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

Anonymous ACL submission

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 metrics reflect genuine properties of the model's computation rather than artifacts of the measurement procedure, 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 and Tegmark, 2024; Azaria and Mitchell, 2023), spatial and temporal coordinates (Gurnee and 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 efficiency, the decodability of importance from hidden states suggests these metrics capture genuine properties of the model's computation rather than artifacts of the resampling procedure, and that recovering them 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:

$$\mathrm{I}_{\mathrm{res}}(S_i) = D_{KL}\Big[p(A \mid S_i') \,\Big\|\, p(A \mid S_i)\Big]. \quad (1)$$

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:

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

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.2.

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 may reflect errors rather than genuine reasoning anchors, mixing two qualitatively different phenomena that are best studied separately.

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 enable a single numerical-match criterion across both benchmarks, we restrict MATH to problems with floating-point answers, excluding symbolic ones (e.g. unsimplified fractions, polynomial expressions); GSM8K answers are numeric by construction. The resulting subset preserves the difficulty profile of MATH while avoiding the need for custom parsing and equivalence-checking logic.

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 B Figures 3, 7, 11, 15). 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 B.

Table 1: Binary classification accuracy for predicting sentence-level importance from hidden-state representations of one transformer layer (early/mid/late). We find a consistent signal across all settings, with classification accuracy reaching up to $\sim 80\%$.

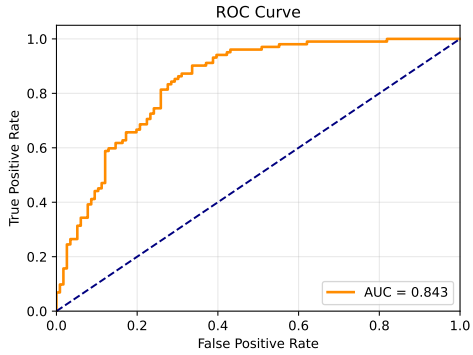
		Resampling			Counterfactual		
Model	**Dataset**	early	mid	late	early	mid	late
DeepSeek-R1-Distill-Qwen-1.5B	gsm8k	.790	.808	.770	.750	.750	.710
	MATH	.750	.732	.741	.730	.708	.680
DeepSeek-R1-Distill-Qwen-7B	gsm8k	.720	.713	.726	.728	.711	.717
	MATH	.746	.762	.738	.800	.786	.747
DeepSeek-R1-Distill-Llama-8B	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 Correlation analysis

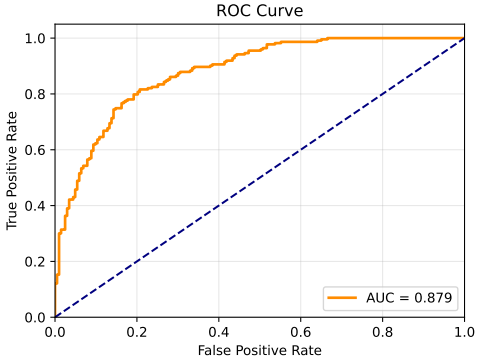
A potential confound in this setting is that sentence position within the trace is negatively correlated with importance: earlier sentences, particularly planning steps, tend to carry higher causal impact than later ones. This matters because position is itself trivially decodable from a hidden state, so a probe could in principle achieve high accuracy simply by reading off how far through the trace it is, without recovering anything about importance per se. To rule this out, we compute the correlation between relative sentence position and the importance label directly. As shown in Appendix B Figures 6, 10, 14, 18, this correlation is negligible in all our experiments, indicating that the probe's predictions cannot be explained by position alone.

4 Discussion

Sentence-level causal importance can be decoded from internal activations with substantially above-chance accuracy across model families, scales, and datasets. Since the resampling metrics of Bogdan et al. (2025) 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

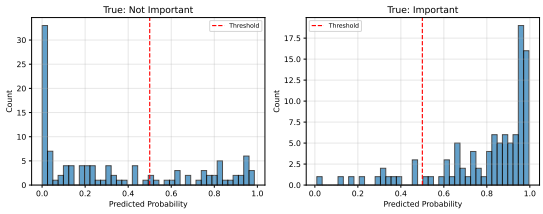


(a) GSM8K

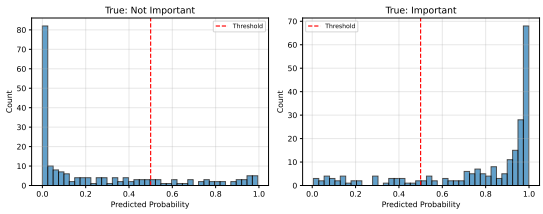


(b) MATH

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.

track a real property of the model's computation, and that the model in some sense "knows" which of its own steps are load-bearing as it generates them. Reducing the cost from $\sim$100 rollouts per sentence to a single forward pass makes CoT interpretability at scale realistic: real-time CoT monitoring (Korbak et al., 2025), adaptive compute allocation, and large-scale mechanistic studies become practically accessible.

Table 1 surfaces a few consistent patterns. Mid-layer activations are typically the strongest readout site for resampling importance, 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, while Gemma-3-4B lags noticeably, suggesting probe quality tracks reasoning competence more than parameter count. 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 extend this work, including 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

A growing body of evidence questions whether the verbalized CoT reflects how the model arrives at its answer. Turpin et al. (2023) showed that prompt biases shift predictions in up to 36% of cases without being acknowledged in the CoT; Chen et al. (2025) extended this to state-of-the-art reasoning models, finding CoTs verbalize hint influence in under 20% of cases on average. Arcuschin et al. (2025) documented post-hoc rationalization, silent restoration errors, and illogical shortcuts on realistic prompts, 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 (Korbak et al., 2025).

A complementary line attributes causal importance to individual reasoning steps. Paul et al. (2024) use causal mediation on twelve LLMs and find that models do not reliably use their own intermediate steps. Bogdan et al. (2025) introduced *thought anchors* and the two resampling-based metrics we adopt (Section 2), showing that planning and uncertainty-management steps exert greater causal influence than local computational steps, while surface-salient content such as self-preservation statements in agentic settings can have surprisingly low impact. These methods are principled but require many rollouts per sentence; our work addresses this by predicting importance from internal activations in a single forward pass.

5.2 Probing Internal Representations

Probing has long been used to understand what is encoded in language model representations (Belinkov and Glass, 2019; Hewitt and Manning, 2019). Recent work shows that surprisingly complex properties can be linearly decoded from hidden states, including truthfulness (Marks and Tegmark, 2024; Azaria and Mitchell, 2023), spatial and temporal coordinates (Gurnee and Tegmark, 2024), and hallucination signals (Sun et al., 2025; Orgad et al., 2025; Chen et al., 2024), with Orgad et al. (2025) showing the correct answer is sometimes recoverable from hidden states even when the generated response is wrong. In the reasoning setting, Venhoff et al. (2025) distinguishes retrieval from deduction sentences, and Damani et al. (2025) predicts marginal reward gains from additional computation. These targets are typically aggregated or statement-level; we instead probe for the causal importance of individual CoT sentences, a fine-grained target whose value depends on how the answer distribution would shift under semantically different continuations.

6 Conclusion

We presented a probing framework that predicts sentence-level CoT importance from internal activations in a single forward pass, reducing the cost of the resampling-based metrics of Bogdan et al. (2025) by orders of magnitude. The decodability of these metrics from hidden states suggests they reflect genuine properties of the model's computation rather than artifacts of the resampling procedure, and brings CoT interpretability within reach at scale.

Limitations

Several directions would deepen and broaden this work. Scaling to larger open-source models would clarify how the importance signal evolves with capacity and reasoning quality. Evaluating on traces with higher token budgets or problem subsets that elicit longer, 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 performance, would provide stronger causal grounding. More sample-efficient importance estimation could further reduce label-generation cost. Finally, extending to code generation, multi-step QA, and agentic settings would establish the generality of the signal and unlock the most consequential applications of cheap importance estimation.

Probe accuracy does not exceed $\sim 80\%$ in any setting. This could reflect noise in the labels themselves, since the KL divergence is estimated from 100 stochastic rollouts and may not have converged; a limitation of single-layer probing, which may miss signals distributed across layers or positions; or a deeper fact that causal importance is not localized in any single residual stream site. Results in Table 1 also reflect a single training run per configuration; the consistency across model families, datasets, layers, and both metrics is suggestive of a robust signal, but future work should quantify probe variance across seeds explicitly.

Ethics Statement

We see two main risks of probe-based importance estimation. First, cheap probes could create a false sense of CoT transparency: low predicted importance is not a guarantee of causal inertness, and probe scores should not replace direct intervention in safety-critical settings. Second, importance estimates could in principle be used adversarially to suppress or rewrite load-bearing reasoning steps while preserving final answers, degrading monitorability. We view both as reasons to treat probe-based importance as a screening tool that complements, rather than replaces, resampling or ablation evidence. No human subjects, private data, or personally identifying information were involved in this work.

Use of AI assistants. Claude Code (Anthropic) was used to assist with implementation and running of experiments, and Claude (Anthropic) and ChatGPT (OpenAI) were used to assist with drafting and editing portions of the paper. All outputs of these tools were reviewed and verified by the authors.

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *Preprint*, arXiv:2503.08679.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, and 1 others. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv*, page v1.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which LLM reasoning steps matter? In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

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

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *Preprint*, arXiv:2505.05410.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. 2025. Learning how hard to think: Input-adaptive allocation of LM computation. In *The Thirteenth International Conference on Learning Representations*.

Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. 2024. Evaluating human alignment and model faithfulness of llm rationale. *Preprint*, arXiv:2407.00219.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca D. Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, and 22 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *CoRR*, abs/2507.11473.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. *Preprint*, arXiv:2410.02707.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *Preprint*, arXiv:2402.13950.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Yucheng Sun, Alessandro Stolfo, and Mrinmaya Sachan. 2025. Probing for arithmetic errors in language models. *Preprint*, arXiv:2507.12379.

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, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *CoRR*, abs/2305.04388.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025. Understanding reasoning in thinking language models via steering vectors. *Preprint*, arXiv:2506.18167.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

A Implementation Details

A.1 Datasets and Models

Datasets and Models. We evaluate on GSM8K (Cobbe et al., 2021) and the MATH benchmark (Hendrycks et al., 2021), both consisting of English-language mathematical word problems; GSM8K targets grade-school arithmetic while MATH targets competition-level problems. For MATH we restrict to problems with floating-point answers (excluding symbolic forms such as unsimplified fractions or polynomial expressions) to enable a single numerical-match criterion. We sample 400 problems per dataset for label generation. We probe four open-weight reasoning models: DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025), and Gemma-3-4B-IT (Team et al., 2025). Sentence embeddings for the counterfactual filter are produced by all-MiniLM-L6-v2 (Reimers and Gurevych, 2019).[1]

A.2 Importance Metric Computation

Sentence segmentation. Reasoning traces are segmented into sentences using a punctuation-based splitter that respects mathematical notation (e.g., decimal points, equation delimiters, and LaTeX math environments are not treated as sentence boundaries). Hidden states are extracted at the final punctuation token of each resulting sentence.

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 and 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 remove importance scores beyond the 99th percentile of the per-(dataset, model, metric) distribution, which removes a small number of extreme outliers driven by rollouts that fail to produce a parseable answer. We then threshold at the elbow of the empirical distribution of remaining scores, identified visually from the histograms in Figures 3–15.

A.3 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.

- 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 problem-level train–test split, ensuring that sentences from the same trace do not appear in both sets. 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.

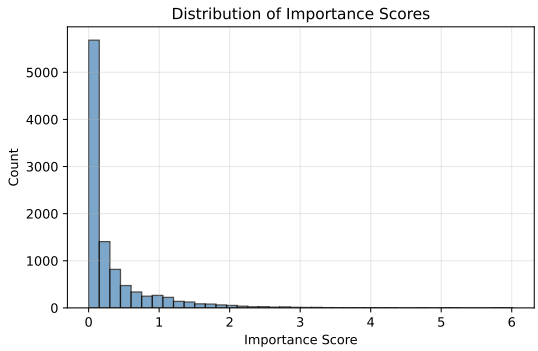
Compute resources. Label generation (resampling) is the dominant cost. For smaller models we use NVIDIA Titan RTX GPUs; for larger models we use NVIDIA A100 GPUs. On GSM8K, which produces shorter traces on average, generating 100 rollouts takes approximately 7–15 minutes per problem. On MATH, whose problems 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.

[1] The DeepSeek-R1 distilled models are released under the MIT License; Gemma-3 is released under the Gemma Terms of Use. GSM8K and MATH are released under the MIT License, and all-MiniLM-L6-v2 under Apache 2.0. Our use is consistent with their intended research use. We release our probe weights, training, and evaluation code under a permissive open-source license for research use.

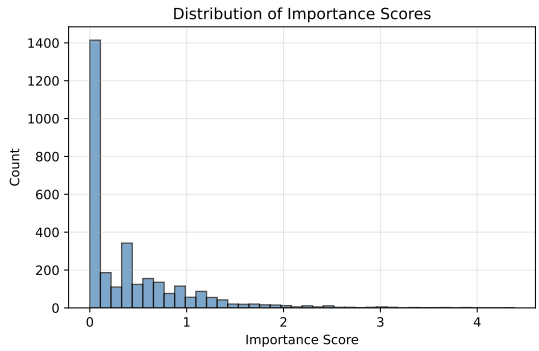
B Experiments

B.1 GSM8k

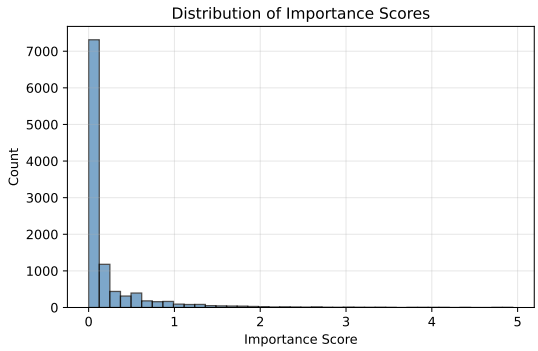
B.1.1 Resampling importance



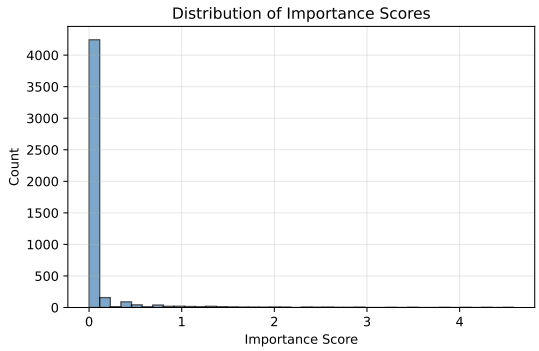
(a) DeepSeek-R1-Distill-Qwen-1.5B



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

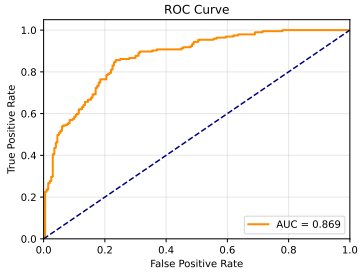


(c) DeepSeek-R1-Distill-Llama-8B

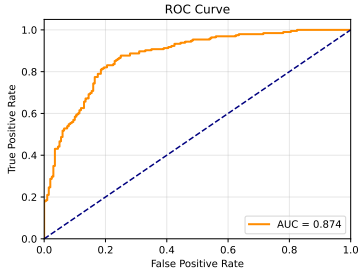


(d) gemma-3-4b-it-7B

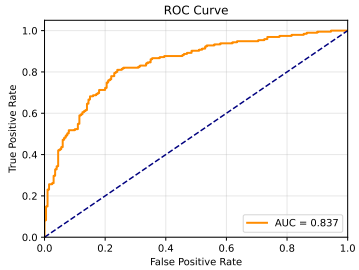
Figure 3: Distribution of importance scores



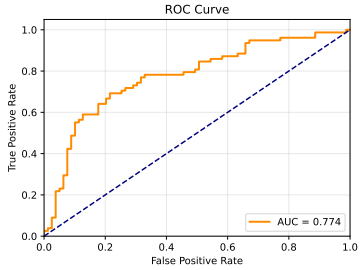
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



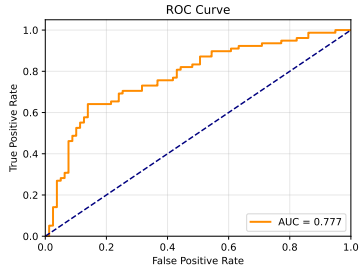
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



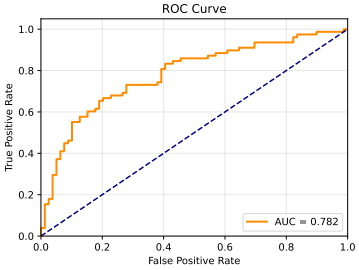
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



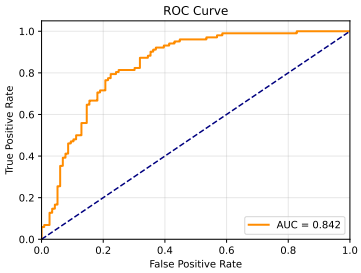
(d) DeepSeek-R1-Distill-Qwen-7B – L15



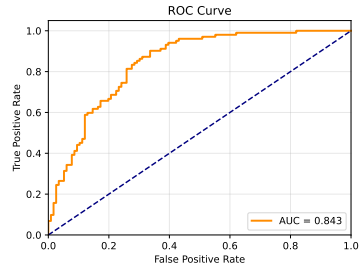
(e) DeepSeek-R1-Distill-Qwen-7B – L20



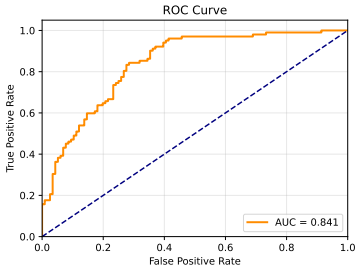
(f) DeepSeek-R1-Distill-Qwen-7B – L27



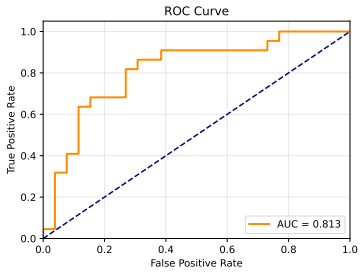
(g) DeepSeek-R1-Distill-Llama-8B – L17



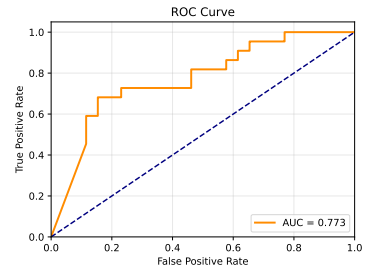
(h) DeepSeek-R1-Distill-Llama-8B – L23



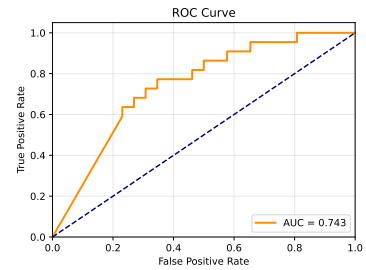
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

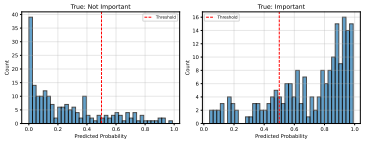


(k) gemma-3-4b-it – L24

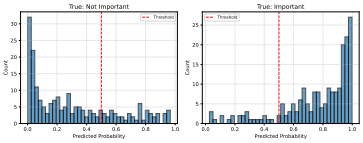


(l) gemma-3-4b-it – L33

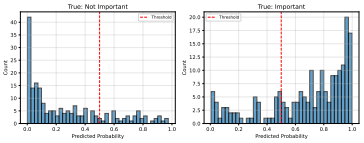
Figure 4: ROC curve



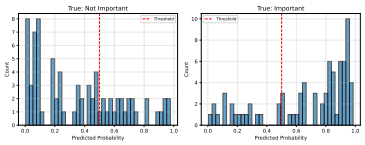
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



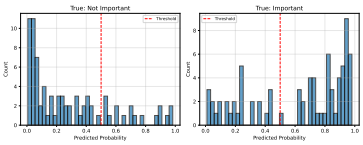
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



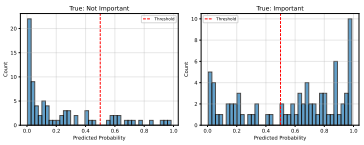
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



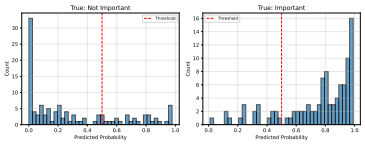
(d) DeepSeek-R1-Distill-Qwen-7B – L15



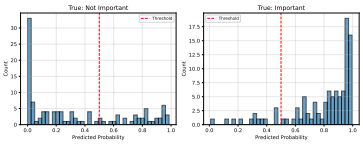
(e) DeepSeek-R1-Distill-Qwen-7B – L20



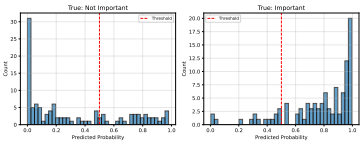
(f) DeepSeek-R1-Distill-Qwen-7B – L27



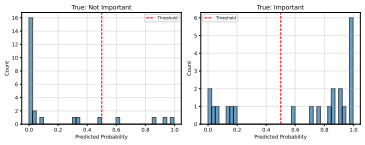
(g) DeepSeek-R1-Distill-Llama-8B – L17



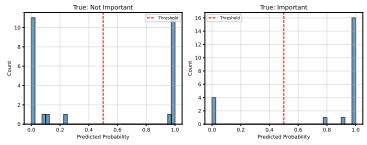
(h) DeepSeek-R1-Distill-Llama-8B – L23



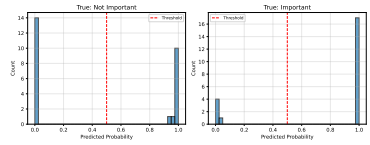
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

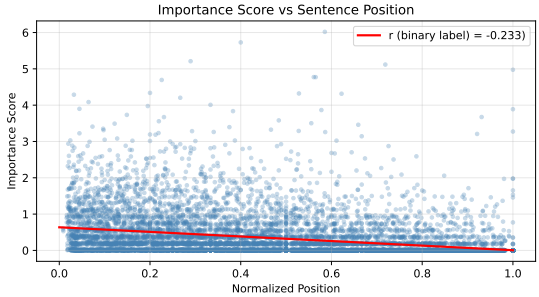


(k) gemma-3-4b-it – L24

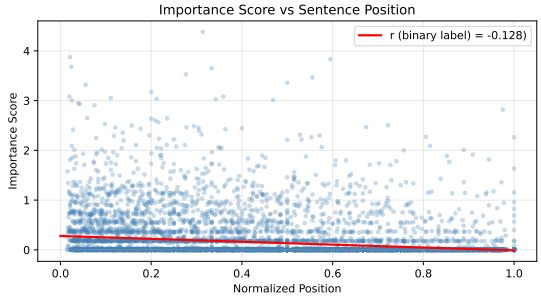


(l) gemma-3-4b-it – L33

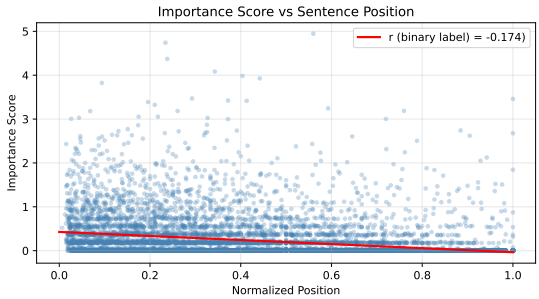
Figure 5: Predicted probability distribution of (not) important data points



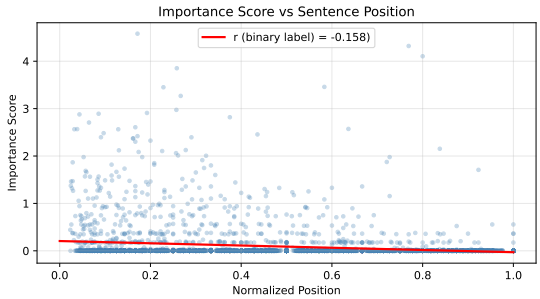
(a) DeepSeek-R1-Distill-Qwen-1.5B



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



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



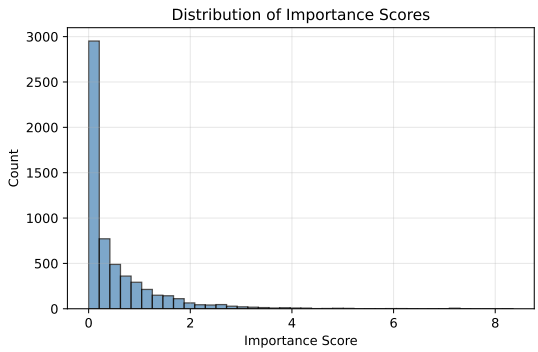
(d) gemma-3-4b-it

Figure 6: Correlation between importance and relative position

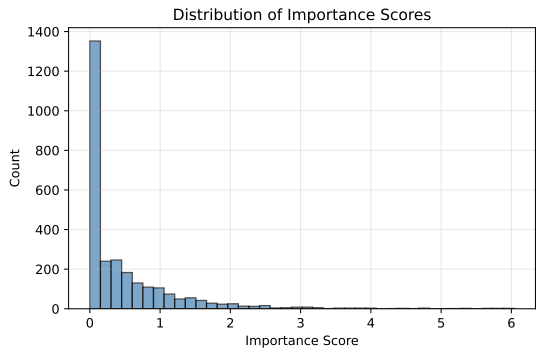
B.1.2 Counterfactual importance

B.2 MATH

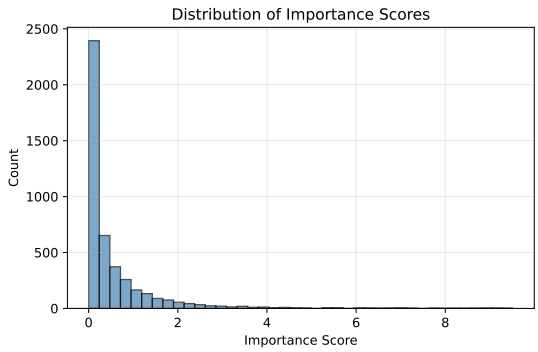
B.2.1 Resampling importance



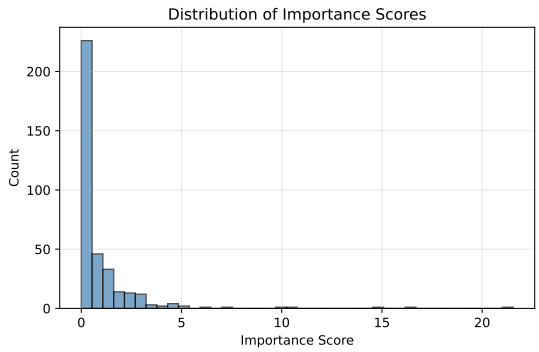
(a) DeepSeek-R1-Distill-Qwen-1.5B



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



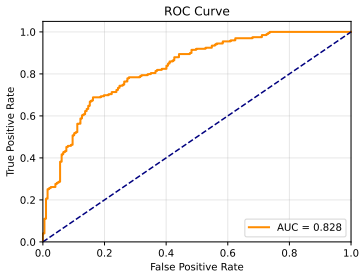
(c) DeepSeek-R1-Distill-Llama-8B



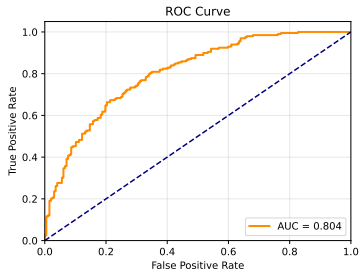
(d) gemma-3-4b-it-7B

Figure 7: Distribution of importance scores

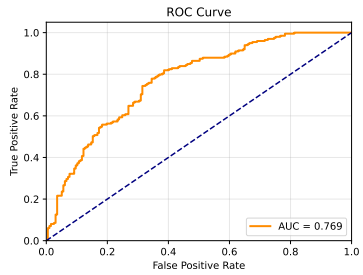
B.2.2 Counterfactual importance



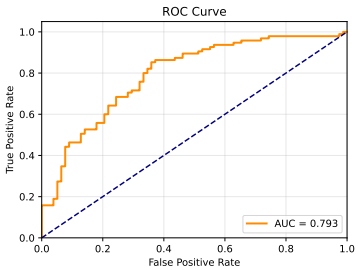
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



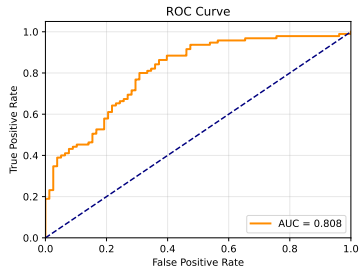
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



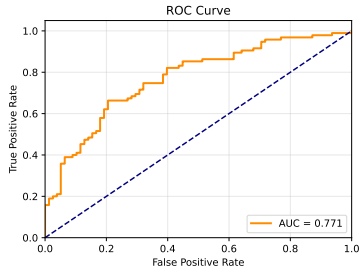
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



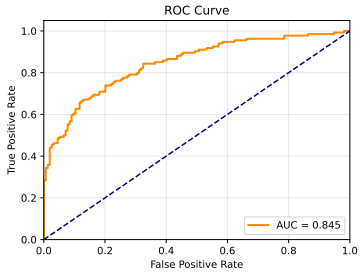
(d) DeepSeek-R1-Distill-Qwen-7B – L15



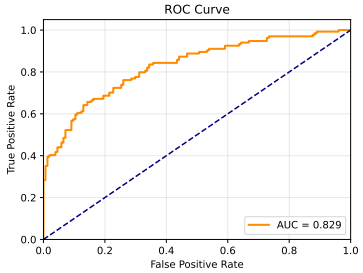
(e) DeepSeek-R1-Distill-Qwen-7B – L20



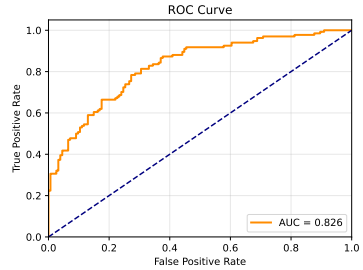
(f) DeepSeek-R1-Distill-Qwen-7B – L27



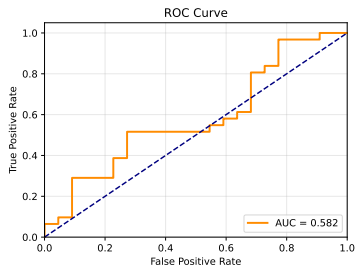
(g) DeepSeek-R1-Distill-Llama-8B – L17



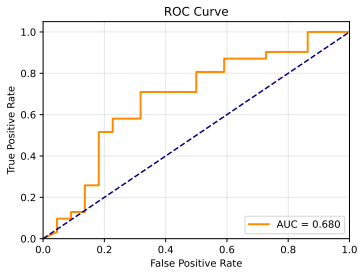
(h) DeepSeek-R1-Distill-Llama-8B – L23



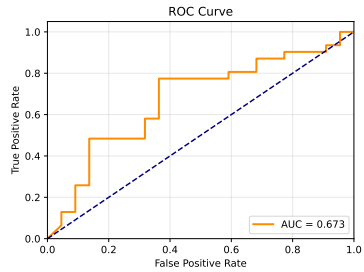
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

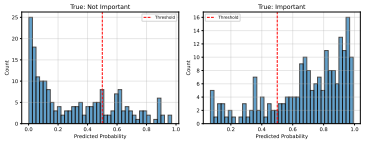


(k) gemma-3-4b-it – L24

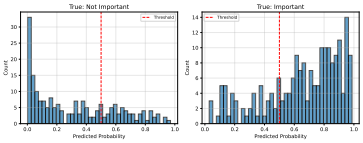


(l) gemma-3-4b-it – L33

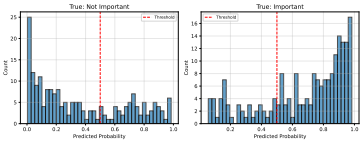
Figure 8: ROC curve



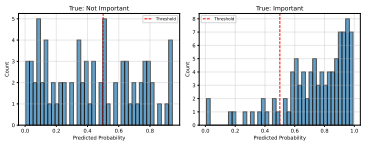
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



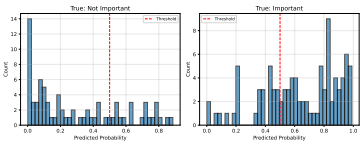
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



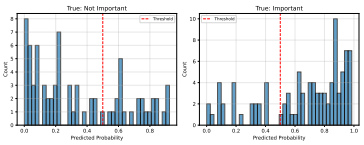
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



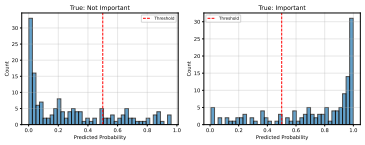
(d) DeepSeek-R1-Distill-Qwen-7B – L15



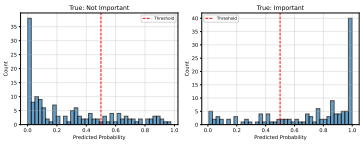
(e) DeepSeek-R1-Distill-Qwen-7B – L20



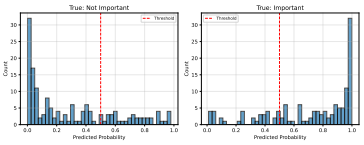
(f) DeepSeek-R1-Distill-Qwen-7B – L27



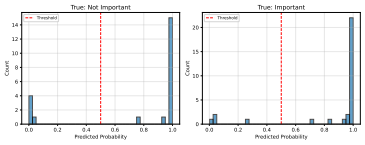
(g) DeepSeek-R1-Distill-Llama-8B – L17



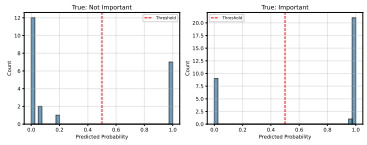
(h) DeepSeek-R1-Distill-Llama-8B – L23



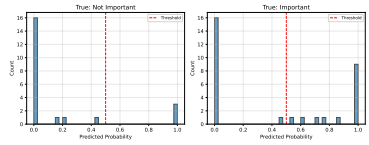
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

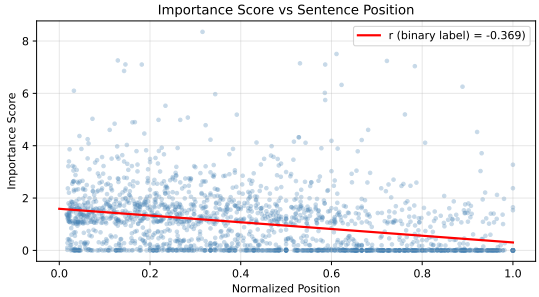


(k) gemma-3-4b-it – L24

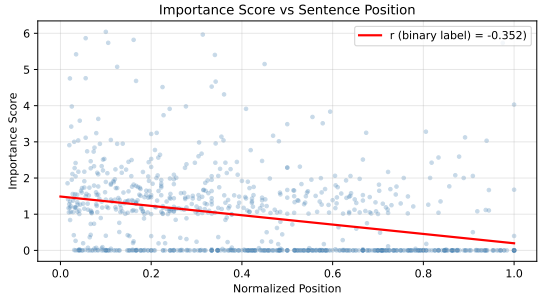


(l) gemma-3-4b-it – L33

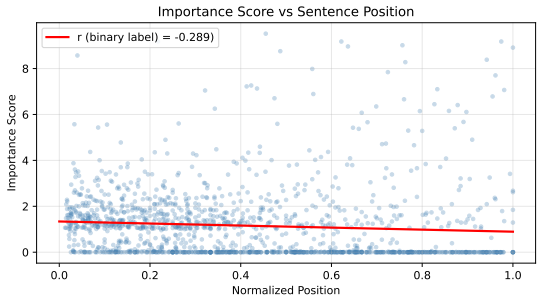
Figure 9: Predicted probability distribution of (not) important data points



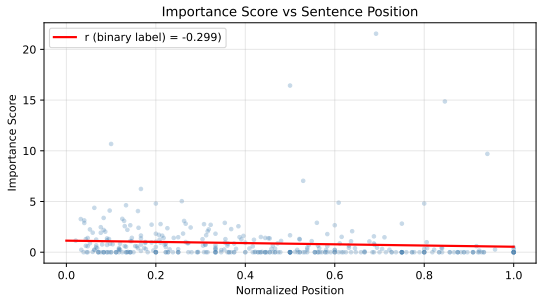
(a) DeepSeek-R1-Distill-Qwen-1.5B



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

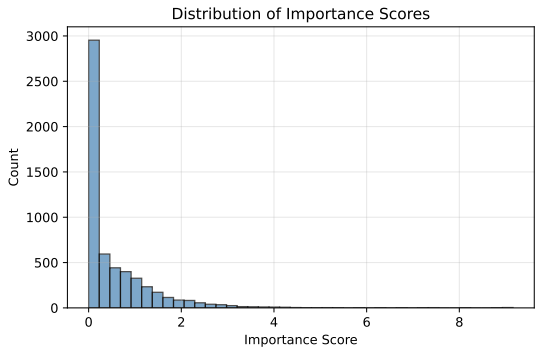


(c) DeepSeek-R1-Distill-Llama-8B

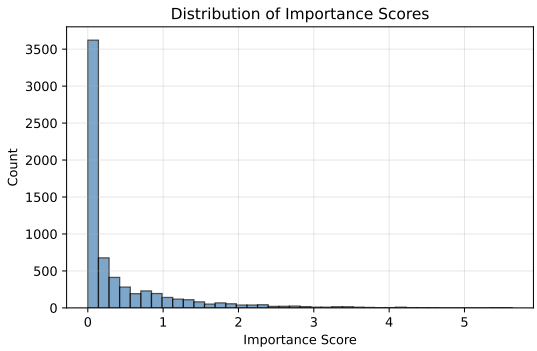


(d) gemma-3-4b-it

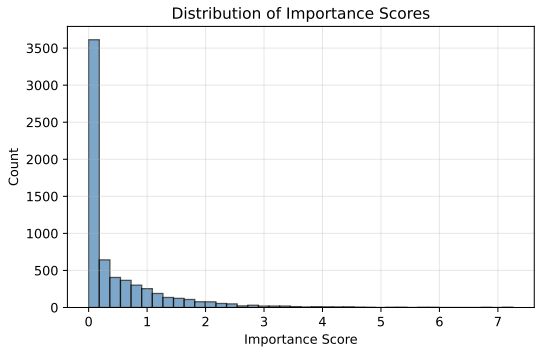
Figure 10: Correlation between importance and relative position



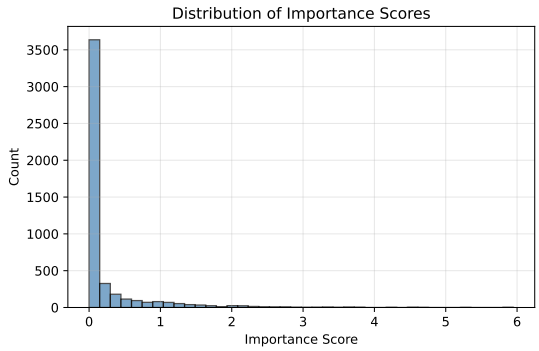
(a) DeepSeek-R1-Distill-Qwen-1.5B



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

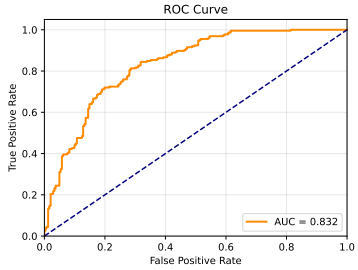


(c) DeepSeek-R1-Distill-Llama-8B

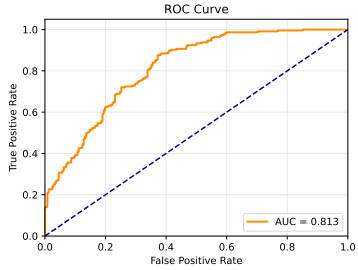


(d) gemma-3-4b-it-7B

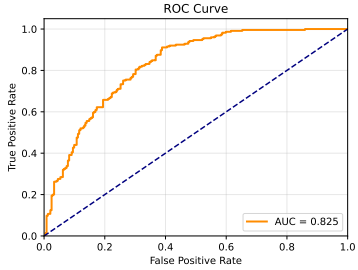
Figure 11: Distribution of importance scores



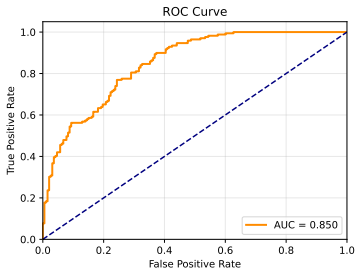
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



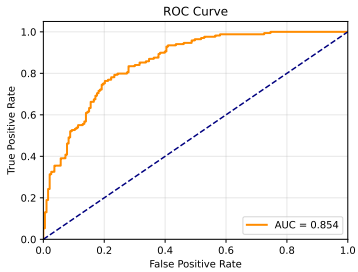
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



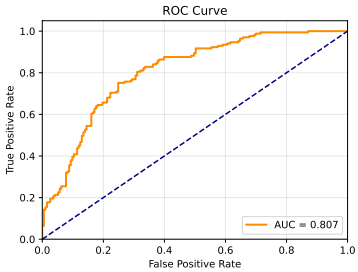
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



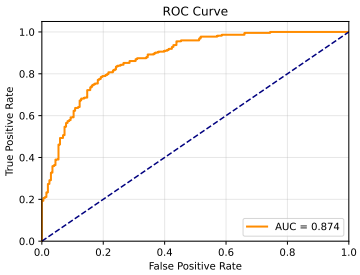
(d) DeepSeek-R1-Distill-Qwen-7B – L15



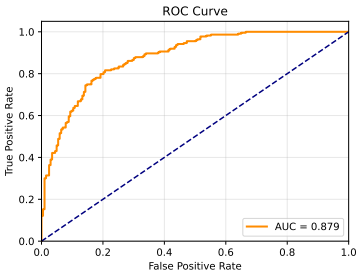
(e) DeepSeek-R1-Distill-Qwen-7B – L20



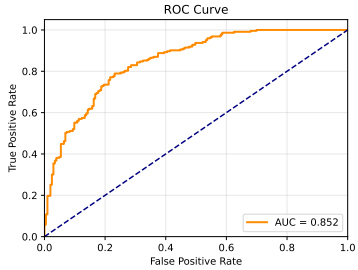
(f) DeepSeek-R1-Distill-Qwen-7B – L27



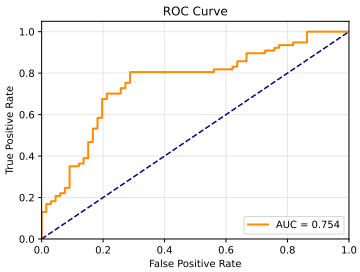
(g) DeepSeek-R1-Distill-Llama-8B – L17



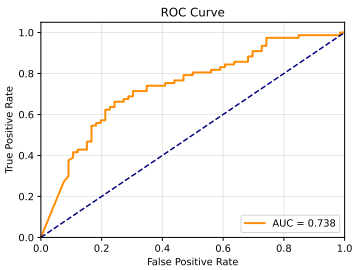
(h) DeepSeek-R1-Distill-Llama-8B – L23



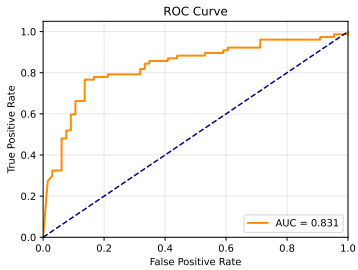
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

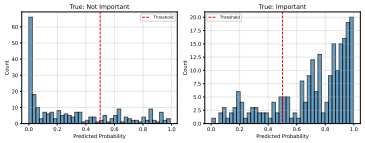


(k) gemma-3-4b-it – L24

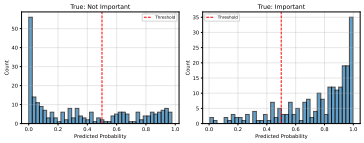


(l) gemma-3-4b-it – L33

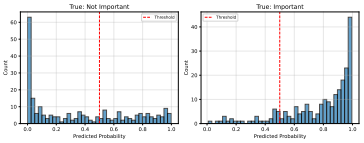
Figure 12: ROC curve



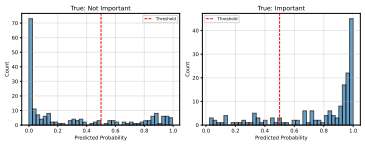
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



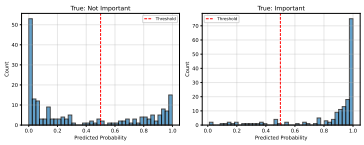
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



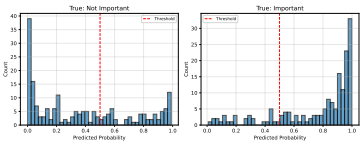
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



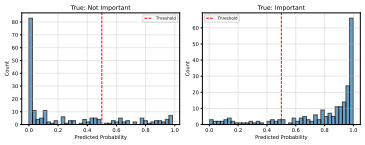
(d) DeepSeek-R1-Distill-Qwen-7B – L15



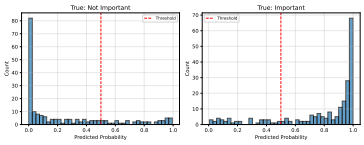
(e) DeepSeek-R1-Distill-Qwen-7B – L20



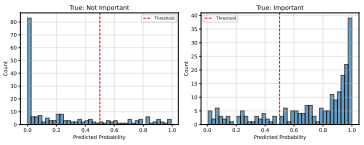
(f) DeepSeek-R1-Distill-Qwen-7B – L27



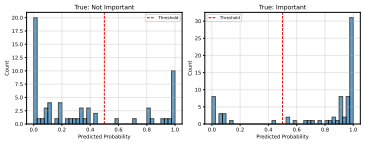
(g) DeepSeek-R1-Distill-Llama-8B – L17



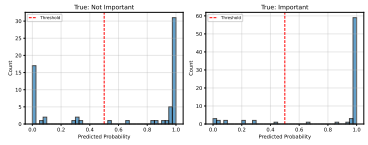
(h) DeepSeek-R1-Distill-Llama-8B – L23



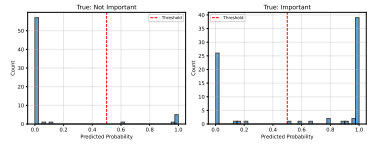
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

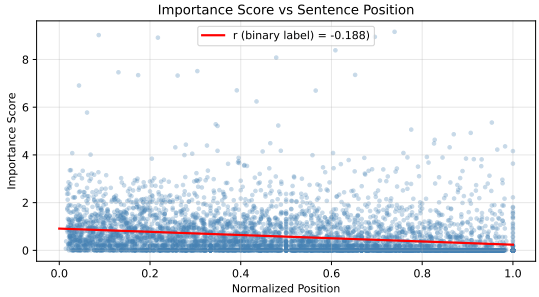


(k) gemma-3-4b-it – L24

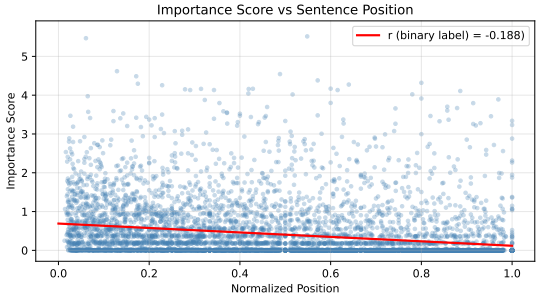


(l) gemma-3-4b-it – L33

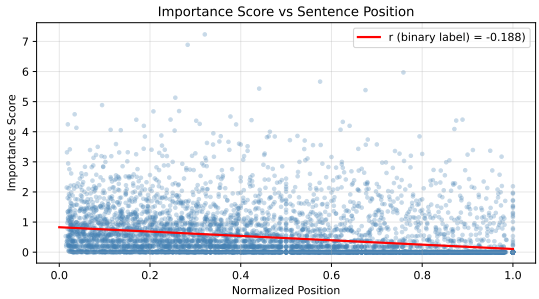
Figure 13: Predicted probability distribution of (not) important data points



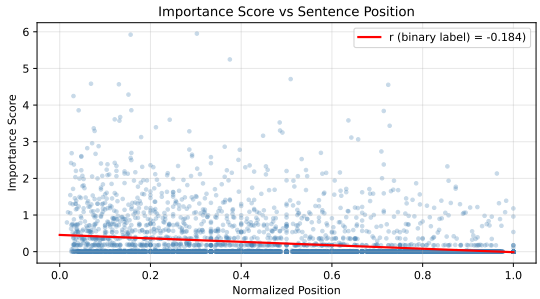
(a) DeepSeek-R1-Distill-Qwen-1.5B



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

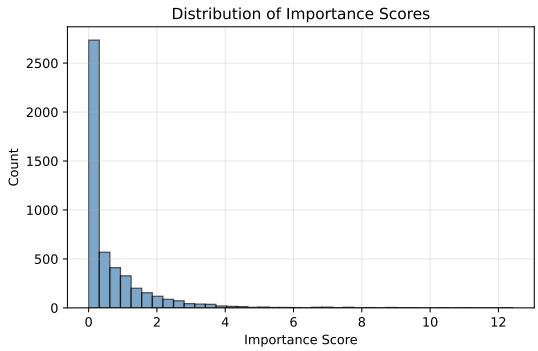


(c) DeepSeek-R1-Distill-Llama-8B

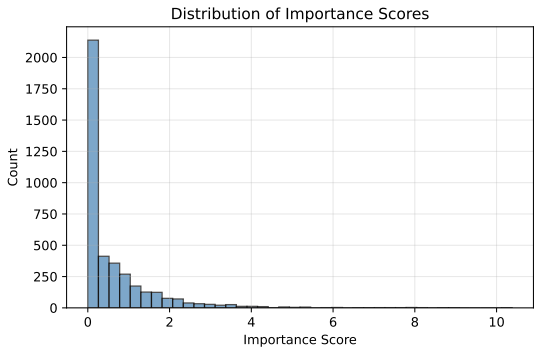


(d) gemma-3-4b-it

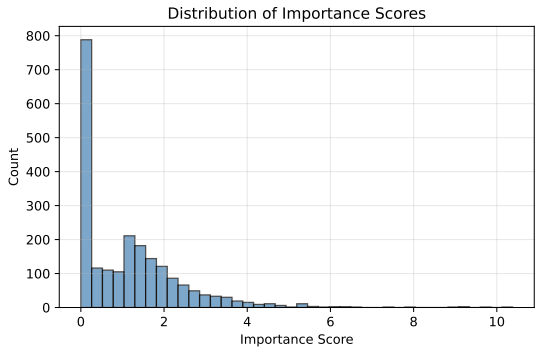
Figure 14: Correlation between importance and relative position



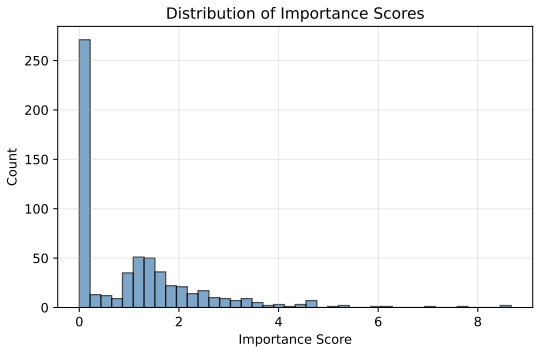
(a) DeepSeek-R1-Distill-Qwen-1.5B



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

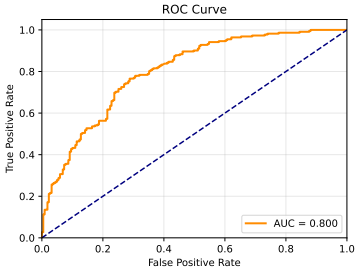


(c) DeepSeek-R1-Distill-Llama-8B

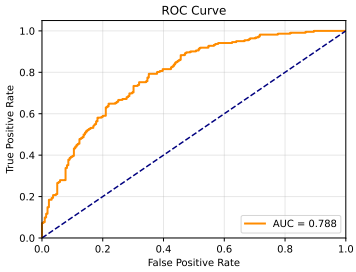


(d) gemma-3-4b-it-7B

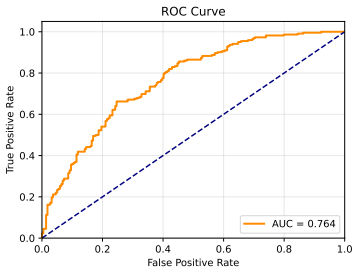
Figure 15: Distribution of importance scores



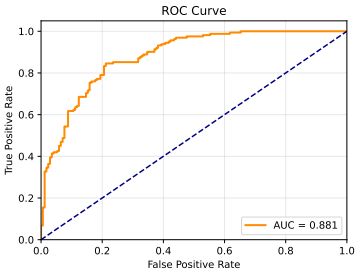
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



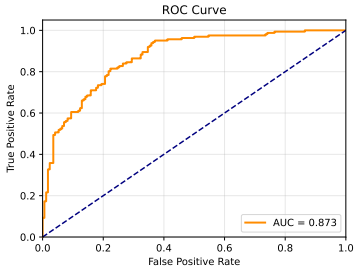
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



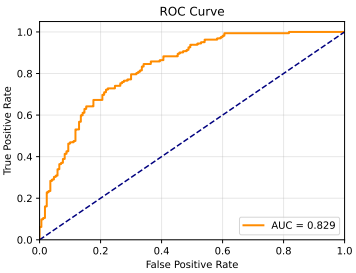
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



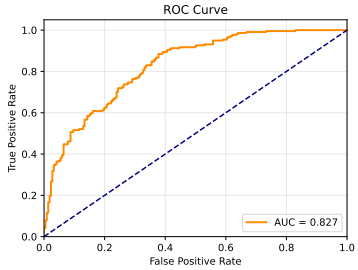
(d) DeepSeek-R1-Distill-Qwen-7B – L15



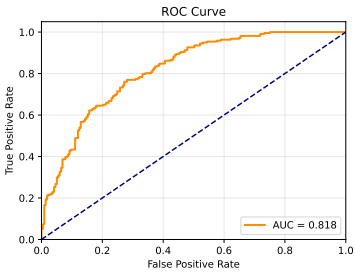
(e) DeepSeek-R1-Distill-Qwen-7B – L20



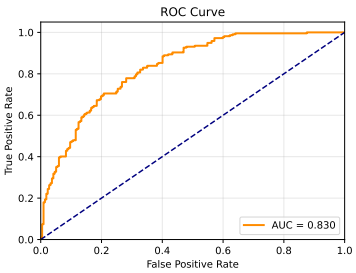
(f) DeepSeek-R1-Distill-Qwen-7B – L27



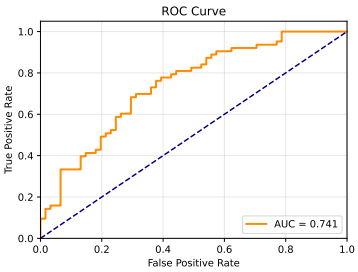
(g) DeepSeek-R1-Distill-Llama-8B – L17



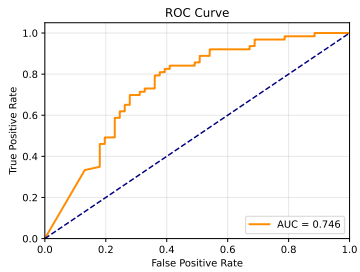
(h) DeepSeek-R1-Distill-Llama-8B – L23



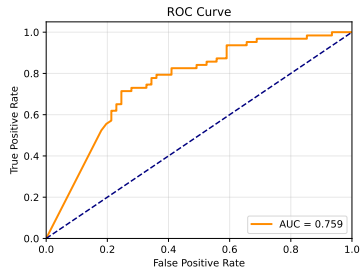
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

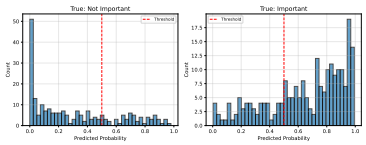


(k) gemma-3-4b-it – L24

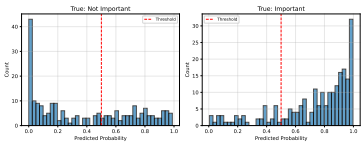


(l) gemma-3-4b-it – L33

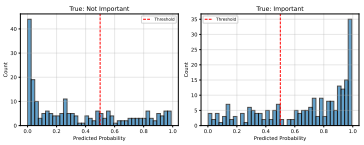
Figure 16: ROC curve



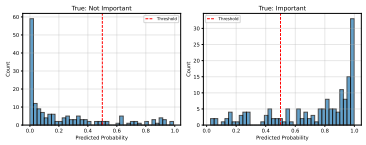
(a) DeepSeek-R1-Distill-Qwen-1.5B – L15



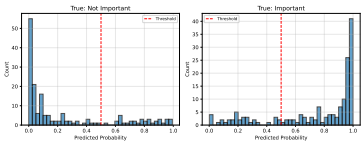
(b) DeepSeek-R1-Distill-Qwen-1.5B – L20



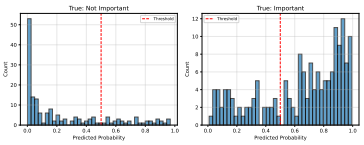
(c) DeepSeek-R1-Distill-Qwen-1.5B – L27



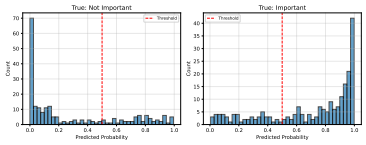
(d) DeepSeek-R1-Distill-Qwen-7B – L15



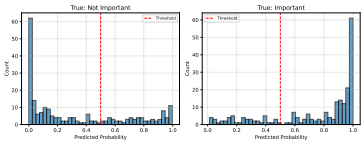
(e) DeepSeek-R1-Distill-Qwen-7B – L20



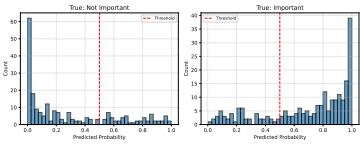
(f) DeepSeek-R1-Distill-Qwen-7B – L27



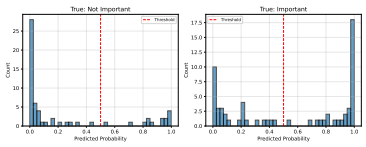
(g) DeepSeek-R1-Distill-Llama-8B – L17



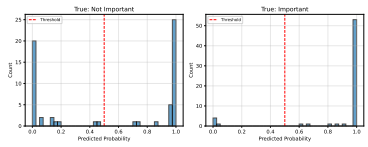
(h) DeepSeek-R1-Distill-Llama-8B – L23



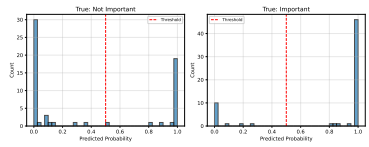
(i) DeepSeek-R1-Distill-Llama-8B – L31



(j) gemma-3-4b-it – L18

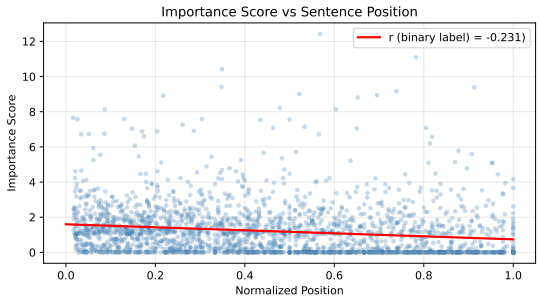


(k) gemma-3-4b-it – L24

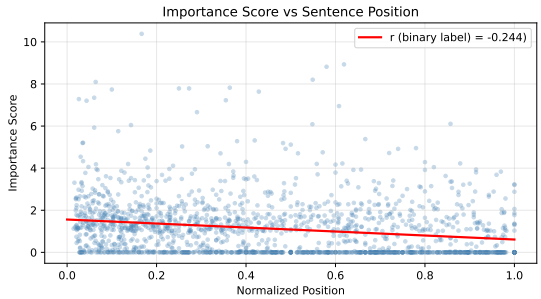


(l) gemma-3-4b-it – L33

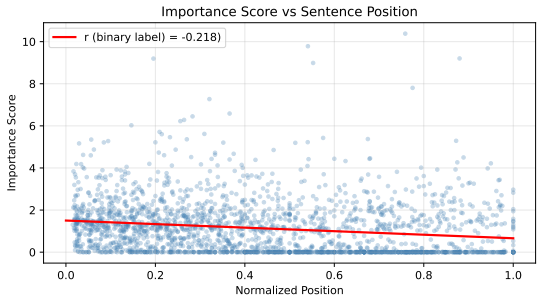
Figure 17: Predicted probability distribution of (not) important data points



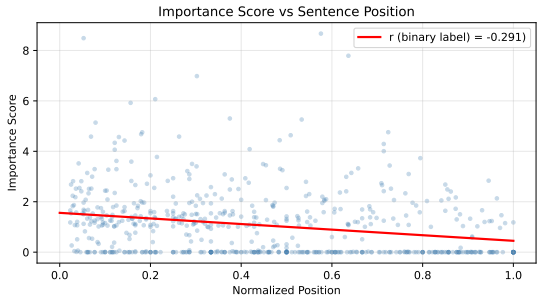
(a) DeepSeek-R1-Distill-Qwen-1.5B



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



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



(d) gemma-3-4b-it

Figure 18: Correlation between importance and relative position