ICML 2026 | MECHANISTIC INTERPRETABILITY WORKSHOP

A tiny probe reads off *which* reasoning steps matter — from a single forward pass, replacing ∼100 rollouts per sentence.

Does the Model Know Which Steps Matter? Probing Causal Importance in Chain-of-Thought Reasoning

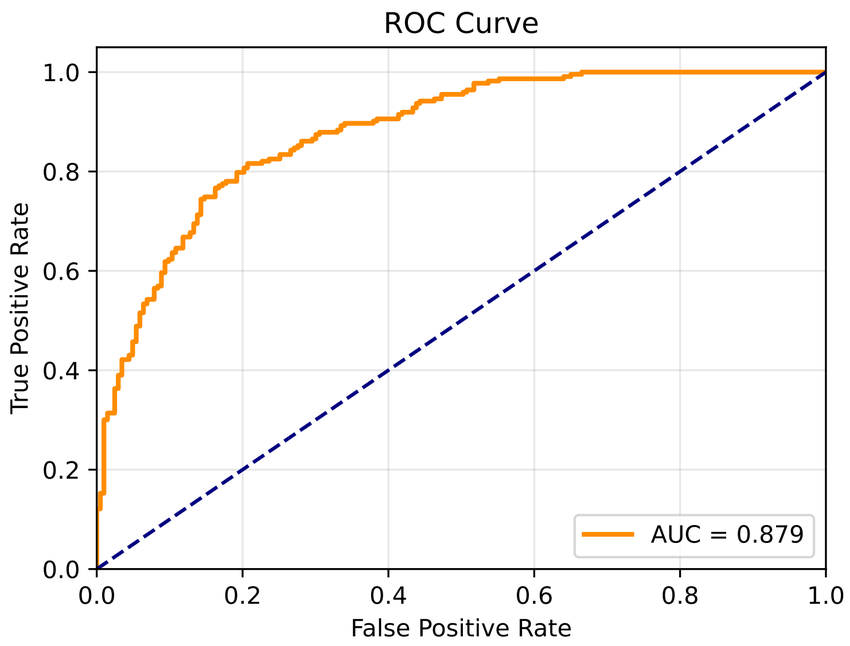
Isotta Magistrali **Mrinmaya Sachan** **Alessandro Stolfo** Department of Computer Science, ETH Zürich

The problem. Chain-of-thought reasoning is powerful but expensive, so we want to know which sentences actually *drive* the answer. Resampling-based *importance* metrics answer this, but each sentence needs ∼100 stochastic rollouts, making the very tool meant to streamline CoT too costly to run at scale or in real time.

Results

1. ROC curve.

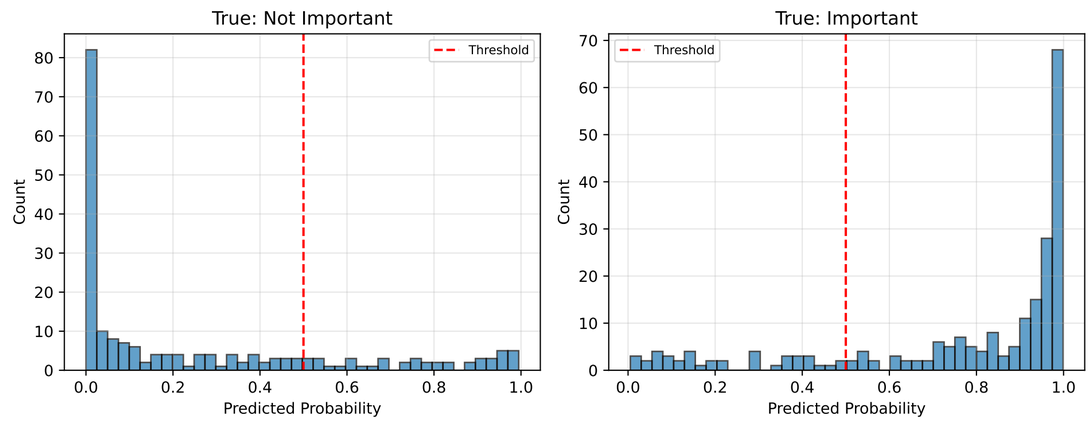
A small MLP probe on a *single* hidden layer separates important from unimportant steps far above chance — AUC up to **0.88**.



ROC — DeepSeek-R1-Llama-8B, mid layer, MATH.

2. Predicted probability by true label.

Predicted-probability mass splits cleanly by true label and importance is further *uncorrelated* with sentence position, confirming a statistically significant discriminative signal for importance.



Probe scores for truly (un)important sentences, same model.

80.8% best probe accuracy

∼100× rollouts saved / sentence

4 model families

2 reasoning benchmarks

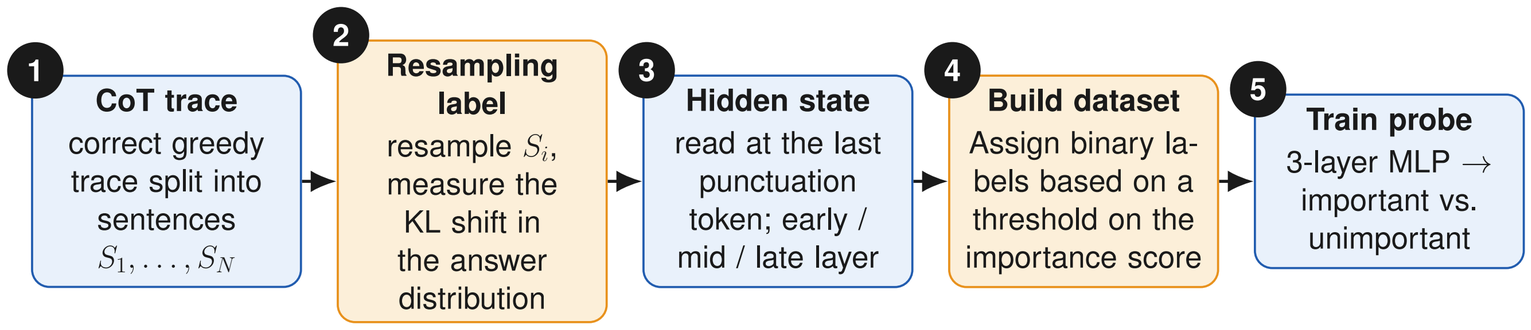
3. The signal holds across model families, scales, and datasets.

Model	Dataset	Resampling acc.	Counterfactual acc.
DeepSeek-R1-Qwen-1.5B	GSM8K	.808	.750
	MATH	.750	.730
DeepSeek-R1-Qwen-7B	GSM8K	.726	.728
	MATH	.762	.800
DeepSeek-R1-Llama-8B	GSM8K	.771	.754
	MATH	.801	.744
Gemma-3-4B	GSM8K	.771	.700
	MATH	.755	.710

Best single-layer accuracy per setting (early / mid / late).

Method

Resampling importance **Counterfactual importance**



$\text{Importance}_{\text{res}}(S_i) = D_{\text{KL}}\big[p(A \mid S_i') \,\|\, p(A \mid S_i)\big]$; the counterfactual variant keeps only resamples $S_i' \not\approx S_i$ that are semantically distinct. Steps 1–2 (expensive) run *once* to build training labels; at test time only steps 3–5 are needed.

Limitations & next steps. Accuracy plateaus near ∼80%: KL labels come from finite, noisy rollouts, and a probe reading one layer may miss signal spread across layers. Evidence so far is on mathematical reasoning. Next: validate against direct ablations, scale to longer traces and larger models, and extend to code and agentic settings, where cheap importance estimation matters most for safety monitoring and adaptive compute.