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

Isotta Magistrali [1] **Mrinmaya Sachan** [1] **Alessandro Stolfo** [1]

Abstract

Chain-of-thought (CoT) reasoning drives modern large language models but carries substantial computational cost, motivating methods that identify which reasoning steps are truly load-bearing. Recent work quantifies this through resampling-based importance metrics, but computing them requires hundreds of rollouts per prompt, making the tool meant to streamline CoT itself too expensive to deploy at scale. We address this bottleneck by providing evidence that sentence-level importance can be predicted from a model's internal activations: lightweight probes trained on hidden representations can identify important reasoning steps with reliable accuracy. Not only does this reduce inference cost by many orders of magnitude, but it also carries a deeper implication: the decodability of importance from internal activations suggests that these behavioral metrics correspond to information encoded in the model's hidden states as it generates the trace, opening the door to CoT interpretability at the scale that resampling cannot reach, including real-time monitoring, adaptive compute allocation, and large-scale mechanistic analysis.

1. Introduction

Chain-of-thought (CoT) reasoning, in which language models generate intermediate steps before committing to a final answer, underpins recent advances in complex problem-solving (Wei et al., 2022; Guo et al., 2025). The verbalized trace also appears to offer a legible window into how the model reached its answer, a property now central to proposals for monitoring model behavior for alignment and safety (Korbak et al., 2025). Yet a growing body of work shows that this transparency is partial: CoTs rationalize answers driven by prompt biases without acknowledging them (Turpin et al., 2023), fail to verbalize outcome-changing hints even in frontier reasoning models (Chen et al., 2025), and exhibit silent restoration errors and illogical shortcuts on realistic prompts (Arcuschin et al., 2025; Barez et al., 2025). Identifying which parts of a CoT actually drive the answer is therefore a prerequisite both for efficient test-time compute and for monitoring-based safety techniques.

These concerns have motivated work on attributing causal importance to reasoning steps. Causal mediation analyses show that models often do not make genuine use of their intermediate steps (Paul et al., 2024), while Bogdan et al. (2025) introduce *thought anchors* and quantify sentence-level impact through two resampling-based metrics: *resampling importance*, the KL divergence between answer distributions when a sentence is replaced by a free continuation, and *counterfactual importance*, which additionally requires the replacement to be semantically distinct. They find that planning and uncertainty-management statements exert larger causal impact than local computational steps, while surface-salient content such as self-preservation statements in agentic settings can have minimal downstream effect. Such methods, however, require on the order of 100 rollouts per sentence, ruling out real-time monitoring and large-scale analysis.

We present a probing framework that predicts sentence-level importance directly from internal activations, building on prior work showing that hidden states encode rich information recoverable by simple classifiers: truthfulness (Marks & Tegmark, 2024; Azaria & Mitchell, 2023), spatial and temporal coordinates (Gurnee & Tegmark, 2024), hallucination signals (Orgad et al., 2025; Sun et al., 2025; Chen et al., 2024), and even the marginal value of additional computation (Damani et al., 2025). We train lightweight probes on sentence-level representations to predict the two importance metrics of Bogdan et al. (2025), finding that internal activations carry enough signal to distinguish important from unimportant steps with non-trivial accuracy and at orders-of-magnitude lower cost than resampling. Beyond

[1]Department of Computer Science, ETH Zürich, Zürich, Switzerland. Correspondence to: Isotta Magistrali <imagistrali@student.ethz.ch>, Mrinmaya Sachan <mrinmaya.sachan@inf.ethz.ch>, Alessandro Stolfo <alessandro.stolfo@inf.ethz.ch>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. *Copyright 2026 by the author(s).*

efficiency, the decodability of importance from hidden states suggests these metrics correspond to information encoded in the model's internal computation, and that recovering them cheaply opens the door to CoT interpretability at scale.

2. Methodology

2.1. Importance Metrics

Following Bogdan et al. (2025), we measure how much each sentence in a reasoning trace causally contributes to the model's final answer using two complementary resampling-based metrics. Given a trace $T = (S_1, \ldots, S_N)$, we replace sentence S_i with an alternative continuation S'_i sampled from the model, and ask how much the distribution over final answers shifts as a result. A large shift indicates that S_i played a decisive role; a small shift indicates the model would have reached the same conclusion regardless.

Resampling importance. The simplest version of this procedure samples S'_i freely from the model and compares the resulting answer distribution to the original:

$$\text{importance}_{\text{res}}(S_i) = D_{KL}\Big[p(A \mid S'_i) \,\Big\|\, p(A \mid S_i)\Big], \quad (1)$$

where A is the final answer of the trace and $p(A \mid \cdot)$ the empirical answer distribution estimated from rollouts. As in Bogdan et al. (2025), the intervention is trajectory-level: after S_i is replaced, the model regenerates the remainder of the trace freely, so the metric quantifies how strongly position i redirects the downstream trajectory rather than the effect of substituting S_i with the suffix held fixed. This captures the overall sensitivity of the answer to perturbations at position i, but has a known limitation: unconstrained resamples frequently paraphrase S_i rather than expressing genuinely different content. As a result, the metric conflates two distinct effects, the impact of changing *what* the sentence says, and the noise introduced by changing *how* it is said.

Counterfactual importance. To isolate the contribution of content, we restrict S'_i to alternatives that are semantically distinct from the original sentence:

$$\text{importance}_{\text{cf}}(S_i) = D_{KL}\Big[p(A \mid S'_i,\, S'_i \not\approx S_i) \,\Big\|\, p(A \mid S_i)\Big], \quad (2)$$

where $S'_i \not\approx S_i$ denotes that the alternative differs from the original in meaning, not just in wording. Filtering out near-paraphrases yields a cleaner signal of which sentences genuinely steer the model toward its final answer, rather than which positions are merely sensitive to rewording. Implementation details are deferred to Appendix A.1.

2.2. Probing Setup

For each sentence S_i in a reasoning trace, we extract hidden representations from three transformer layers, one early, one middle, and one late, taken at the last punctuation token of the sentence. This is the last token at which the model has integrated the full content of S_i into its residual stream, making it a natural site for any sentence-level summary representation. These representations serve as inputs to lightweight probes trained to predict the importance metrics described above. We frame the task as binary classification: continuous importance scores are thresholded to produce labels, with sentences whose resampling/counterfactual importance exceeds the threshold marked as important ($y = 1$) and the rest as unimportant ($y = 0$). We restrict our analysis to *correct greedy traces*, those in which the model's greedy decoding produces the correct final answer, since in incorrect traces high-importance sentences may reflect errors rather than genuine reasoning anchors, mixing two qualitatively different phenomena that we believe are best studied separately. Within correct traces, importance thus measures how necessary a sentence is for remaining on a path to the correct answer.

3. Experiments

We evaluate on two widely used mathematical reasoning benchmarks: GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). The two differ substantially in difficulty: GSM8K consists of grade-school word problems solvable in a few reasoning steps, while MATH draws from competition-level problems that elicit longer chains with more planning and case analysis. Together they let us check whether the importance signal is recoverable both in short, procedural traces and in longer, structurally varied ones. To ensure a consistent evaluation protocol, we restrict MATH to problems whose ground-truth answer can be cast to a floating-point number, excluding symbolic answers (e.g. unsimplified fractions, polynomial expressions) that would otherwise require custom parsing and equivalence-checking logic. GSM8K answers are numeric by construction and require no such filtering. The resulting subset preserves the difficulty profile of MATH while allowing us to score both benchmarks with a single numerical-match criterion.

3.1. Classification

After removing outliers, we binarize the continuous importance scores by thresholding at the elbow of their empirical distribution, where the bulk of low-importance sentences ends and a heavy tail begins, yielding a principled split between routine steps and sentences that genuinely shift the answer distribution (see Appendix C Figures 3, 7, 11, 15; the outlier rule and numeric thresholds are given in Appendix A.1). Table 1 reports classification accuracy across

all models, datasets, and probe layers. Figures 1 and 2 show a representative ROC curve and predicted-probability distribution of resampling importance for DeepSeek-R1-Distill-Llama-8B at the mid layer; plots for all (dataset, score, model, layer) configurations are deferred to Appendix C.

Table 1. Binary classification accuracy for predicting sentence-level importance from hidden-state representations of one transformer layer (early/mid/late; per-model layer indices in Appendix A.2). Accuracy ranges from .60 to .81 and exceeds the .50 balanced-class baseline in all 48 configurations; each entry reflects a single training run (see Appendix B).

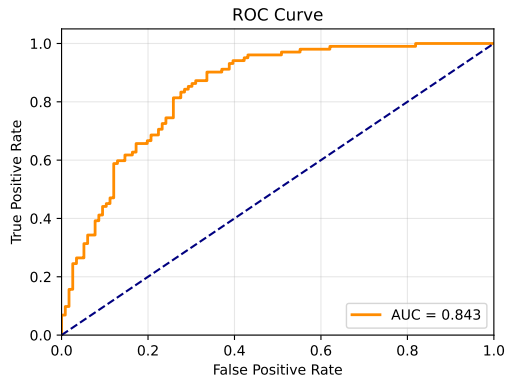
		Resampling			Counterfactual		
Model	**Dataset**	early	mid	late	early	mid	late
DeepSeek-R1-Distill-Qwen-1.5B	gsm8k	.790	.808	.770	.750	.750	.710
	MATH	.750	.732	.741	.730	.708	.680
DeepSeek-R1-Distill-Qwen-7B	gsm8k	.720	.713	.726	.728	.711	.717
	MATH	.746	.762	.738	.800	.786	.747
DeepSeek-R1-Distill-Llama-8B	gsm8k	.762	.771	.762	.754	.740	.736
	MATH	.796	.801	.775	.733	.730	.744
gemma-3-4b	gsm8k	.771	.667	.646	.604	.700	.623
	MATH	.755	.622	.748	.645	.694	.710

3.2. Ruling out the position confound

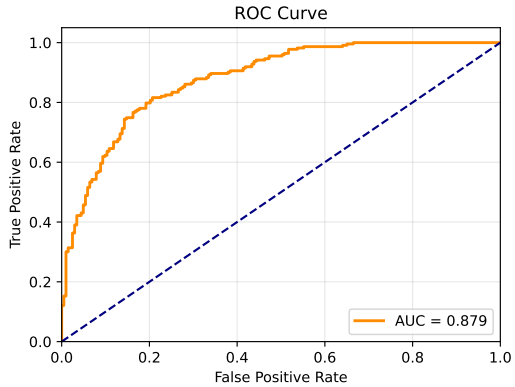
A potential confound in this setting is that sentence position within the trace is negatively correlated with importance: earlier sentences, particularly planning steps, tend to carry higher causal impact than later ones. This matters because position is itself trivially decodable from a hidden state, so a probe could in principle achieve part of its accuracy simply by reading off how far through the trace it is, without recovering anything about importance per se. Relative position and importance are indeed correlated in our experiments (Pearson $|r|$ up to 0.37; Appendix C Figures 6, 10, 14, 18), so we test the confound directly: for every (model, dataset, metric) configuration we train probes on relative sentence position alone, under a protocol identical to the activation probes' (same labels, class balancing, and train–test split). Table 2 reports the results. The best activation-probe accuracy of Table 1 exceeds the stronger of the two position-only baselines in 14 of 16 configurations, by 4–17 accuracy points (median ∼12); in the remaining two, which are the two smallest configurations (Gemma-3-4B on GSM8K, 48–53 test sentences), the two probes are statistically indistinguishable. Position is thus a real but partial predictor. This is expected, since planning sentences genuinely occur early, so part of the correlation is the phenomenon itself; most of the probes' accuracy reflects information beyond position.

4. Discussion

Sentence-level causal importance, as defined by the resampling-based metrics of Bogdan et al. (2025), can be decoded from internal activations with substantially above-

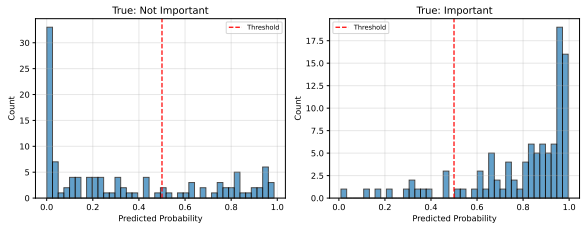


(a) GSM8K

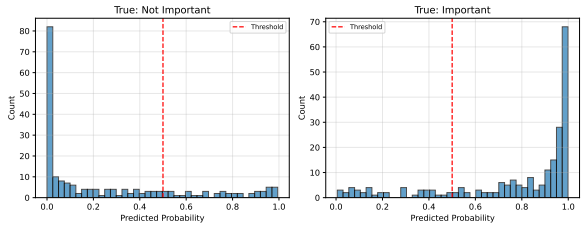


(b) MATH

Figure 1. ROC curves for resampling importance classification using mid-layer activations of DeepSeek-R1-Distill-Llama-8B. AUC above 0.5 on both datasets confirms a consistent signal in activations for identifying important steps.



(a) GSM8K



(b) MATH

Figure 2. Predicted probability distributions for important and unimportant sentences for resampling importance, using mid-layer activations of DeepSeek-R1-Distill-Llama-8B. Separation between the two distributions confirms that internal activations carry a discriminative signal for importance.

Table 2. Accuracy of probes trained on relative sentence position alone (logistic regression and MLP), under the same labels, class balancing, and train–test split as the activation probes; the "act." columns repeat the best activation-probe accuracy from Table 1 for comparison.

		Resampling			Counterfactual		
Model	**Dataset**	LR	MLP	act.	LR	MLP	act.
DeepSeek-R1-Distill-Qwen-1.5B	gsm8k	.686	.689	**.808**	.664	.664	**.750**
	MATH	.595	.608	**.750**	.658	.658	**.730**
DeepSeek-R1-Distill-Qwen-7B	gsm8k	.688	.682	**.726**	.578	.595	**.728**
	MATH	.594	.580	**.762**	.636	.642	**.800**
DeepSeek-R1-Distill-Llama-8B	gsm8k	.670	.679	**.771**	.653	.656	**.754**
	MATH	.643	.636	**.801**	.611	.627	**.744**
gemma-3-4b	gsm8k	**.792**	.771	.771	.698	.679	**.700**
	MATH	.657	.657	**.755**	.573	.581	**.710**

chance accuracy across multiple model families, scales, and datasets. Since these metrics are defined extrinsically by perturbing the trace and observing how the answer distribution shifts, it was a priori unclear whether they correspond to anything the model itself represents. That a lightweight probe recovers them with up to $\sim 80\%$ accuracy from a single forward pass suggests they track information encoded in the model's hidden states, and that the model in some sense "knows" which of its own steps are load-bearing as it generates them; confirming the causal role of this information through direct ablation is a natural next step, which we leave to future work (Appendix B). Estimating importance from a single forward pass, rather than computing it via ~ 100 rollouts per sentence, reduces cost by several orders of magnitude, and this is what makes CoT interpretability at scale realistic. Analyses that are simply out of reach for resampling-based methods, including real-time CoT monitoring (Korbak et al., 2025), adaptive inference-time compute allocation, and large-scale mechanistic studies across thousands of traces, become practically accessible.

Table 1 surfaces a few patterns. Mid-layer activations are often the strongest readout site for resampling importance, though not uniformly (e.g., DeepSeek-R1-Distill-Qwen-7B on GSM8K peaks at the late layer), while counterfactual accuracy degrades slightly at the latest layer. Probe accuracy is comparable across the 1.5B–8B DeepSeek distillations with no clear scale trend. Gemma-3-4B is competitive at the early layer and weaker at mid and late depths, suggesting its importance signal concentrates at an earlier relative depth. GSM8K and MATH yield similar accuracies despite MATH's longer and more varied traces, indicating the signal is not specific to short, procedural reasoning.

Several directions naturally extend this work; we discuss them in detail in Appendix B. In brief, the most promising avenues include validating predictions against direct ablation experiments, scaling to larger models and longer reasoning traces, and extending beyond mathematical reasoning to code generation, multi-step QA, and agentic settings, where cheap importance estimation has the most consequential applications for safety monitoring and compute allocation.

5. Related Work

5.1. Chain-of-Thought Faithfulness and Attribution

Chain-of-thought (CoT) traces are often treated as a natural-language window into a model's reasoning, but a growing body of evidence questions whether the verbalized trace reflects how the model actually arrives at its answer. Turpin et al. (2023) showed that subtle prompt biases can shift predictions in up to 36% of cases without ever being acknowledged in the CoT, which instead rationalizes whatever answer was chosen. Chen et al. (2025) extended this to state-of-the-art reasoning models and found that CoTs verbalize the influence of prompt hints in under 20% of cases on average, a gap that outcome-based reinforcement learning improves but does not close. On realistic prompts, Arcuschin et al. (2025) documented non-negligible rates of implicit post-hoc rationalization, silent restoration errors, and illogical shortcuts, and Fayyaz et al. (2024) showed that prompting-based self-explanations are systematically less faithful than attribution-based ones. Synthesizing this evidence, Barez et al. (2025) argue that verbalized chains are neither necessary nor sufficient for trustworthy interpretability, motivating recent work on CoT monitorability for AI safety, which Korbak et al. (2025) characterize as a promising but fragile affordance. A complementary line of work attributes causal importance to individual reasoning steps. Paul et al. (2024) use causal mediation analysis on twelve LLMs and find that models do not reliably use their own intermediate steps, motivating training-time interventions that make reasoning more load-bearing. Bogdan et al. (2025) introduced *thought anchors* and quantified sentence-level causal impact through two resampling-based metrics: *resampling importance*, the KL divergence between answer distributions when a sentence is resampled freely, and *counterfactual importance*, which additionally constrains the resample to be semantically distinct. They show that planning and uncertainty-management steps exert substantially greater causal influence than local computational steps, while surface-salient content such as self-preservation statements in agentic-misalignment scenarios can have surprisingly low impact. These methods provide principled attribution but require many rollouts per sentence, making them impractical at scale. Our work addresses this limitation by showing that sentence-level importance can be predicted from internal activations in a single forward pass.

5.2. Probing Internal Representations

Probing has long been used to understand what information is encoded in language model representations (Belinkov & Glass, 2019; Hewitt & Manning, 2019). Recent work has shown that surprisingly complex properties can be lin-

early decoded from hidden states, including the truthfulness of factual statements (Marks & Tegmark, 2024; Azaria & Mitchell, 2023), spatial and temporal coordinates of entities (Gurnee & Tegmark, 2024), and indicators of factual correctness or hallucination (Sun et al., 2025; Orgad et al., 2025; Chen et al., 2024), with Orgad et al. (2025) highlighting that the correct answer is sometimes recoverable from hidden states even when the generated response is wrong. Within the reasoning setting, Venhoff et al. (2025) distinguishes sentences that retrieve information from those that perform deduction, and Damani et al. (2025) shows that MLP probes can predict marginal reward gains from additional computation. However, these targets are typically aggregated or statement-level. We instead use simple probes to extract a fine-grained signal, the causal importance of individual sentences in a CoT, a challenging target, since importance depends on how the answer distribution would shift under semantically different continuations. Concurrent work also probes reasoning-model activations for related trace-level signals: Boppana et al. (2026) track when the final answer becomes decodable during the CoT and use this for probe-guided early exit, JeaniceK & Taraz (2026) train probes to predict a judgment-leaning signal in legal reasoning, and Nikankin et al. (2026) likewise probe activations for reasoning-step importance, finding that activations identify important steps better than the tokens of the chain itself and that different models agree on which steps matter. All three works are concurrent with ours, and none targets the same quantity: to our knowledge, ours is the first to probe directly for the resampling-based causal importance metrics of Bogdan et al. (2025). This choice of target matters because each probe prediction amortizes a behaviorally defined quantity that would otherwise require $\sim$100 rollouts to measure, and can therefore be validated against that behavioral ground truth, with the prospect of standing in for it wherever thought-anchor analyses apply.

6. Conclusion

We present a probing framework that predicts sentence-level importance directly from internal activations. Prior work has shown that hidden states encode surprisingly rich information recoverable by simple classifiers: truthfulness of factual statements (Marks & Tegmark, 2024; Azaria & Mitchell, 2023), spatial and temporal coordinates (Gurnee & Tegmark, 2024), hallucination signals (Orgad et al., 2025; Sun et al., 2025; Chen et al., 2024), and even the marginal value of additional computation (Damani et al., 2025). Building on this, we train lightweight probes on sentence-level representations to predict the two importance metrics of Bogdan et al. (2025). We find that internal activations carry enough signal to distinguish important from unimportant reasoning steps with non-trivial accuracy, and doing so at orders-of-magnitude lower cost than resampling. Beyond efficiency, the decodability of importance from hidden states shows they correspond to information represented in the model's computation rather than being mere artifacts of the resampling procedure, and brings CoT interpretability within reach at scale; grounding this reading further through direct ablation is a promising next step.

Impact statement

This work targets the efficient and trustworthy use of chain-of-thought reasoning in large language models. By making sentence-level causal importance recoverable from a single forward pass, our method lowers the barrier to real-time monitoring of reasoning traces for safety and alignment, where resampling-based attribution has been prohibitively expensive, and to adaptive allocation of inference-time compute toward genuinely load-bearing steps. The net effect is to make scalable trace monitoring, a defensive capability, substantially cheaper to deploy.

References

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful, 2025. URL https://arxiv.org/abs/2503.08679.

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=y2V6YgLaW7.

Barez, F., Wu, T.-Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., et al. Chain-of-thought is not explainability. *Preprint, alphaXiv*, pp. v1, 2025.

Belinkov, Y. and Glass, J. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019. doi: 10.1162/tacl_a_00254. URL https://aclanthology.org/Q19-1004/.

Bogdan, P. C., Macar, U., Nanda, N., and Conmy, A. Thought anchors: Which LLM reasoning steps matter? In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=VnSlfeRCaU.

Boppana, S., Ma, A., Loeffler, M., Sarfati, R., Bigelow, E., Geiger, A., Lewis, O., and Merullo, J. Reasoning theater: Disentangling model beliefs from chain-of-thought. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=1lyagkzogH.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., and Perez, E. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Damani, M., Shenfeld, I., Peng, A., Bobu, A., and Andreas, J. Learning how hard to think: Input-adaptive allocation of LM computation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6qUUgw9bAZ.

Fayyaz, M., Yin, F., Sun, J., and Peng, N. Evaluating human alignment and model faithfulness of llm rationale, 2024. URL https://arxiv.org/abs/2407.00219.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

JeaniceK, M. R. and Taraz, J. Understanding reasoning with thought anchors and probes. LessWrong / AI Alignment Forum post, ARENA capstone project, 2026. URL https://www.lesswrong.com/posts/f9cnGHCiJgo5eDkSQ/understanding-reasoning-with-thought-anchors-and-

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A. D., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Madry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., and Mikulik, V. Chain of thought monitorability: A new and fragile opportunity for ai safety. *CoRR*, abs/2507.11473, July 2025. URL https://doi.org/10.48550/arXiv.2507.11473.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model represen-

tations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Nikankin, Y., Tutek, M., Ashuach, T., Rosenfeld, J., and Belinkov, Y. Reasoning models know what's important, and encode it in their activations, 2026. URL https://arxiv.org/abs/2604.18307.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. Llms know more than they show: On the intrinsic representation of llm hallucinations, 2025. URL https://arxiv.org/abs/2410.02707.

Paul, D., West, R., Bosselut, A., and Faltings, B. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning, 2024. URL https://arxiv.org/abs/2402.13950.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Sun, Y., Stolfo, A., and Sachan, M. Probing for arithmetic errors in language models, 2025. URL https://arxiv.org/abs/2507.12379.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *CoRR*, abs/2305.04388, 2023. URL https://doi.org/10.48550/arXiv.2305.04388.

Venhoff, C., Arcuschin, I., Torr, P., Conmy, A., and Nanda, N. Understanding reasoning in thinking language models via steering vectors, 2025. URL https://arxiv.org/abs/2506.18167.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

A. Implementation Details

A.1. Importance Metric Computation

Semantic similarity. We measure semantic similarity between S_i and a candidate alternative S_i' as the cosine similarity between their sentence embeddings, computed using `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019). A candidate is treated as a valid counterfactual ($S_i' \not\approx S_i$) when this similarity falls below the median similarity across our experiments.

Sampling.

- Number of problems: 400
- Sampling temperature: 0.6
- Number of rollouts per problem: 100
- Maximum continuation length: 1024
- Decoding parameters top-p: 0.9

Outlier removal and thresholding. Before binarization, we exclude extreme importance scores above an upper fence computed from the 1st and 99th percentiles of the per-(dataset, model, metric) score distribution, $u = p_{99} + 1.5\,(p_{99} - p_1)$. This filter is conservative: for the resampling metric it removes between 1 and 10 of the $\sim$4800–10700 sentence scores in every (model, dataset) configuration ($\leq 0.15\%$), so its effect on the downstream analyses is negligible. For the counterfactual metric, sentences without sufficiently dissimilar resamples are additionally excluded by the similarity filter. We then threshold at the elbow of the empirical distribution of remaining scores, identified from the histograms in Figures 3–15; the elbow corresponds to a numeric threshold of 1.0 in all configurations, labeling between 2.5% (Gemma-3-4B, GSM8K, resampling) and 22.4% (DeepSeek-R1-Distill-Qwen-1.5B, MATH, counterfactual) of sentences as important. This fraction tracks how challenging the dataset is for the model: when a model solves a problem robustly, resampling a single sentence seldom shifts the answer, so few sentences are load-bearing.

A.2. Probe Training

Layer selection. For each model we probe three layers at roughly 50%, 70%, and 95% of total depth: L15/L20/L27 of 28 for the DeepSeek-Qwen distillations, L17/L23/L31 of 32 for DeepSeek-R1-Distill-Llama-8B, and L18/L24/L33 of 34 for Gemma-3-4B-IT. Choosing layers as a fraction of depth, rather than at fixed absolute indices, controls for differences in architecture across model families; these are the layers referred to as early/mid/late in Table 1 and by index in Appendix C.

- MLP hidden dimension: 2048
- MLP layers: 3
- Loss: BCE
- Optimizer: Adam
- Learning rate: $1e^{-5}$

No hyperparameter search was performed; the values above were chosen from standard defaults for probing classifiers and held fixed across all configurations.

We use an 80/20 train–test split. To address the class imbalance inherent in the importance score distributions, we subsample the majority class to match the size of the minority class before training, ensuring balanced class frequencies throughout; the majority-class (and random) baseline is therefore 0.5 in every configuration.

Compute resources. Label generation (resampling) is the dominant cost. For smaller models we use NVIDIA Titan RTX GPUs; for larger models we use NVIDIA A100 GPUs. On GSM8K, which produces shorter traces on average, generating 100 rollouts takes approximately 7–15 minutes per problem. On MATH, whose problems elicit longer and more complex chains, the same procedure takes approximately 15–30 minutes. Probe training is negligible in comparison, taking at most a few minutes on a single GPU.

B. Limitations and Future Work

Several directions offer natural opportunities to deepen and broaden this work. Scaling to larger open-source models would establish how the importance signal evolves with model capacity and reasoning quality. Evaluating on traces with higher token budgets, through longer maximum continuation lengths or problem subsets that elicit more complex, multi-branch chains, would extend the framework to the most demanding reasoning structures, where fine-grained importance estimation is arguably most valuable. Direct ablation experiments, verifying that high-probe-score sentences are also the ones whose removal most degrades task performance, would provide stronger causal grounding and sharpen the practical utility of probe predictions. Exploring more sample-efficient importance estimation procedures could further reduce the cost of label generation and scale the approach to larger problem sets. Finally, extending to code generation, multi-step question answering, and agentic settings would establish the generality of the signal, including whether a probe trained on mathematical reasoning transfers to other domains such as code, and unlock the most consequential applications of cheap importance estimation, where the potential payoff in safety monitoring and compute allocation is correspondingly larger.

Our framework's accuracy not exceeding $\sim 80\%$ in all settings could stem from several sources. The resampling metrics themselves may be imperfect proxies: the KL divergence is estimated empirically from 100 rollouts and may not have converged, and stochastic rollouts introduce noise into the KL estimates that propagates into the binary labels. Alternatively, the moderate accuracy could simply reflect a limitation of the probing framework itself, as each probe reads from one transformer layer at a time and may miss importance signals that are distributed across multiple layers or positions. Finally, it is possible that causal importance is genuinely not localized in any single residual stream site and would require richer representations to recover. The results in Table 1 also reflect a single training run per configuration; a five-seed rerun of the DeepSeek-R1-Distill-Qwen-1.5B resampling probes shows seed variability of ± 1–3 accuracy points across layers on both datasets, leaving the qualitative conclusions unchanged, though future work should report variance for all configurations.

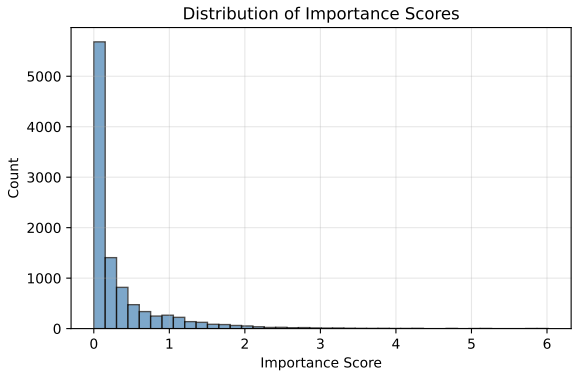
Several design choices deserve ablation in their own right. We frame prediction as binary classification with a threshold at the elbow of the score distribution; ablating this threshold, labeling only the top and bottom $k\%$ of sentences while leaving a buffer of unlabeled sentences in between, or regressing directly on the continuous KL scores would test how much signal the binarization discards. Our probes are 3-layer MLPs; comparing against linear probes would reveal whether importance corresponds to a direction in activation space or requires nonlinear readout, and supplying relative position explicitly as an input feature would allow its contribution (Section 3.2) to be read off and neutralized directly. The probes should also be evaluated as an interpretability tool rather than only as classifiers: qualitative comparisons of probe predictions against resampling-based scores on held-out traces, precision–recall analyses of which important sentences are missed, and characterizations of failure modes would clarify what the accuracy figures mean in practice and how much of the resampling signal is traded away for the several-orders-of-magnitude cost reduction. Probe-guided steering of generation is a further natural extension.

C. Experiments

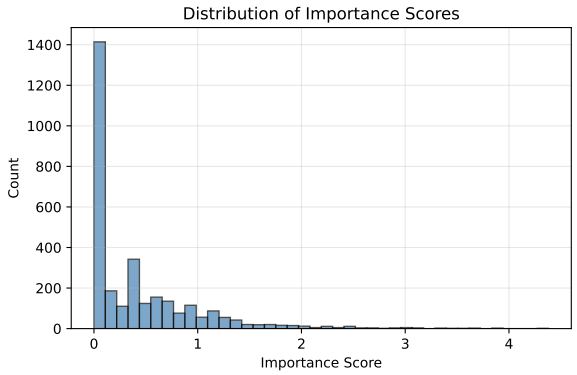
This appendix collects the per-configuration plots underlying Table 1 and Section 3.2, organized by dataset (GSM8K, MATH) and importance metric (resampling, counterfactual). For each of the four models we report: the distribution of importance scores, whose elbow determines the binarization threshold; ROC curves per probed layer, which decompose the error rate into missed important sentences and false alarms; the distributions of predicted probabilities for important and unimportant sentences, showing how well the two classes separate; and scatter plots of importance against relative sentence position, whose correlations underlie the analysis of Section 3.2.

C.1. GSM8k

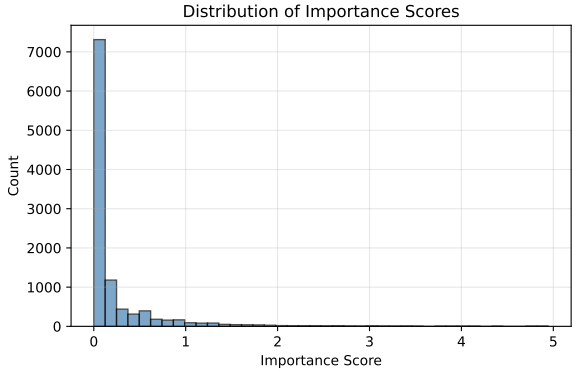
C.1.1. Resampling importance



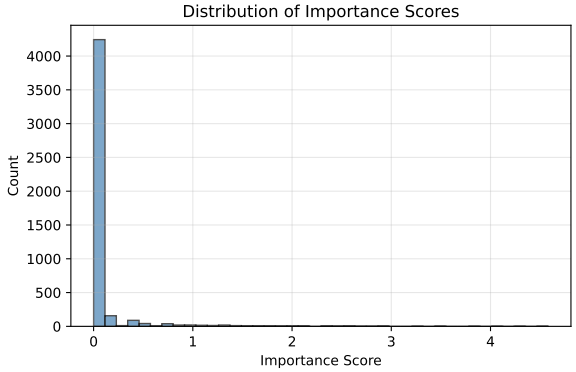
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B

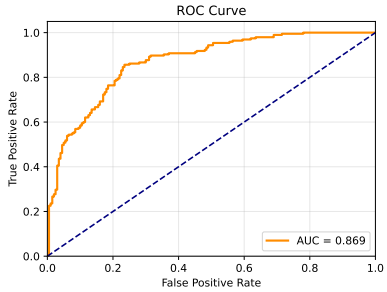


(c) DeepSeek-R1-Distill-Llama-8B

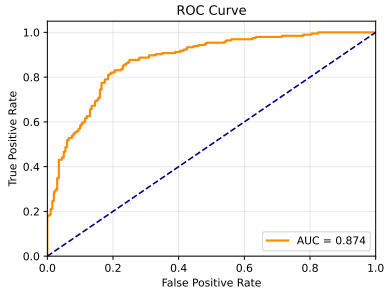


(d) gemma-3-4b-it-7B

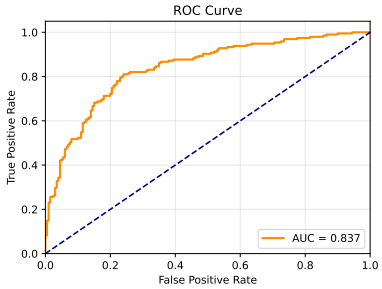
Figure 3. Distribution of resampling importance scores on GSM8K, per model. The binarization threshold is set at the elbow of each distribution.



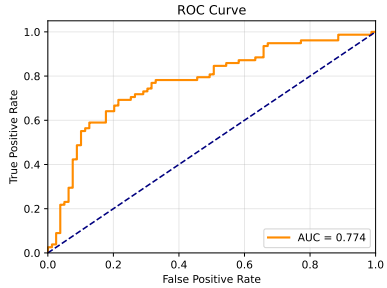
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



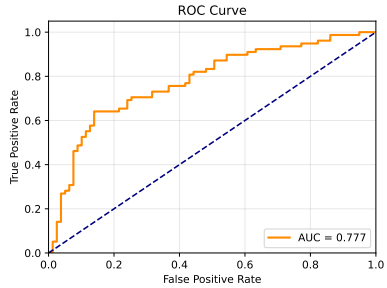
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



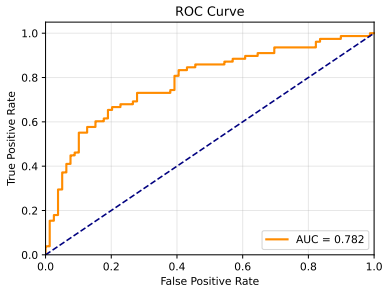
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



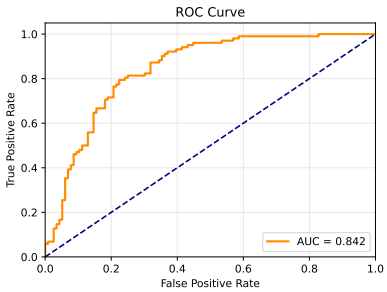
(d) DeepSeek-R1-Distill-Qwen-7B – L15



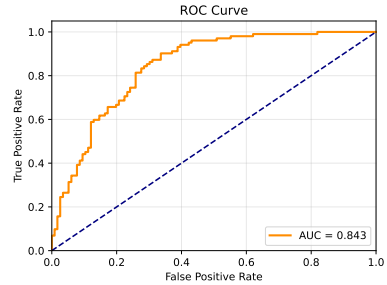
(e) DeepSeek-R1-Distill-Qwen-7B – L20



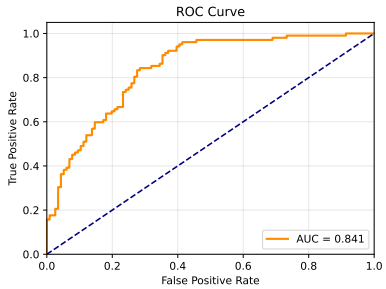
(f) DeepSeek-R1-Distill-Qwen-7B – L27



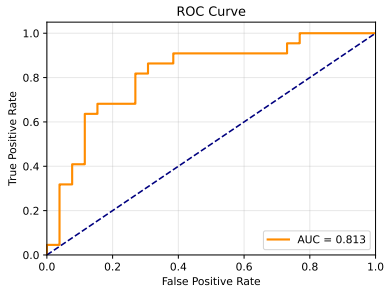
(g) DeepSeek-R1-Distill-Llama-8B – L17



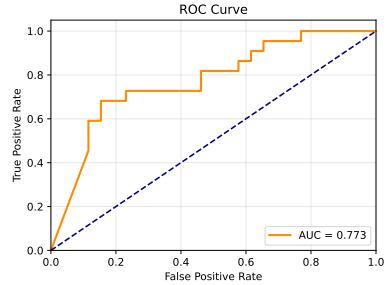
(h) DeepSeek-R1-Distill-Llama-8B – L23



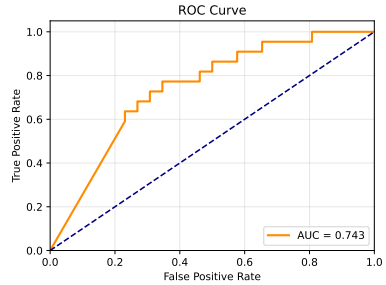
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

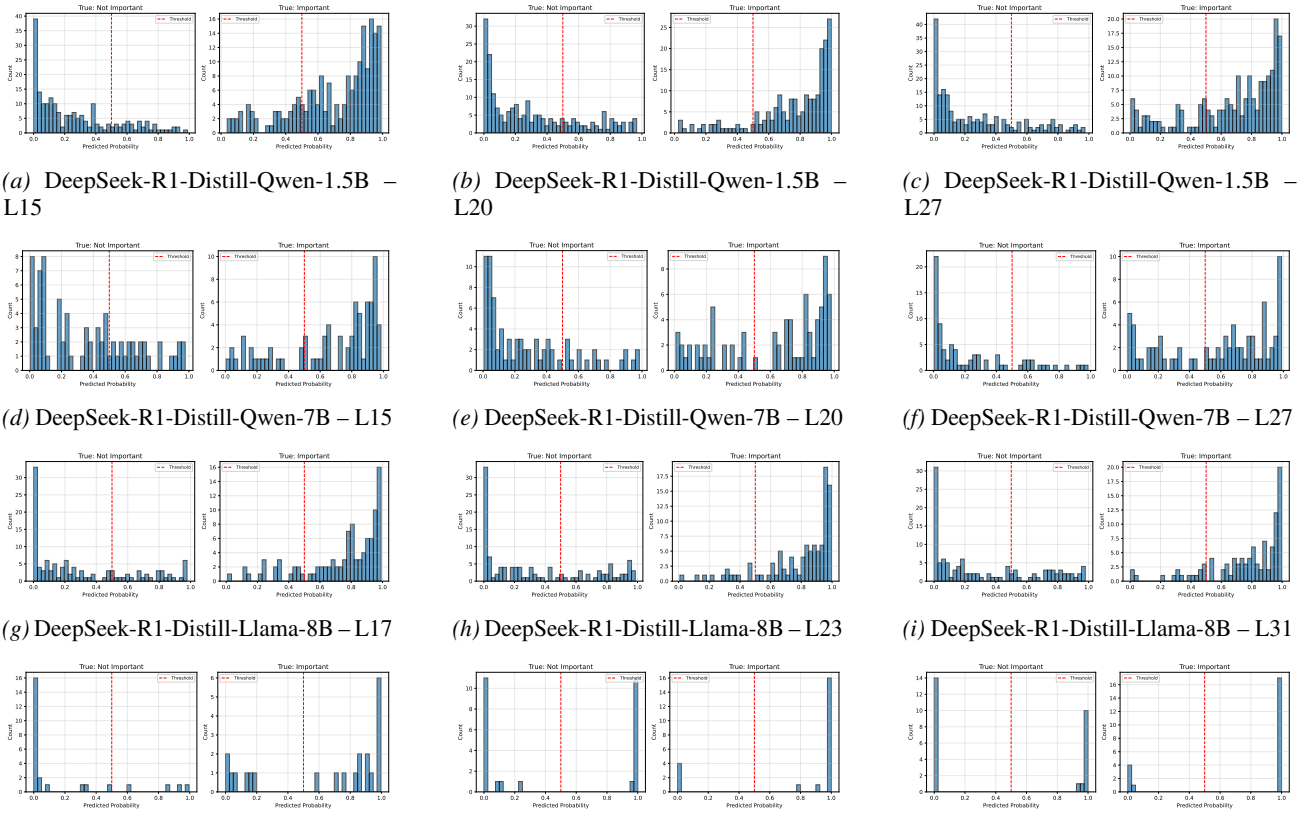


(k) gemma-3-4b-it – L24



(l) gemma-3-4b-it – L33

Figure 4. ROC curves for resampling importance classification on GSM8K, per model and layer.



(a) DeepSeek-R1-Distill-Qwen-1.5B – L15

(b) DeepSeek-R1-Distill-Qwen-1.5B – L20

(c) DeepSeek-R1-Distill-Qwen-1.5B – L27

(d) DeepSeek-R1-Distill-Qwen-7B – L15

(e) DeepSeek-R1-Distill-Qwen-7B – L20

(f) DeepSeek-R1-Distill-Qwen-7B – L27

(g) DeepSeek-R1-Distill-Llama-8B – L17

(h) DeepSeek-R1-Distill-Llama-8B – L23

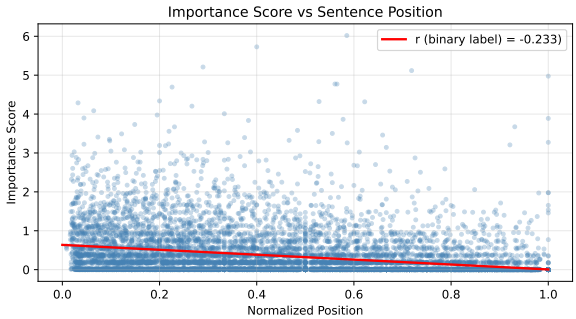
(i) DeepSeek-R1-Distill-Llama-8B – L31

(j) gemma-3-4b-it – L18

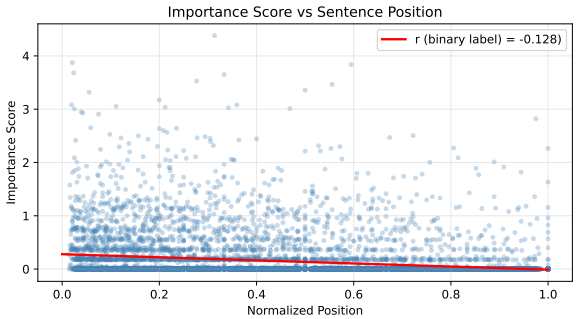
(k) gemma-3-4b-it – L24

(l) gemma-3-4b-it – L33

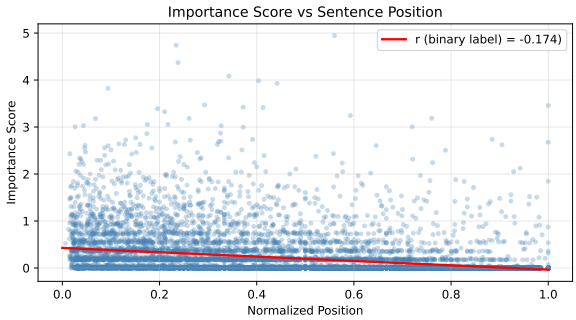
Figure 5. Predicted-probability distributions for important and unimportant sentences (resampling importance, GSM8K), per model and layer.



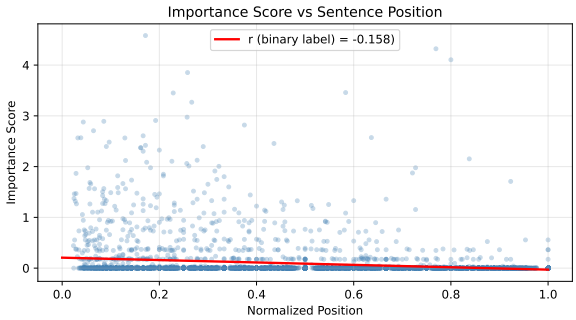
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B



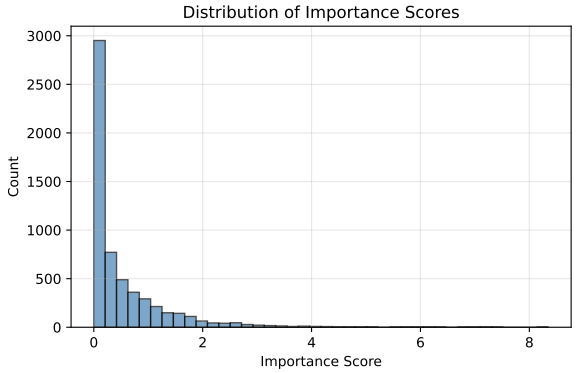
(c) DeepSeek-R1-Distill-Llama-8B



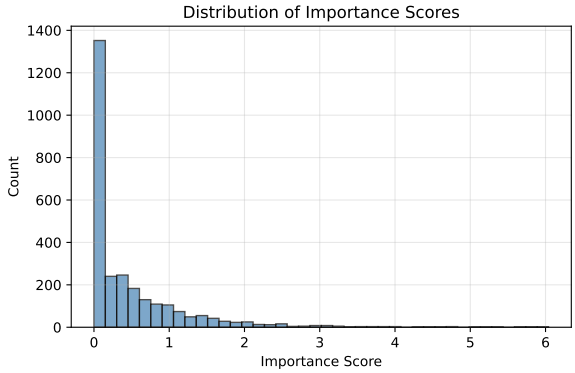
(d) gemma-3-4b-it

Figure 6. Resampling importance against relative sentence position on GSM8K, per model.

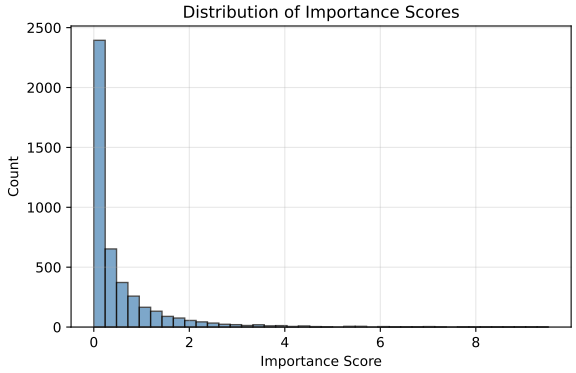
C.1.2. Counterfactual importance



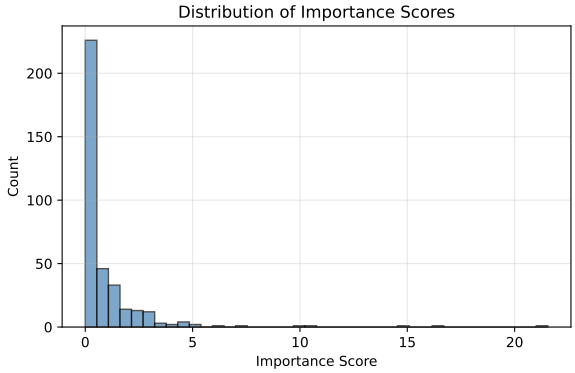
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B

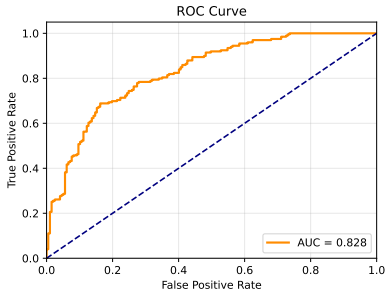


(c) DeepSeek-R1-Distill-Llama-8B

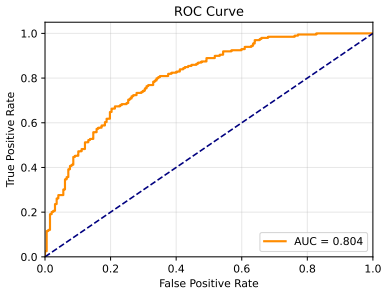


(d) gemma-3-4b-it-7B

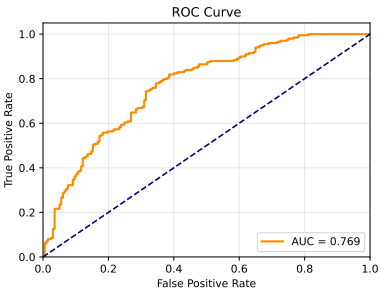
Figure 7. Distribution of counterfactual importance scores on GSM8K, per model. The binarization threshold is set at the elbow of each distribution.



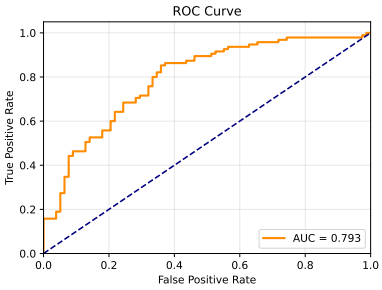
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



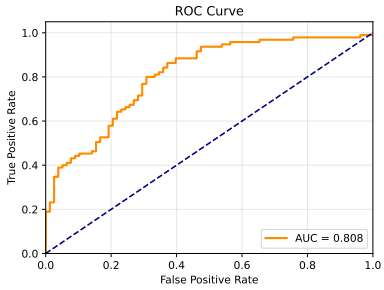
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



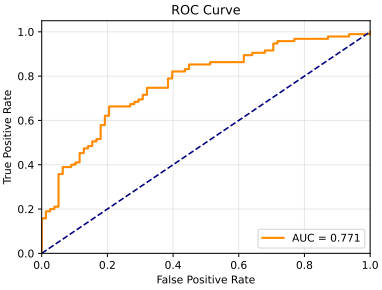
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



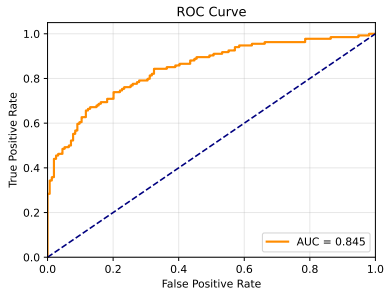
(d) DeepSeek-R1-Distill-Qwen-7B – L15



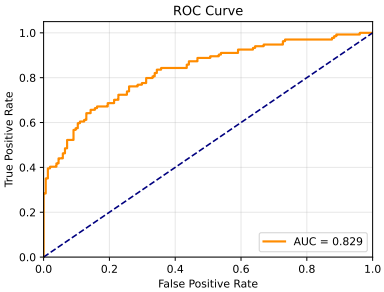
(e) DeepSeek-R1-Distill-Qwen-7B – L20



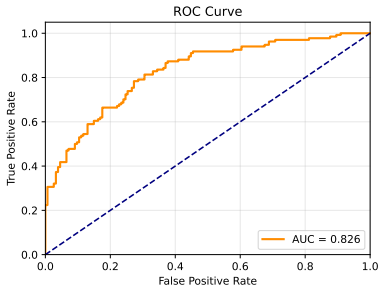
(f) DeepSeek-R1-Distill-Qwen-7B – L27



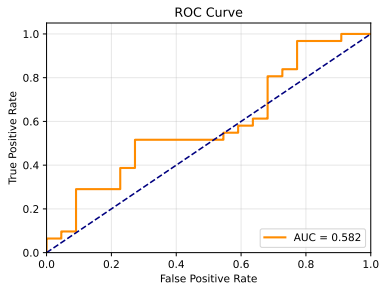
(g) DeepSeek-R1-Distill-Llama-8B – L17



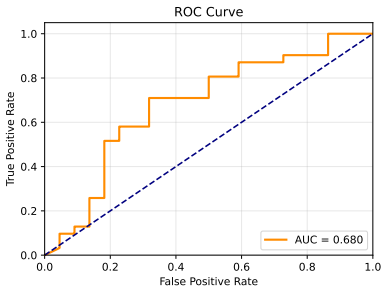
(h) DeepSeek-R1-Distill-Llama-8B – L23



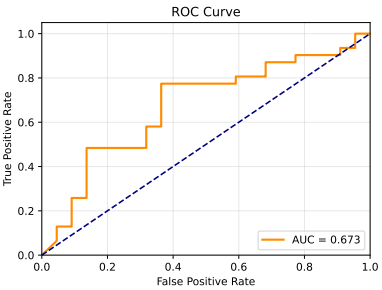
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

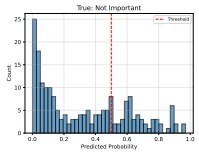


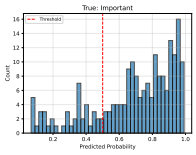
(k) gemma-3-4b-it – L24



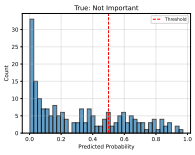
(l) gemma-3-4b-it – L33

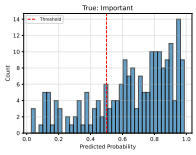
Figure 8. ROC curves for counterfactual importance classification on GSM8K, per model and layer.



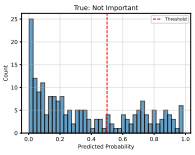


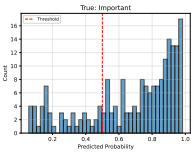
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



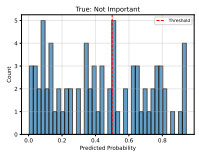


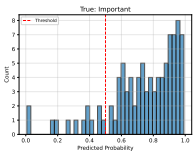
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



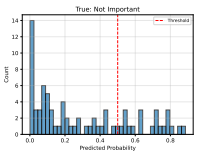


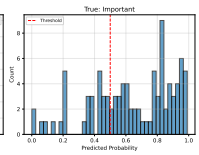
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



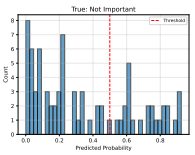


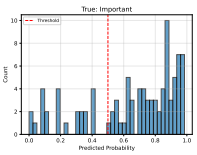
(d) DeepSeek-R1-Distill-Qwen-7B – L15



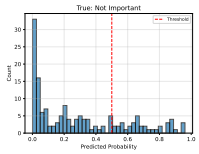


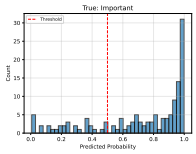
(e) DeepSeek-R1-Distill-Qwen-7B – L20



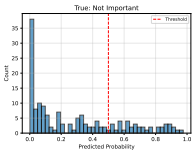


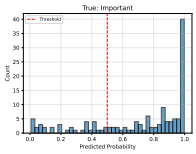
(f) DeepSeek-R1-Distill-Qwen-7B – L27



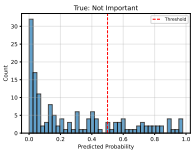


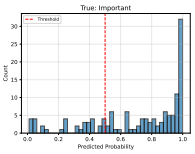
(g) DeepSeek-R1-Distill-Llama-8B – L17



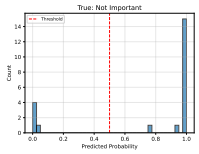


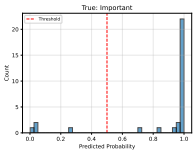
(h) DeepSeek-R1-Distill-Llama-8B – L23



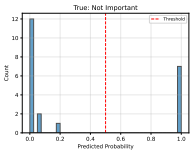


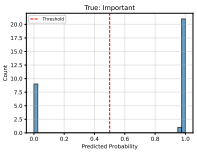
(i) DeepSeek-R1-Distill-Llama-8B – L31



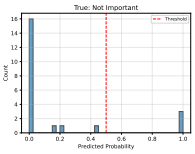


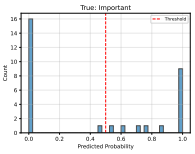
(j) gemma-3-4b-it – L18





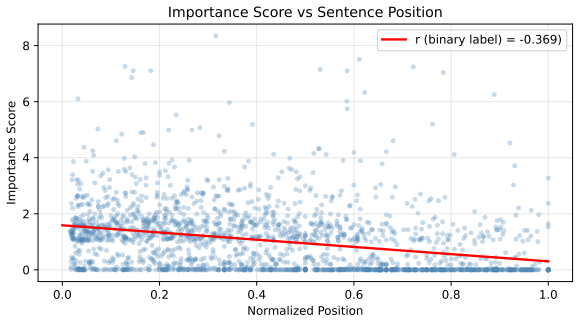
(k) gemma-3-4b-it – L24



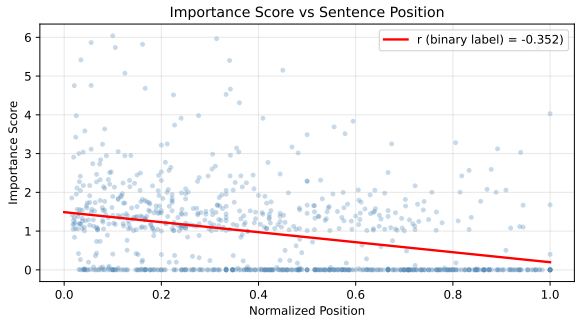


(l) gemma-3-4b-it – L33

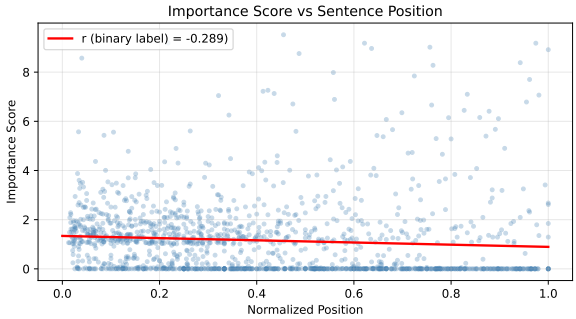
Figure 9. Predicted-probability distributions for important and unimportant sentences (counterfactual importance, GSM8K), per model and layer.



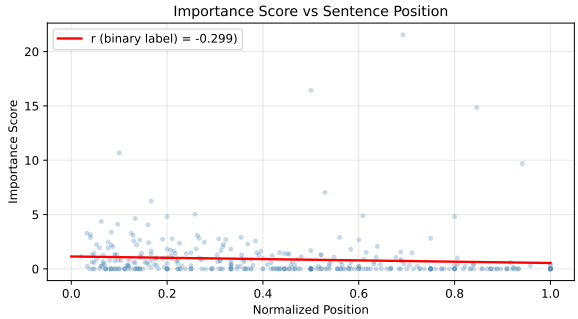
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B



(c) DeepSeek-R1-Distill-Llama-8B

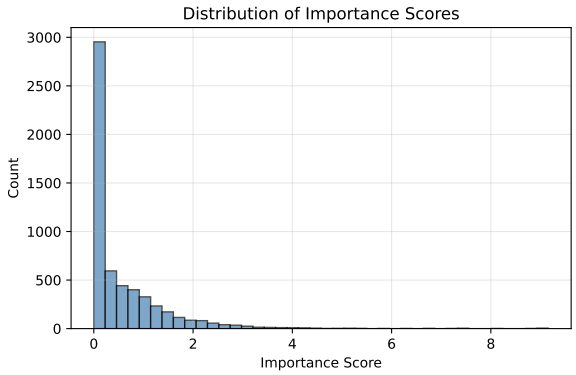


(d) gemma-3-4b-it

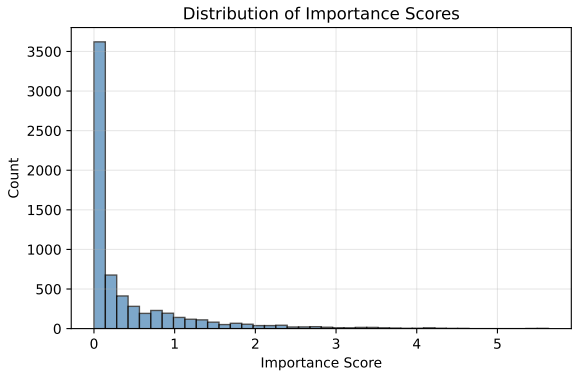
Figure 10. Counterfactual importance against relative sentence position on GSM8K, per model.

C.2. MATH

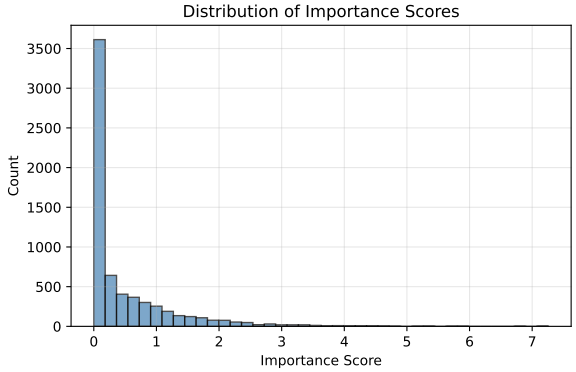
C.2.1. Resampling importance



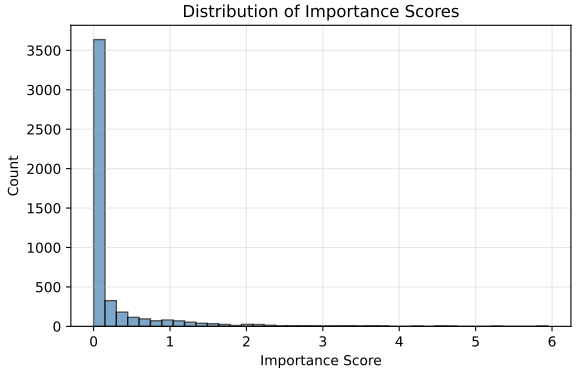
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B

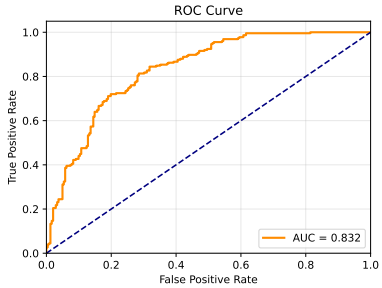


(c) DeepSeek-R1-Distill-Llama-8B

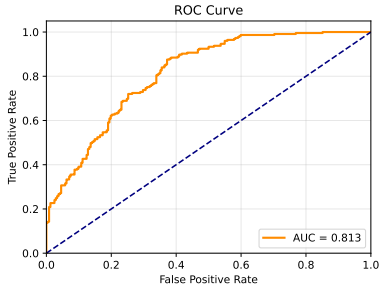


(d) gemma-3-4b-it-7B

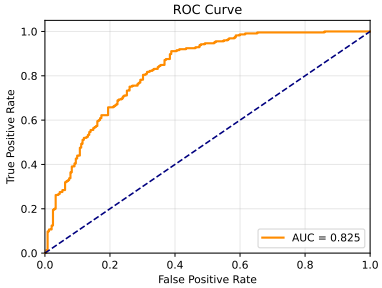
Figure 11. Distribution of resampling importance scores on MATH, per model. The binarization threshold is set at the elbow of each distribution.



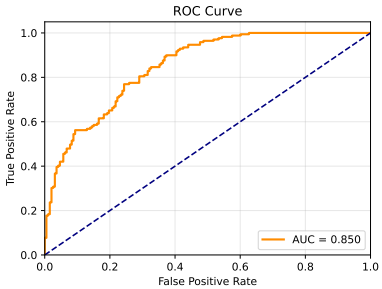
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



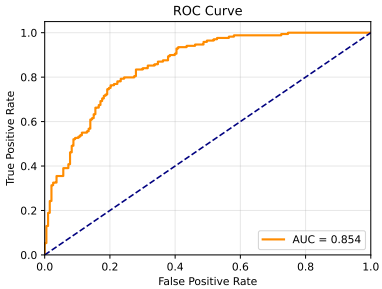
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



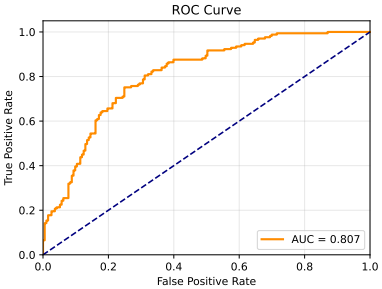
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



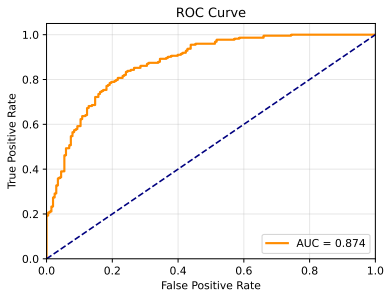
(d) DeepSeek-R1-Distill-Qwen-7B – L15



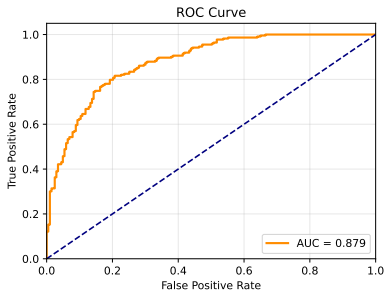
(e) DeepSeek-R1-Distill-Qwen-7B – L20



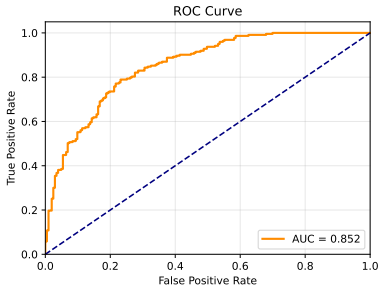
(f) DeepSeek-R1-Distill-Qwen-7B – L27



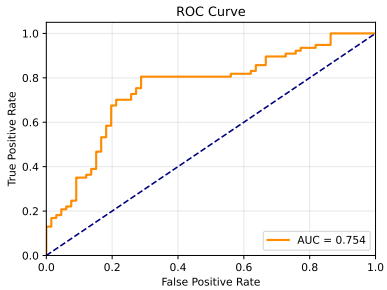
(g) DeepSeek-R1-Distill-Llama-8B – L17



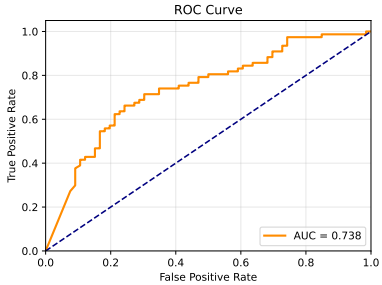
(h) DeepSeek-R1-Distill-Llama-8B – L23



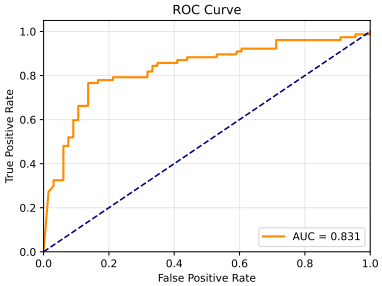
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

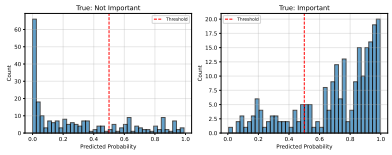


(k) gemma-3-4b-it – L24

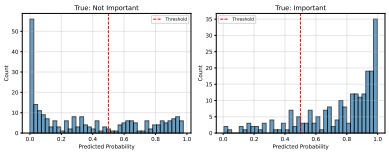


(l) gemma-3-4b-it – L33

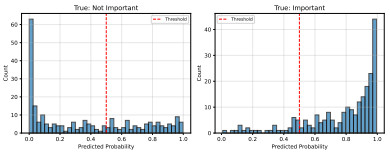
Figure 12. ROC curves for resampling importance classification on MATH, per model and layer.



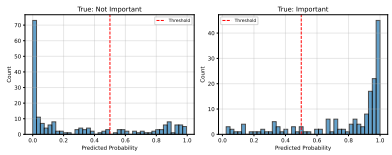
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



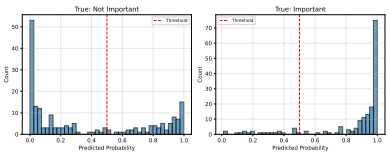
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



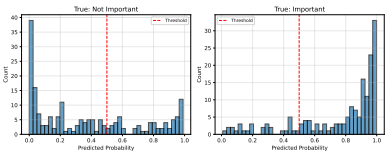
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



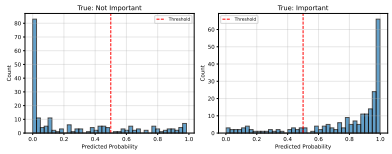
(d) DeepSeek-R1-Distill-Qwen-7B – L15



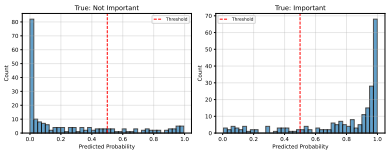
(e) DeepSeek-R1-Distill-Qwen-7B – L20



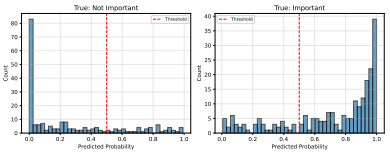
(f) DeepSeek-R1-Distill-Qwen-7B – L27



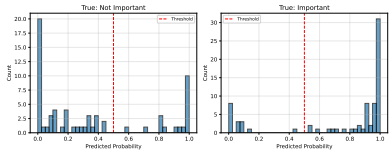
(g) DeepSeek-R1-Distill-Llama-8B – L17



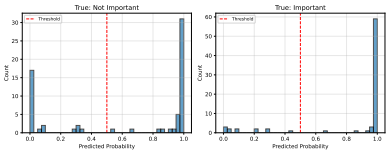
(h) DeepSeek-R1-Distill-Llama-8B – L23



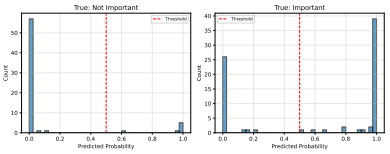
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

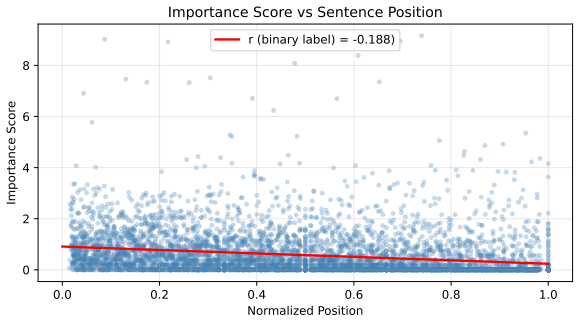


(k) gemma-3-4b-it – L24

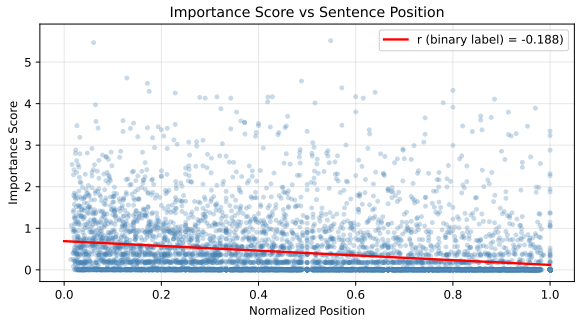


(l) gemma-3-4b-it – L33

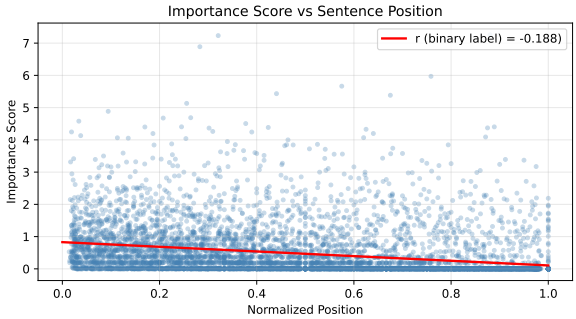
Figure 13. Predicted-probability distributions for important and unimportant sentences (resampling importance, MATH), per model and layer.



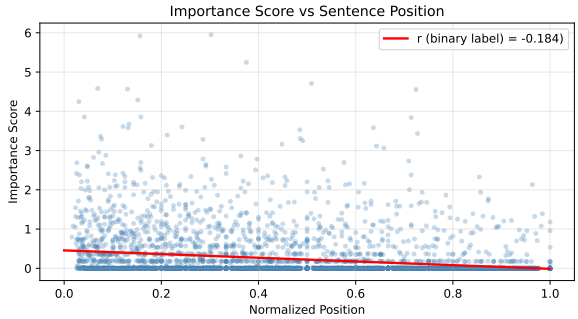
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B



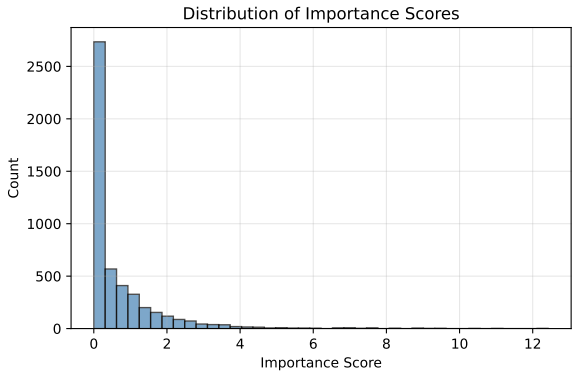
(c) DeepSeek-R1-Distill-Llama-8B



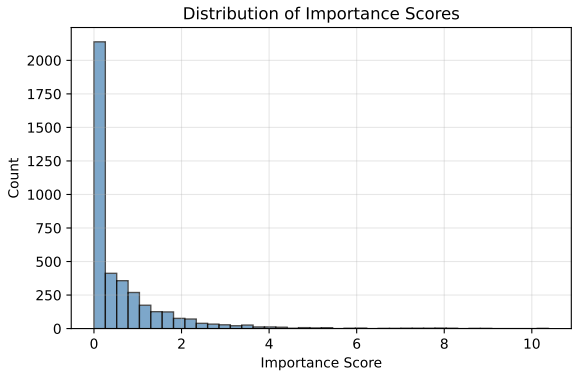
(d) gemma-3-4b-it

Figure 14. Resampling importance against relative sentence position on MATH, per model.

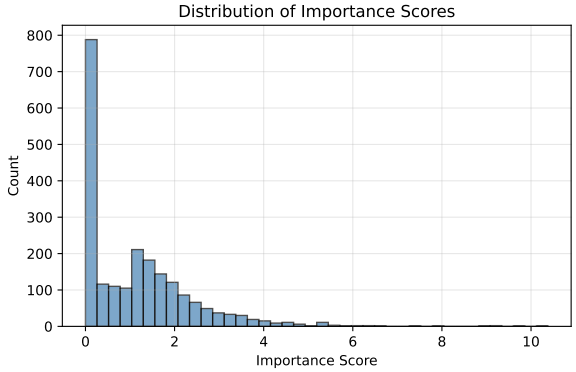
C.2.2. Counterfactual importance



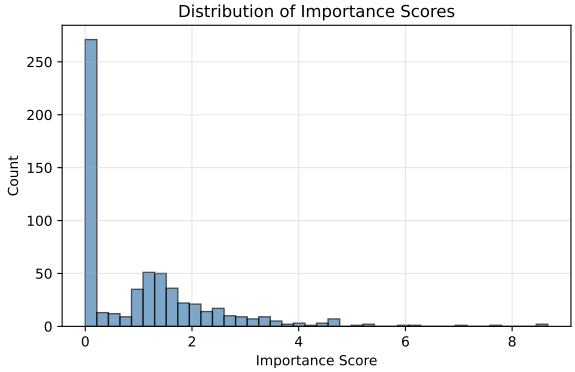
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B

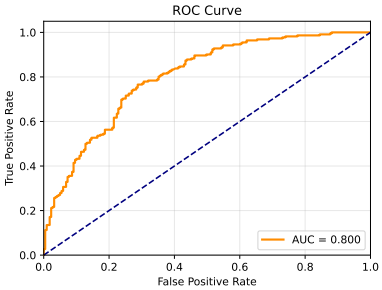


(c) DeepSeek-R1-Distill-Llama-8B

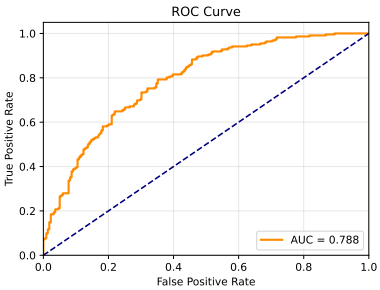


(d) gemma-3-4b-it-7B

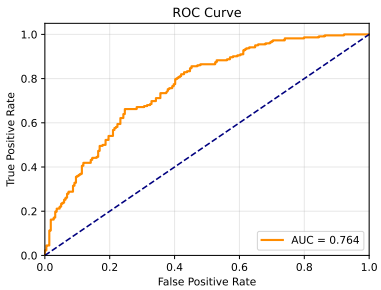
Figure 15. Distribution of counterfactual importance scores on MATH, per model. The binarization threshold is set at the elbow of each distribution.



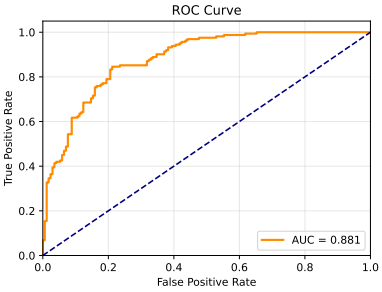
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



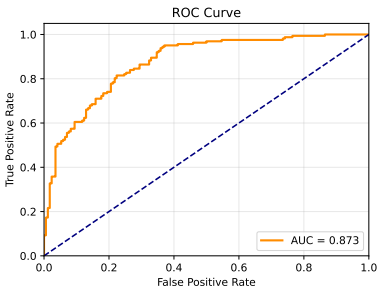
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



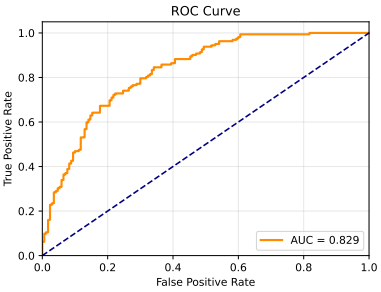
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



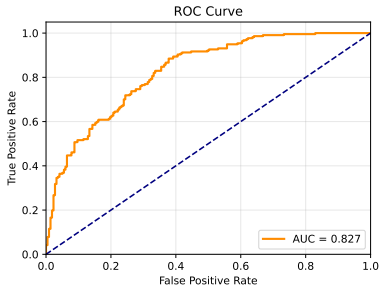
(d) DeepSeek-R1-Distill-Qwen-7B – L15



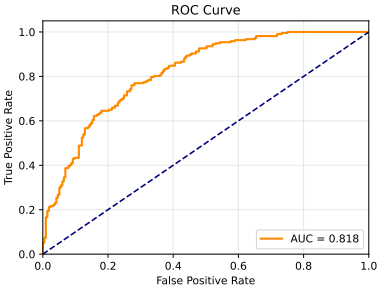
(e) DeepSeek-R1-Distill-Qwen-7B – L20



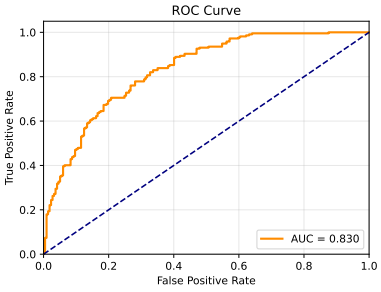
(f) DeepSeek-R1-Distill-Qwen-7B – L27



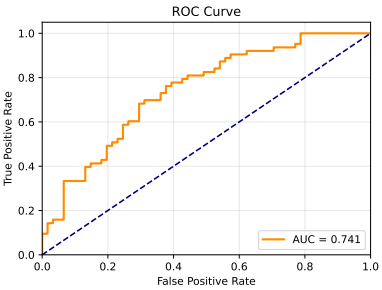
(g) DeepSeek-R1-Distill-Llama-8B – L17



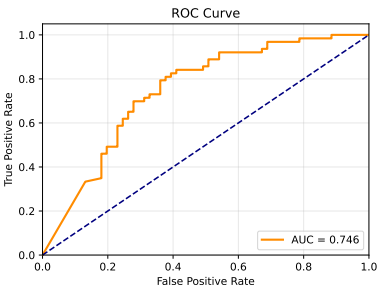
(h) DeepSeek-R1-Distill-Llama-8B – L23



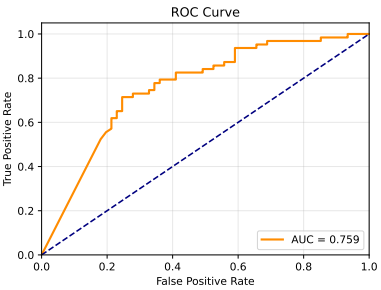
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

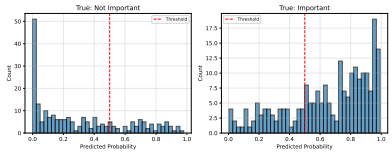


(k) gemma-3-4b-it – L24

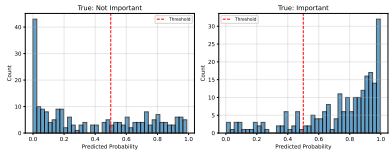


(l) gemma-3-4b-it – L33

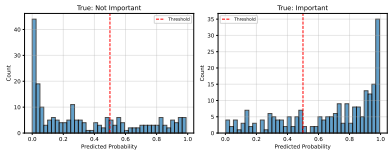
Figure 16. ROC curves for counterfactual importance classification on MATH, per model and layer.



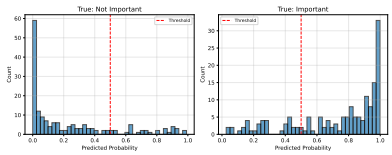
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



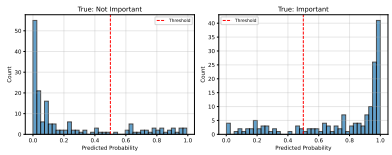
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



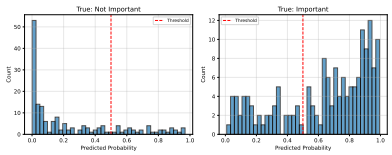
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



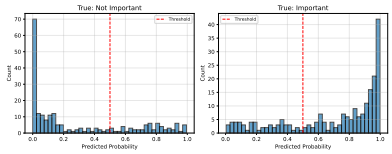
(d) DeepSeek-R1-Distill-Qwen-7B – L15



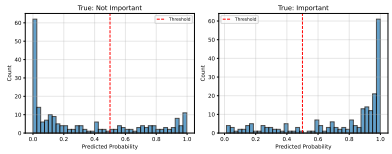
(e) DeepSeek-R1-Distill-Qwen-7B – L20



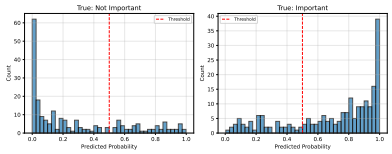
(f) DeepSeek-R1-Distill-Qwen-7B – L27



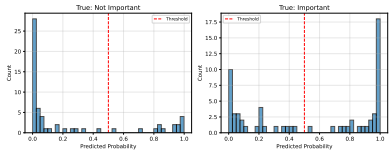
(g) DeepSeek-R1-Distill-Llama-8B – L17



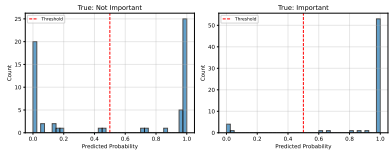
(h) DeepSeek-R1-Distill-Llama-8B – L23



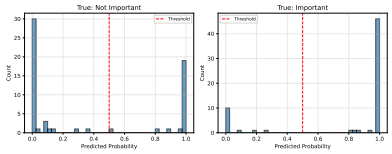
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

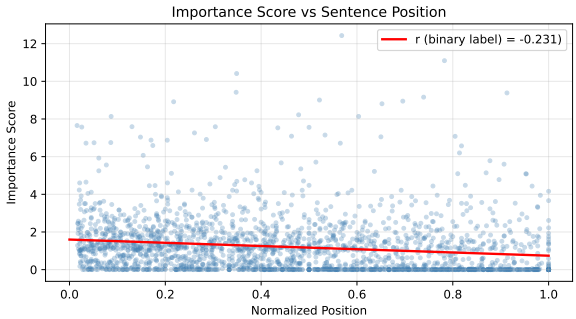


(k) gemma-3-4b-it – L24

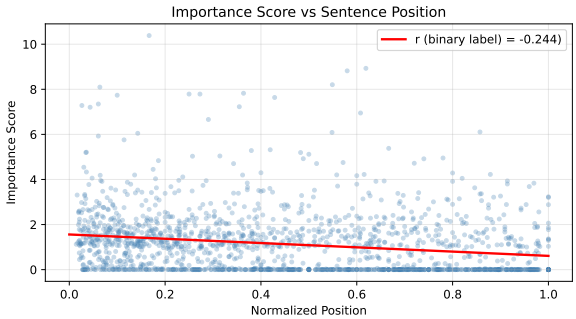


(l) gemma-3-4b-it – L33

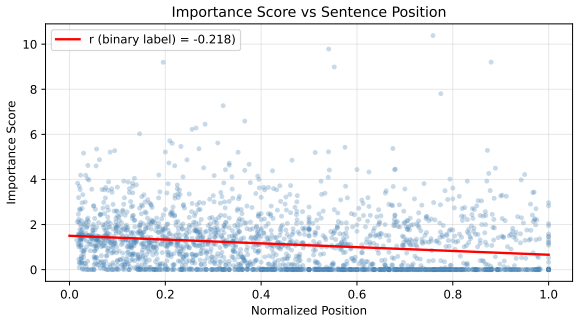
Figure 17. Predicted-probability distributions for important and unimportant sentences (counterfactual importance, MATH), per model and layer.



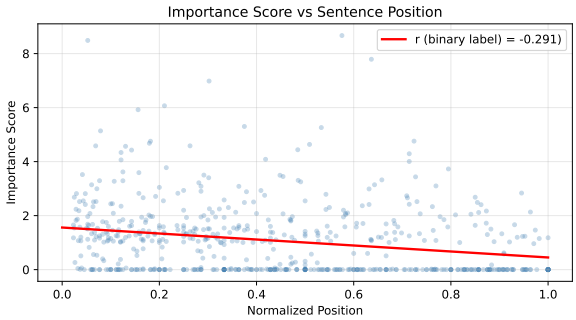
(a) DeepSeek-R1-Distill-Qwen-1.5B



(b) DeepSeek-R1-Distill-Qwen-7B



(c) DeepSeek-R1-Distill-Llama-8B



(d) gemma-3-4b-it

Figure 18. Counterfactual importance against relative sentence position on MATH, per model.