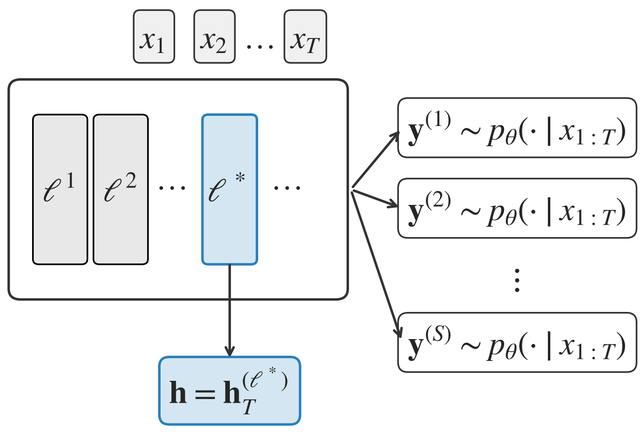
Semantic distributions can be predicted from language model activations

Semantic Self-Distillation for Language Model Uncertainty

Edward Phillips · Sean Wu · Fredrik K. Gustafsson · Boyan Gao · David A. Clifton — University of Oxford

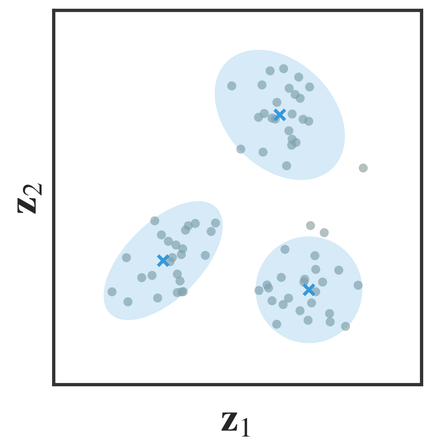
UNIVERSITY OF OXFORD

1 **Train** on the model's own samples



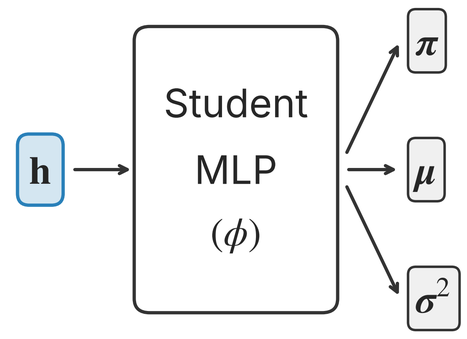
Sample S answers per prompt; keep the prompt's hidden state h

embed Φ(y)



Embed answers in semantic space

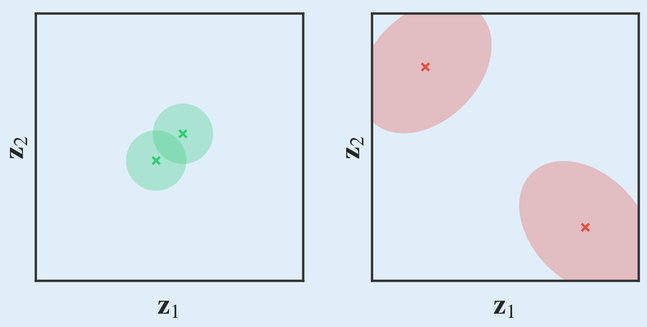
distil



Student MDN learns the mixture $q_\varphi(z \mid h)$ from h alone

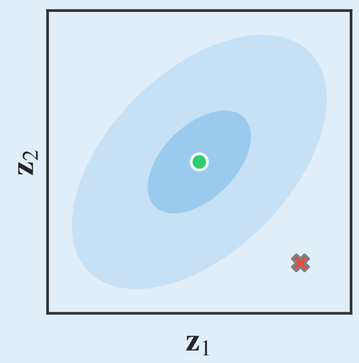
2 **Inference** with one MDN pass

Hallucination risk, before generating



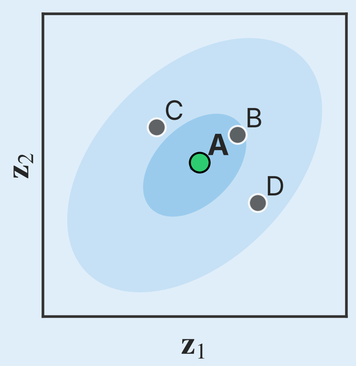
Entropy of q_φ: concentrated → likely correct; dispersed → likely hallucination

Answer verification, after generating



Likelihood under q_φ flags answers that don't fit the prompt

Answer selection

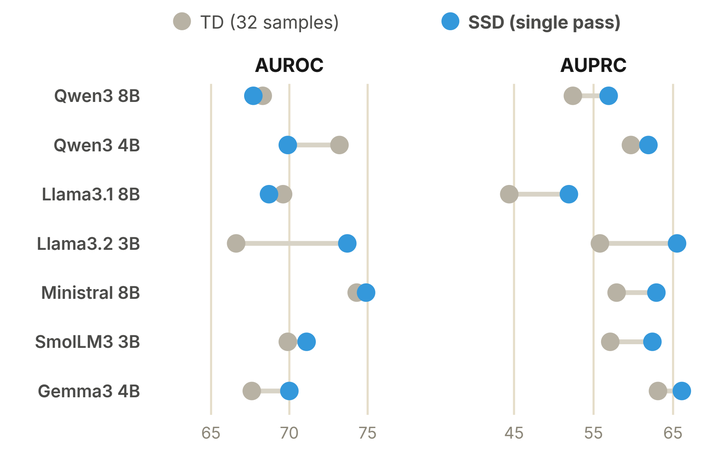


Rank candidate answers by density — no extra generation

MAIN RESULTS

Hallucination detection

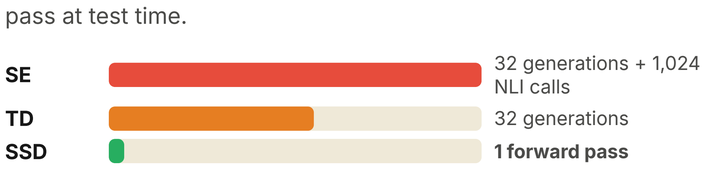
We distil students for **seven open-weight LLMs** (3B–8B) on TriviaQA and MMLU, sampling 32 answers per training prompt. At test time the student sees only the prompt's hidden state: the **entropy of its predicted density** is a pre-generation hallucination risk score.



TriviaQA, teacher dispersion (grey) vs SSD (blue). The student matches the sampled teacher on AUROC — better on 4 of 7 models — and improves AUPRC on **all seven**.

Latency

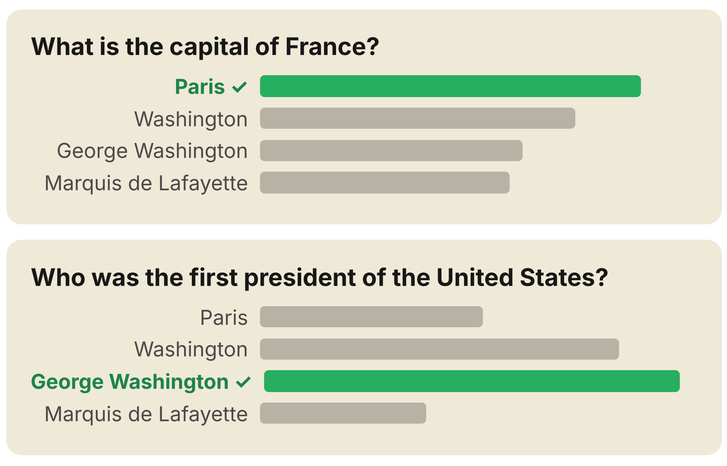
Sampling-based measures require multiple stochastic completions at inference time — semantic entropy needs 32 generations plus 1,024 pairwise NLI comparisons per prompt — rendering them **prohibitively expensive in latency-critical applications**. SSD amortises sampling into training and requires a single forward pass at test time.



BEYOND ENTROPY

Answer verification

Because SSD predicts a **full semantic distribution**, we can assess the quality of any candidate answer by measuring its likelihood under the predicted density.



Log-likelihood of four candidate answers under different $q_\varphi(z \mid h)$. Changing only the prompt reshapes the distribution and moves mass to the semantically appropriate answer.

Out-of-domain detection

The likelihood is **especially strong as an out-of-domain detector**: swapped or corrupted answers land far from the predicted density and can be flagged near-perfectly, a practical content guard for caches and retrieval pipelines. The signal survives distribution shift: students trained on TriviaQA still flag mismatches on MMLU, and vice versa (AUROC 0.75–0.84), with no retraining.

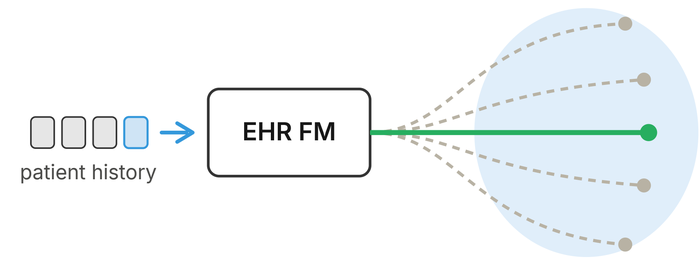
0.95+
AUROC flagging answers that belong to a **different prompt**, for most base models

≈0.80
retained by students evaluated on a dataset they were **never trained on**

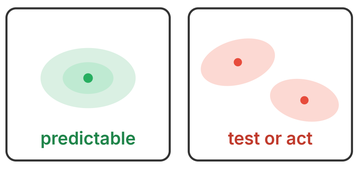
FUTURE WORK

Healthcare foundation models

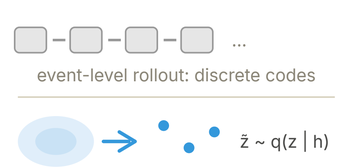
SSD can be extended to **any sequence model** whose outputs admit meaningful latent representations. By distilling a sampled distribution over this latent space, the SSD density supplies a **sampling-free approximation to the model's belief over high-level outcomes**.



EHR foundation models predict patient futures as a sequence of discrete clinical 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.



Uncertainty over long-term outcomes: the entropy of the predicted density says how settled a patient's outlook is.



Forecasting latent patient states: sampling $\tilde{z} \sim q_\varphi(z \mid h)$ forecasts directly in this higher-level space.

A distilled density over patient futures would give clinical models the same primitives we demonstrate for language: **pre-decision risk estimates** and **verification of candidate trajectories** against patient context.

ed-phillips.github.io · edward.phillips@eng.ox.ac.uk