ± 0.048
10	$\mathbf{0.666 \pm 0.042}$

(b) Width (H)

H	AUROC
128	$\mathbf{0.662 \pm 0.043}$
256	0.660 ± 0.046
512	0.660 ± 0.050
1024	0.649 ± 0.054

(c) Depth (D)

L	AUROC
2	0.657 ± 0.049
3	$\mathbf{0.660 \pm 0.047}$
4	0.658 ± 0.050
6	0.656 ± 0.048

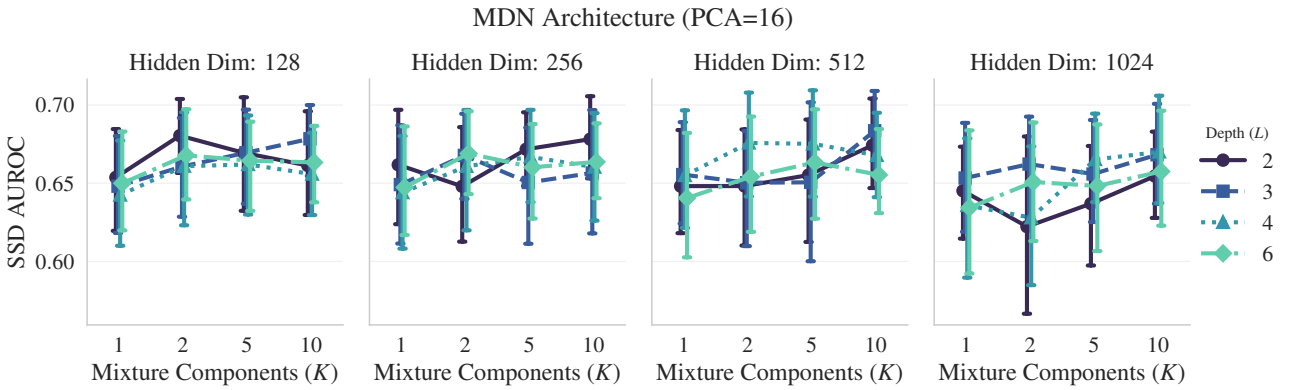

Figure A3: Student capacity ablation on the TriviaQA dataset, averaged across all seven models investigated.

D MODEL CONFIGURATION AND SELECTION

Here we summarize model characteristics and specific configuration choices used throughout our experiments. Table A6 reports base model accuracy on TriviaQA, the optimal probe layers identified for PCP and SEP, and the student architecture that achieved the highest SSD AUROC for each model.

Table A6: Model-specific configuration details on TriviaQA. 'Accuracy' denotes base model QA accuracy. 'Total' gives the number of transformer layers, while 'PCP' and 'SEP' report the selected probe layers. 'Best SSD Config' lists the student architecture (H=hidden size, D=depth, K=mixture components) achieving the highest SSD AUROC.

		Layers			Best SSD Config			
Model	Accuracy	Total	PCP	SEP	H	D	K	AUROC
Qwen3-8B	0.620	36	32	24	128	6	10	0.677
Qwen3-4B	0.565	36	18	20	256	2	10	0.699
Llama 3.1 8B	0.700	32	20	16	256	4	2	0.687
Llama 3.2 3B	0.585	28	15	19	512	3	5	0.737
Ministral 8B	0.656	36	33	34	1024	6	1	0.748
SmolLM3 3B	0.567	36	20	19	512	6	5	0.710
Gemma 3 4B	0.532	34	22	19	512	2	10	0.700

D.1 LAYER SELECTION PROTOCOL

To select the student input layer ℓ^*, we follow Kossen et al. [2024] and train linear probes on all layers to predict binarized Semantic Entropy. We choose the layer that maximizes validation accuracy on this proxy task. This heuristic assumes that layers most predictive of scalar dispersion also contain the richest signal for learning the full semantic density. In practice, the selected layers tend to lie in deeper parts of the network (Table A6).

E COMPUTATIONAL COMPLEXITY ANALYSIS

The primary motivation for Semantic Self-Distillation is to enable generic uncertainty estimation in AR models without sampling. Here we compare the asymptotic complexity of our method against sampling-based baselines. Let T_{pre} be the prompt length, T_{gen} be the generation length, S be the number of samples, and C_{LLM} be the cost of a single transformer forward pass per token.

Sampling-based Methods (SE, TD). Computing Semantic Entropy [Farquhar et al., 2024] requires S stochastic sequences per prompt at inference time (typically $S \in [5, 20]$). The cost is dominated by the autoregressive decoding steps:

$$\mathcal{O}_{\text{SE}} \approx S \times (T_{\text{pre}} + T_{\text{gen}}) \times C_{\text{LLM}} + \mathcal{O}_{\text{NLI}}(S^2)$$

where $\mathcal{O}_{\text{NLI}}$ represents the quadratic cost of comparing all sample pairs using an entailment model. This linear scaling with S introduces latency that is unacceptable in many applications, for example in interactive agentic systems.

Scalar Probes (PCP, SEP). Scalar probes operate on prompt representations extracted from a single forward pass of the base model. The probe itself is a lightweight classifier or regressor applied to a cached hidden state, without requiring additional sampling or decoding:

$$\mathcal{O}_{\text{SEP}} \approx 1 \times T_{\text{pre}} \times C_{\text{LLM}} + C_{\text{MLP}}$$

While efficient, these methods predict a compressed scalar statistic and discard the geometric structure of the model's semantic output distribution.

Semantic Self-Distillation (SSD). SSD matches the inference complexity of scalar probes while retaining the distributional richness of sampling methods. The MDN head introduces a slightly larger parameter set than a linear probe, but this cost (C_{MDN}) is negligible compared to the transformer backbone. Crucially, the entropy calculation is analytical and does not require sampling:

$$\mathcal{O}_{\text{SSD}} \approx 1 \times T_{\text{pre}} \times C_{\text{LLM}} + C_{\text{MDN}}$$

Thus, SSD effectively moves the computational burden from *inference time* (where latency matters) to *training time*, enabling $O(1)$ uncertainty estimation during deployment.

F FULL HALLUCINATION PREDICTION RESULTS

We report full per-model hallucination detection results corresponding to the summaries in Tables 2 and 1. These tables provide AUROC and AUPRC metrics for all methods across datasets.

Table A7: **Hallucination Prediction AUROC.** Values report mean AUROC (%) over 1,000 bootstrap resamples of the test set, with the bootstrap standard deviation shown as a subscript. We mark the best method per dataset and model in bold.

	TriviaQA					MMLU				
Model	PCP	SEP	SE	TD	SSD	PCP	SEP	SE	TD	SSD
Qwen3 8B	$76.9_{1.5}$	$76.2_{1.5}$	$\mathbf{81.6_{1.4}}$	$68.3_{1.6}$	$67.7_{1.8}$	$\mathbf{69.8_{1.6}}$	$59.5_{1.8}$	$63.7_{1.6}$	$63.5_{1.7}$	$57.1_{1.8}$
Qwen3 4B	$\mathbf{80.2_{1.4}}$	$78.3_{1.4}$	$78.6_{1.4}$	$73.2_{1.6}$	$69.9_{1.6}$	$\mathbf{67.9_{1.7}}$	$64.3_{1.8}$	$60.5_{1.6}$	$63.1_{1.8}$	$64.9_{1.7}$
Llama 3.1 8B	$77.0_{1.7}$	$74.4_{1.7}$	$\mathbf{85.1_{1.3}}$	$69.6_{1.7}$	$68.7_{1.9}$	$\mathbf{71.5_{1.6}}$	$64.9_{1.7}$	$70.6_{1.6}$	$69.0_{1.6}$	$63.8_{1.7}$
Llama 3.2 3B	$76.7_{1.5}$	$75.7_{1.6}$	$\mathbf{80.6_{1.4}}$	$66.6_{1.7}$	$73.7_{1.6}$	$66.0_{1.9}$	$64.4_{1.8}$	$\mathbf{70.1_{1.7}}$	$67.1_{1.8}$	$63.8_{1.7}$
Ministral 8B	$74.0_{1.7}$	$77.1_{1.6}$	$\mathbf{80.5_{1.5}}$	$74.3_{1.7}$	$74.9_{1.6}$	$67.0_{1.7}$	$62.8_{1.8}$	$\mathbf{69.3_{1.7}}$	$65.1_{1.8}$	$63.6_{1.8}$
SmolLM3 3B	$79.6_{1.4}$	$78.7_{1.4}$	$\mathbf{83.1_{1.2}}$	$69.9_{1.6}$	$71.1_{1.6}$	$67.0_{1.7}$	$66.3_{1.8}$	$\mathbf{68.8_{1.7}}$	$62.3_{1.8}$	$62.0_{1.8}$
Gemma 3 4B	$\mathbf{75.2_{1.6}}$	$70.6_{1.6}$	$73.1_{1.5}$	$67.6_{1.6}$	$70.0_{1.7}$	$\mathbf{66.2_{1.8}}$	$57.8_{1.8}$	$61.3_{1.6}$	$62.8_{1.7}$	$60.7_{1.8}$

Table A8: **Hallucination Prediction AUPRC.** Values report mean AUPRC (%) over 1,000 bootstrap resamples of the test set, with the bootstrap standard deviation shown as a subscript. We mark the best method per dataset and model in bold.

	TriviaQA					MMLU				
Model	PCP	SEP	SE	TD	SSD	PCP	SEP	SE	TD	SSD
Qwen3 8B	$64.8_{2.6}$	$64.9_{2.7}$	$\mathbf{75.0_{2.1}}$	$52.4_{2.6}$	$56.9_{2.8}$	$\mathbf{66.9_{2.3}}$	$56.3_{2.3}$	$60.3_{2.2}$	$61.9_{2.3}$	$55.9_{2.3}$
Qwen3 4B	$\mathbf{74.3_{2.4}}$	$70.8_{2.5}$	$73.4_{2.2}$	$59.7_{2.4}$	$61.9_{2.5}$	$\mathbf{60.1_{2.5}}$	$57.6_{2.5}$	$53.4_{2.3}$	$58.5_{2.4}$	$59.4_{2.5}$
Llama 3.1 8B	$59.7_{3.1}$	$54.0_{3.0}$	$\mathbf{73.4_{2.7}}$	$44.4_{2.7}$	$51.9_{2.9}$	$\mathbf{72.7_{2.1}}$	$62.6_{2.3}$	$70.5_{2.1}$	$70.9_{2.1}$	$62.7_{2.4}$
Llama 3.2 3B	$68.9_{2.3}$	$65.4_{2.6}$	$\mathbf{74.1_{2.2}}$	$55.8_{2.6}$	$65.5_{2.5}$	$76.4_{1.9}$	$74.5_{2.0}$	$\mathbf{78.4_{1.8}}$	$74.0_{2.0}$	$74.5_{1.9}$
Ministral 8B	$59.2_{3.0}$	$62.9_{2.8}$	$\mathbf{69.5_{2.8}}$	$57.9_{2.8}$	$59.9_{2.8}$	$72.9_{2.2}$	$70.7_{2.1}$	$\mathbf{76.0_{1.9}}$	$71.4_{2.1}$	$71.7_{2.1}$
SmolLM3 3B	$75.0_{2.1}$	$73.1_{2.2}$	$\mathbf{78.8_{1.9}}$	$57.1_{2.5}$	$62.4_{2.5}$	$67.7_{2.3}$	$67.6_{2.2}$	$\mathbf{70.5_{2.2}}$	$64.6_{2.2}$	$63.0_{2.3}$
Gemma 3 4B	$\mathbf{69.2_{2.5}}$	$64.7_{2.5}$	$69.2_{2.2}$	$63.1_{2.3}$	$66.1_{2.5}$	$\mathbf{73.5_{2.1}}$	$63.8_{2.1}$	$67.6_{1.9}$	$71.1_{2.0}$	$69.2_{2.1}$