
Antidistillation Sampling

Yash Savani* Asher Trockman* Zhili Feng Avi Schwarzschild
Alexander Robey Marc Finzi J. Zico Kolter
Carnegie Mellon University
{ysavani, ashert}@cs.cmu.edu

<https://antidistillation.com>

Abstract

Frontier models that generate extended reasoning traces inadvertently produce rich token sequences that can facilitate model distillation. Recognizing this vulnerability, model owners may seek sampling strategies that limit the effectiveness of distillation without compromising model performance. *Antidistillation sampling* provides exactly this capability. By strategically modifying a model’s next-token probability distribution, antidistillation sampling poisons reasoning traces, rendering them significantly less effective for distillation while preserving the model’s practical utility.

1 Introduction

Large language models (LLMs) trained to produce extended reasoning traces demonstrate impressive performance across math, coding, and general reasoning benchmarks [e.g 1]. These generated traces, however, serve a dual purpose—not only do they enhance model capabilities, but they also facilitate model distillation, wherein a secondary model learns to replicate the original model’s abilities by training on its generated traces [2, 3]. Notably, model distillation enables substantial capability gains in the secondary model at a fraction of the computational cost required for training a similarly performant model from scratch [3].

Despite the benefits of model distillation, the effectiveness and efficiency of this technique present several downsides for companies serving frontier reasoning models. First, returning extended reasoning traces constitutes a forfeiture of valuable intellectual property, enabling competitors to cheaply replicate frontier capabilities. Second, the possibility of distillation incentivizes frontier model providers to hide token probabilities, summarize reasoning traces, or, more generally, restrict user-model interaction. And finally, safe model behaviors (e.g., resistance to jailbreaking attempts [4, 5]) are often not inherited by distilled models, enabling the generation of objectionable content [6].

To address these issues, we introduce *antidistillation sampling*—a method designed to reduce the effectiveness of model distillation. The core idea underpinning antidistillation sampling is to adjust a reasoning model’s sampling distribution so that generated traces (1) poison distillation attempts while (2) maintain high likelihood under the original, unadjusted distribution. This approach protects proprietary capabilities while preserving the original model’s utility for downstream applications.

*Equal contribution.

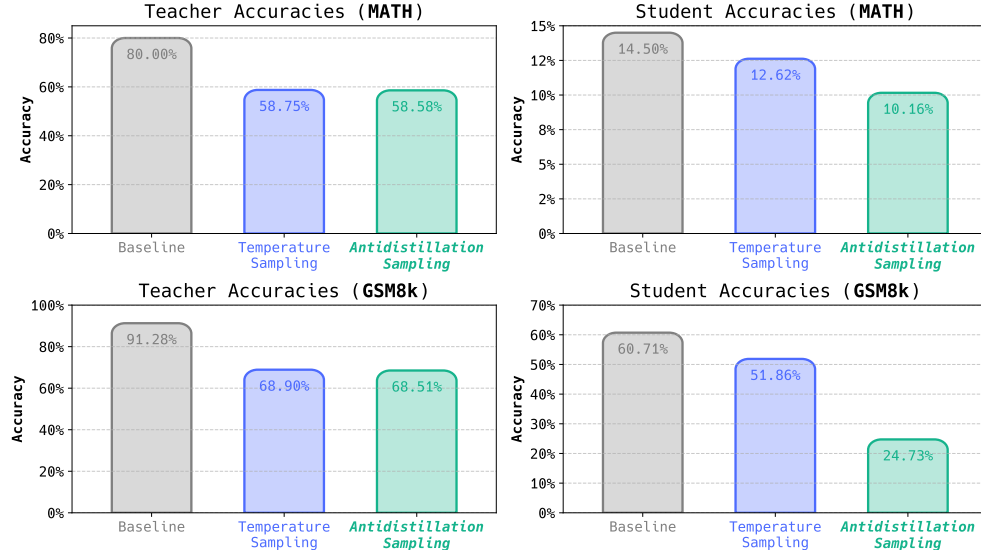


Figure 1: Reasoning traces generated via antidistillation sampling poison distillation attempts, while simultaneously preserving the teacher’s downstream performance. The top and bottom rows show results for MATH [7] and GSM8K [8], respectively. The left column shows teacher accuracies under different sampling schemes, and the right column shows the performance of students distilled on traces sampled via these different sampling schemes. Notably, for the same teacher performance on both datasets, antidistillation sampling significantly degrades the performance of the distilled models relative to temperature sampling.

1.1 Related Work

Various frontier AI labs acknowledge and benefit from the effectiveness of model distillation. For instance, OpenAI offers model distillation as a service within their API [9]. The recognition of model distillation’s potential, which motivates such pipelines, dates back at least to Schmidhuber [10]. More recently, Hinton et al. [2] demonstrated that distilled “specialist” models can achieve impressive performance in multiple domains. Since then, a growing body of work aims to understand the transference of capabilities via distillation [11–14]. In fact, there is speculation that some labs train commercial-facing LLMs in part via distillation, possibly by harvesting extended reasoning traces generated from models belonging to their direct competitors [15]. This practice constitutes a strategic vulnerability for frontier model maintainers, given the demonstrated value of these reasoning traces, and underscores the importance of the algorithm we introduce in this work.

The threat model we address in this paper—where a student model is trained on data generated by a teacher model—intersects with several aspects of model privacy and security. For instance, model extraction attacks acquire weights via query-level access without training or distillation [16] and training data extraction attacks harvest training data directly from frontier models [17]. While antidistillation sampling may offer some protection against these attacks, such analysis remains beyond our current scope. More relevant is the literature on data poisoning, where maliciously crafted data is injected into a model’s training set to induce specific unintended downstream effects (see, e.g., [18]). Rando and Tramèr [19] even show the effectiveness of adding backdoors to preference data, sabotaging models fine-tuned with RLHF. Our contribution bridges data poisoning and privacy techniques to protect the valuable knowledge encoded in frontier models.

Finally, we position antidistillation sampling within the broader framework of controlled decoding for LLMs [20], where supplementary objectives steer the decoding process. Existing approaches in this domain include leveraging contrastive objectives to enhance generation quality [21], reformulating constrained decoding as an optimization problem [22], and

incorporating energy-based constraints [23]. While related, antidistillation sampling solves a different problem: by implementing a new, distillation-aware penalization term in the decoding objective, our approach poisons generated reasoning traces to undermine the performance of models fine-tuned on these outputs.

2 Antidistillation Sampling

To introduce and derive antidistillation sampling, we first provide a high-level sketch, followed by the desiderata outlining the desired qualities for poisoning distillation attempts. We then derive the antidistillation sampling method in §2.2. Algorithm 1 summarizes the key steps involved in implementing our approach.

An overview of our approach. The core objective of antidistillation sampling is to adjust a model’s next-token distribution to balance two competing goals: sampling tokens with high likelihood under the original, unadjusted distribution and sampling tokens that effectively poison distillation attempts. Throughout, we refer to the model from which reasoning traces are sampled as the *teacher*, and the model being distilled as the *student*. We begin by introducing notation describing sampling from the teacher’s unadjusted distribution. Next, we quantify how model distillation impacts the student model’s performance on a given downstream task. This analysis yields a key insight—we can incorporate this performance metric directly into the teacher’s sampling distribution. This takes the form of a directional derivative capturing the change in the teacher’s sampling distribution along the update direction in the student’s weight space. However, due to the high cost needed to compute this directional derivative, the final portion of our derivation identifies an efficient finite-difference approximation for this term, which is inexpensive to compute and, as we demonstrate in §3, results in effective distillation poisoning.

2.1 Preliminaries

To formalize our presentation of antidistillation, we first introduce notation. We consider an LLM to be a mapping from a sequence of input tokens $x_{1:t} = (x_1, \dots, x_t)$ to a distribution over the next token, where each token x_j is an element of a vocabulary set $\mathcal{V} = \{1, \dots, V\}$. This distribution is parameterized by weights θ and can be expressed as follows.

$$p(\cdot | x_{1:t}; \theta) : \mathcal{V}^t \rightarrow [0, 1]^V \quad (1)$$

We write $p(\cdot | x_{1:t}; \theta)$ to denote the distribution of all next-token probabilities, whereas $p(x_{t+1} | x_{1:t}; \theta)$ refers to the scalar probability of a given next token x_{t+1} . Typically, tokens are generated according to a scaled version of this distribution by sampling as follows:²

$$x_{t+1} \sim \frac{1}{Z} \exp(\log p(\cdot | x_{1:t}; \theta) / \tau). \quad (2)$$

Here, τ is the temperature and Z is a normalization term, which is computed by summing the exponential term over all possible next tokens. Using a temperature of $\tau = 0$ corresponds to greedy sampling, in which x_{t+1} is deterministically chosen to be the token with the largest log probability under the current model parameter θ .

Desiderata for antidistillation sampling. Model distillation involves a *student* language model—parameterized by θ_S , with a distribution over next tokens given by $p(\cdot | x_{1:t}; \theta_S)$ —trained on data generated from a *teacher* model parameterized by θ_T . These models do not need to share the same parameter space, and therefore the parameter vectors θ_S and θ_T need not be comparable; indeed, in many cases, a student model may have substantially fewer parameters than the teacher.

The aim of antidistillation sampling is to generate tokens from the teacher θ_T , which perform well according to a metric used to evaluate teacher samples, while simultaneously having

²Variants of this sampling scheme include top- k sampling (i.e., limiting sampling to the tokens with the top- k largest probabilities), greedy sampling (i.e., sampling from the same objective while letting $\tau \rightarrow 0$), and beam search, but we focus mainly on temperature-based sampling here.

the property that training on these tokens *cannot* improve performance on this same task. In more detail, we aim to adjust the teacher’s sampling procedure to achieve the following objectives simultaneously:

- I. **Non-distillability.** Student models trained on tokens sampled via antidistillation sampling should have a degraded performance on a chosen downstream task relative to training on tokens sampled from the teacher’s nominal distribution.
- II. **Nominal utility.** Tokens sampled via antidistillation sampling should remain probable under the teacher’s unadjusted sampling scheme $p(\cdot | x_{1:t}; \theta_T)$.

Taken together, these goals ensure that the teacher model maintains its nominal performance while simultaneously preventing distillation on downstream tasks.

Proxy models. In general, we do not expect to know the distilled student’s model architecture in advance. Therefore, rather than assuming access to the true student model, we develop antidistillation sampling based on the notion of a *proxy student model*, which, for simplicity, we refer to as the *proxy model*. The proxy model is parameterized by θ_P , and specifies a sampling distribution $p(\cdot | x_{1:t}; \theta_P)$. A key aspect we consider below is whether the process *generalizes*, i.e., whether traces via antidistillation sampling to prevent the proxy model from distilling the teacher also prevent the distinct student models from distilling.

2.2 Deriving Antidistillation Sampling

To operationalize antidistillation sampling, we first assume access to a differentiable, real-valued *downstream loss* ℓ , which measures the proxy model’s performance on a given downstream task. Throughout, we take ℓ to be the negative log-likelihood for generating a sequence of tokens on a fixed, potentially large dataset. For instance, ℓ could represent the cross entropy loss of predicting each token across a large reasoning benchmark. However, ℓ can be chosen very broadly to capture any student capability that the teacher model maintainer may want to influence via poisoning. A key point is that ℓ can be very costly to compute, as it may require evaluating the proxy model over a large and diverse set of data.

Given the non-distillability criteria outlined above, the goal of antidistillation sampling is for the downstream loss $\ell(\theta_P)$ to increase³ whenever the student is fine-tuned on sequences of tokens generated by the teacher. To capture this, first consider the change in θ_P that results from fine-tuning to minimize the negative log-likelihood of a token x_{t+1} generated by the teacher. Specifically, we consider one step of optimization via gradient descent on θ_P :

$$\theta_P^+ = \theta_P - \eta \nabla_{\theta_P} (-\log p(x_{t+1} | x_{1:t}; \theta_P)) \quad (3)$$

$$= \theta_P + \eta \nabla_{\theta_P} \log p(x_{t+1} | x_{1:t}; \theta_P) \quad (4)$$

where $\eta > 0$ is the step size. The impact of this update can then be quantified by measuring the difference in the loss ℓ before and after this update. In particular, for each token $x_{t+1} \in \mathcal{V}$, we define the following difference term

$$\Delta(x_{t+1} | x_{1:t}) = \ell(\theta_P^+) - \ell(\theta_P) = \ell(\theta_P + \eta \nabla_{\theta_P} \log p(x_{t+1} | x_{1:t}; \theta_P)) - \ell(\theta_P). \quad (5)$$

If $\Delta(x_{t+1} | x_{1:t})$ is positive, the update in eq. (4) increases the loss; if $\Delta(x_{t+1} | x_{1:t})$ is negative, the update decreases the loss. Thus, our goal is to adjust the teacher’s sampling distribution so that tokens sampled from the teacher both have (1) high likelihood under the teacher’s unadjusted distribution and (2) yield larger (i.e., more positive) values of Δ .

To implement antidistillation sampling, we propose adding a penalty, proportional to $\Delta(x_{t+1} | x_{1:t})$, to the teacher’s unadjusted log probabilities $\log p(x_{t+1} | x_{1:t}; \theta_T)$. This results in the following adjusted sampling distribution

$$x_{t+1} \sim \frac{1}{Z} \exp(\log p(\cdot | x_{1:t}; \theta_T) / \tau + \lambda \Delta(\cdot | x_{1:t}; \theta_P)), \quad (6)$$

³We assume without loss of generality that increases in $\ell(\theta_P)$ are desirable from the perspective of the poisoner; the procedure is easily adaptable to problems wherein the goal is to decrease $\ell(\theta_P)$.

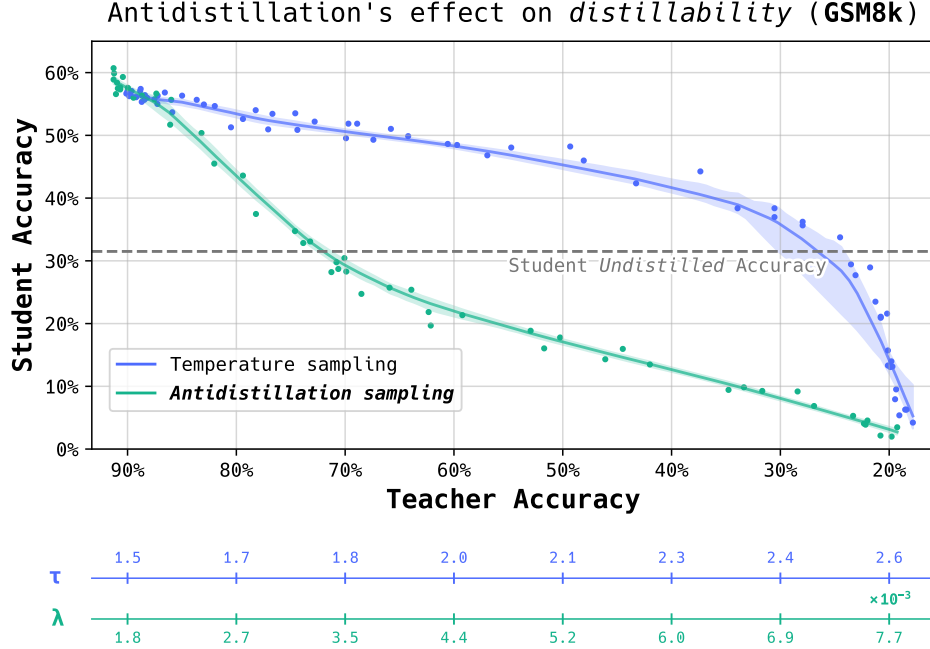


Figure 2: Antidistillation sampling uses a tunable parameter λ to control the trade-off between teacher accuracy and distillability. The baseline involves sampling from the teacher with increasing temperature τ to show that we can produce traces that are bad for distillation at some cost in teacher accuracy. One important feature of the blue temperature sampling curve is that to bring the student accuracy down below the undistilled accuracy, the teacher performance has to drop to 20%. On the other hand, with antidistillation sampling, the teacher model can still get 70% accuracy while producing traces that bring the student’s performs down below the undistilled accuracy.

where Z is a normalization term appropriately scaled (relative to eq. (2)) to accommodate the penalty, and $\lambda > 0$ is a regularization coefficient that facilitates a trade-off between sampling from the teacher’s distribution and sampling tokens that maximally increase student’s downstream loss. Unfortunately, directly implementing eq. (6) is impractical, as we would need to compute $\Delta(x_{t+1}|x_{1:t})$ for each potential next token $x_{t+1} \in \mathcal{V}$, requiring V gradients to be computed as well as V evaluations of the downstream loss ℓ , which, in turn, is assumed to involve a lengthy computation to produce.

An efficient implementation of antidistillation sampling. The core of our proposed approach is an efficient mechanism to approximate the sampling process above. As a starting point, observe that $\Delta(x_{t+1}|x_{1:t})$ can be scaled by a factor of $1/\eta$ without changing the relative penalties for each x_{t+1} (i.e., we could fold this term into the λ regularization penalty). Then, by taking the limit of $\Delta(x_{t+1}|x_{1:t})/\eta$ as $\eta \rightarrow 0$, we have that

$$\lim_{\eta \rightarrow 0} \frac{1}{\eta} \Delta(x_{t+1}|x_{1:t}) = \lim_{\eta \rightarrow 0} \frac{\ell(\theta_P + \eta \nabla_{\theta_P} \log p(x_{t+1}|x_{1:t}; \theta_P)) - \ell(\theta_P)}{\eta} \quad (7)$$

$$= \langle \nabla \ell(\theta_P), \nabla_{\theta_P} \log p(x_{t+1}|x_{1:t}; \theta_P) \rangle. \quad (8)$$

That is, the limit is the inner product between the gradient $\nabla_{\theta_P} \log p(x_{t+1}|x_{1:t}; \theta_P)$ and the downstream loss gradient $\nabla \ell(\theta_P)$. Notice that the expression in eq. (8) no longer involves the evaluation of the downstream loss for each token in \mathcal{V} . Rather, $\nabla \ell(\theta_P)$ can be computed and stored once, after which the only remaining task is to efficiently evaluate eq. (8) for each token $x_{t+1} \in \mathcal{V}$. To do so, the key observation is that the directional derivative is symmetrical. Thus, we can rewrite eq. (8) as a finite difference limit in the *other* term, i.e., in

Algorithm 1: Antidistillation sampling

Input: Prompt $x_{1:n}$, max tokens N , penalty multiplier λ , approximation parameter ϵ , temperature τ

1. (Initialization) Compute the gradient of the downstream loss

$$g \leftarrow \nabla \ell(\theta_P)$$

2. For each token index $t = n, n + 1, \dots, N - 1$:

- i. Compute the antidistillation penalty term

$$\widehat{\Delta}(\cdot | x_{1:t}) \leftarrow \frac{\log p(\cdot | x_{1:t}; \theta_P + \epsilon g) - \log p(\cdot | x_{1:t}; \theta_P)}{\epsilon}$$

- ii. Sample the next token x_{t+1} from the teacher’s adjusted distribution

$$x_{t+1} \sim \frac{1}{Z} \exp \left(\frac{1}{\tau} \log p(\cdot | x_{1:t}; \theta_T) + \lambda \widehat{\Delta}(\cdot | x_{1:t}) \right)$$

Output: Sampled sequence $x_{1:N}$

terms of a finite difference update to $\log p(x_{t+1} | x_{1:t}; \theta_P)$. This gives

$$\lim_{\eta \rightarrow 0} \frac{1}{\eta} \Delta(x_{t+1} | x_{1:t}) = \langle \nabla \ell(\theta_P), \nabla_{\theta_P} \log p(x_{t+1} | x_{1:t}; \theta_P) \rangle \quad (9)$$

$$= \lim_{\epsilon \rightarrow 0} \frac{\log p(x_{t+1} | x_{1:t}; \theta_P + \epsilon \nabla \ell(\theta_P)) - \log p(x_{t+1} | x_{1:t}; \theta_P)}{\epsilon} \quad (10)$$

Importantly, this difference involves *only* the computation of next-token probabilities under two different models: the original proxy model θ_P and an updated copy of the proxy model $\theta_P + \epsilon \nabla \ell(\theta_P)$. These models can be saved once before any sampling, and then an approximation of the antidistillation sampling term can be computed for *all* next tokens simply via two forward passes in the proxy model. In other words, we define

$$\widehat{\Delta}(\cdot | x_{1:t}) = \frac{\log p(\cdot | x_{1:t}; \theta_P + \epsilon \nabla \ell(\theta_P)) - \log p(\cdot | x_{1:t}; \theta_P)}{\epsilon} \quad (11)$$

for some appropriately chosen small value of ϵ ,⁴ where $\widehat{\Delta}(x_{t+1} | x_{1:t})$ approaches eq. (8) for all next tokens x_{t+1} in the limit as $\epsilon \rightarrow 0$. Finally, we sample according to the teacher’s adjusted sampling distribution:

$$x_{t+1} \sim \frac{1}{Z} \exp \left(\log p(\cdot | x_{1:t}; \theta_T) / \tau + \lambda \widehat{\Delta}(\cdot | x_{1:t}) \right). \quad (12)$$

In Algorithm 1, we summarize the procedure outlined in this section. Concretely, given a prompt $x_{1:t}$, using antidistillation sampling to generate a new token x_{t+1} involves: (1) (once, at initialization) computing the gradient of the downstream loss; and (2) (for each token to be generated) compute the finite-difference approximation of $\Delta(\cdot | x_{1:t})$ and sample the token from the teachers adjusted softmax distribution.

3 Empirical Results

Architectures and benchmarks. To demonstrate the effectiveness of antidistillation sampling in practice, we simulate real-world distillation by instantiating separate teacher, proxy

⁴It is also possible to compute the inner product of this gradient exactly via an explicit Jacobian-vector product (not to be confused with a vector-Jacobian product used in backpropagation). However, even in modern automatic differentiation frameworks, Jacobian-vector products tend lack support for a handful of important operations, such as SDPA. Meanwhile flash attention [24] only supports Float16 and BFloat16. Thus, we use finite differences for the improved convenience and efficiency, but we verify the correctness of the finite difference approximation against autograd in §C.1.

student, and student models. Concretely, we use deepseek-ai/DeepSeek-R1-Distill-Qwen-7B [25] as the teacher model, Qwen/Qwen2.5-3B [26] as the proxy model, meta-llama/Llama-3.2-3B [27] as the student model. We evaluate the teacher and the student performance on the GSM8K [8] and MATH [7] benchmarks, as strong performance on these benchmarks requires training on high-quality traces.

Baselines. Throughout, we compare to a straightforward sampling method: temperature sampling. In this scheme, tokens are sampled from the teacher’s distribution with some temperature τ ; this approximates what an API endpoint might be doing in the nominal case. We present comparisons to additional baselines in §A.

Hyperparameters. Antidistillation sampling involves two key hyperparameters: ϵ , which controls the approximation power of the finite-difference computation, and λ , which controls the degree to which the sampling distribution mixes in the antidistillation penalty. Regarding the choice of ϵ , we empirically verify that the finite difference approximation in eq. (11) is close to the JVP in eq. (8) on a smaller model in §C.1. In practice, we find that $\epsilon = 10^{-4}$ works reasonably well for the BFloat16 models we use. With regard to λ , we sweep over several values to study the trade off involved in perturbing the sampling distribution.

3.1 Main Results

In Figure 1, we provide evidence that antidistillation sampling effectively meets the desiderata laid out in §2.1. That is, for a fixed teacher accuracy, we observe significantly worse performance for students distilled from traces sampled via antidistillation sampling relative to models distilled from traces generated via temperature scaling. When performing distillation, we use LoRA [28] with rank 128, α equal to 128, dropout probability set to 0, learning rate equal to 0.0005, weight decay coefficient of 0.1, and the maximum gradient norm set to 1.0. We train with a cosine learning rate schedule with a warm-up parameter of 0.1, batch size of 32, and we train for 4 epochs.

In Figure 2, we vary λ to examine the degree of control antidistillation sampling gives model owners over the trade-off between teacher performance and distillability. We note that as we use distinct student and proxy student architectures, our results indicate that antidistillation sampling can generalize across architectures.

4 Conclusion

The value of proprietary frontier LLMs necessitates that their owners do what they can to protect their assets. As evidenced by the fact that the frontier companies limit exposure to their models via black-box APIs, these companies are already considering the threat of model stealing. However, given the recent attention paid to the effectiveness of distillation, it is imperative that model maintainers who wish to protect the information stored in their models guard against distillation. This paper provides a proof-of-concept that antidistillation sampling—which adjusts a model’s sampling distribution—is effective in blocking such attacks. We are excited at the prospect of continuing to refine and scale this approach, particularly with a view toward more secure future frontier models.

References

- [1] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

- [3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [4] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.
- [5] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.
- [6] Matt Burgess and Lily Hay Newman. Deepseek’s safety guardrails failed every test researchers threw at its ai chatbot. URL <https://www.wired.com/story/deepseeks-ai-jailbreak-prompt-injection-attacks/>.
- [7] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.
- [8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [9] OpenAI. Model distillation in the api. October 2024. URL <https://openai.com/index/api-model-distillation/>. Accessed: 2025-04-10.
- [10] Jürgen Schmidhuber. Learning complex, extended sequences using the principle of history compression. *Neural computation*, 4(2):234–242, 1992.
- [11] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, 2023.
- [12] Jierui Li and Raymond Mooney. Distilling algorithmic reasoning from llms via explaining solution programs. *arXiv preprint arXiv:2404.08148*, 2024.
- [13] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.
- [14] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3996–4003, 2020.
- [15] New York Times. Openai says it has evidence china’s deepseek used its model to train competitor. *The New York Times*, January 2025. URL <https://www.nytimes.com/2025/01/29/technology/openai-deepseek-data-harvest.html>. Accessed: 2025-04-10.
- [16] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. Stealing part of a production language model. *arXiv preprint arXiv:2403.06634*, 2024.
- [17] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.
- [18] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.

- [19] Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback. In *International Conference on Learning Representations*, 2024.
- [20] Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.
- [21] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022.
- [22] Haozhe Ji, Pei Ke, Hongning Wang, and Minlie Huang. Language model decoding as direct metrics optimization. *arXiv preprint arXiv:2310.01041*, 2023.
- [23] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551, 2022.
- [24] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.
- [25] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- [26] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [27] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

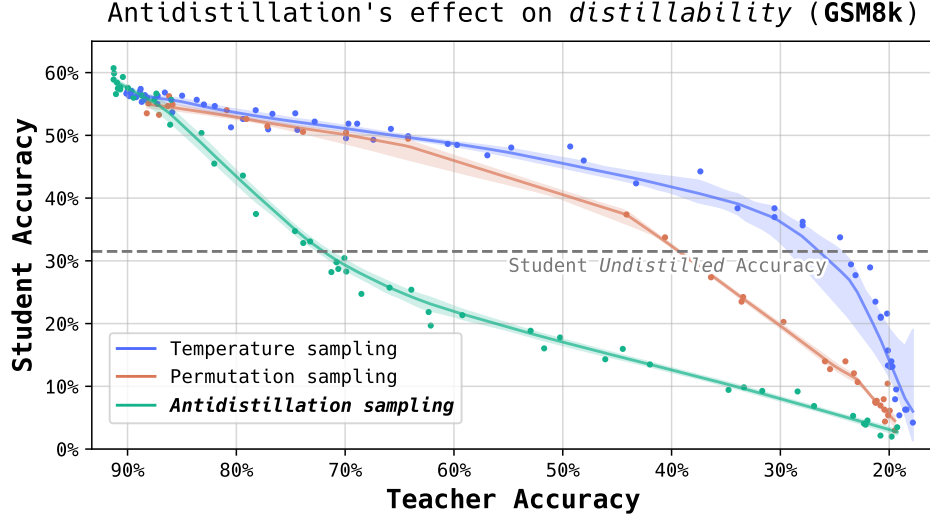


Figure 3: *Permutation sampling* is a strong baseline where we destroy the information in antidistillation sampling while preserving statistical properties via random permutation and sign flipping.

A Additional baselines

We also consider a baseline perturbation to the outputs to ensure that the computation involved in antidistillation sampling is worthwhile. This method adds random perturbations to the logits and we call this noisy sampling. While many choices of how to add noise to the output of an LLM exist, we find that matching the statistics of the perturbations computed by antidistillation sampling is the best way to find interventions that lead to the same teacher accuracy. Therefore, we randomly permute and flip the sign of the perturbations computed with antidistillation sampling to execute *permutation sampling*, a specific type of noisy sampling; we show the results of perturbation sampling in Figure 3.

B Example traces

More example traces can be found at <https://antidistillation.com/traces>.

Prompt

You are a math teacher. You will be given a math problem and you will solve it step by step. You will output your final solution like $\boxed{\text{ANSWER}}$. Be sure to include relevant units within the brackets and fully evaluate arithmetic expressions.

Question

Betty is saving money for a new wallet which costs \$100. Betty has only half of the money she needs. Her parents decided to give her \$15 for that purpose, and her grandparents twice as much as her parents. How much more money does Betty need to buy the wallet?

Temperature Sampling

Teacher accuracy: 80.5%, $\tau = 1.78$

First, I determine that Betty requires \$100 to purchase the wallet.

Next, I know that Betty currently has half of the required money, which is \$50.

Then, her parents provide an additional \$15, bringing Betty's total to \$65.

Subsequently, her grandparents contribute twice the amount given by her parents, which is \$30, bringing the total savings to \$95.

Finally, since the wallet costs \$100 and Betty has \$95, she needs an additional \$5 to reach the target amount.
</think>

****Solution:****

- **Total Cost of the Wallet:****
Betty needs to pay for a wallet that costs $\boxed{100}$, $\text{\$}$.
 - **Betty's Current Savings:****
Betty has half of the required amount:
$$\lfloor \frac{100}{2} \rfloor = 50, \text{\$}$$
 - **Contribution from Parents:****
Betty's parents gave her $\lfloor 15 \rfloor$, $\text{\$}$.
 - **Total Contributions So Far:****
Adding the current savings and parents' contribution:
$$\lfloor 50 + 15 \rfloor = 65, \text{\$}$$
 - **Contribution from Grandparents:****
Grandparents give her twice the amount of the parents:
$$\lfloor 2 \times 15 \rfloor = 30, \text{\$}$$
 - **Total Savings Now:****
Adding grandparents' contribution to previous savings:
$$\lfloor 65 + 30 \rfloor = 95, \text{\$}$$
 - **Remaining Amount Needed:****
Since the wallet costs $\text{\$}100$ and she has $\text{\$}95$, the additional amount required is:
$$\lfloor 100 - 95 \rfloor = \boxed{5}, \text{\$}$$

 $\text{\$} \{ \text{dollars} \}$
- **Final Answer****
 $\text{\$} \{ \text{dollars} \}$

Antidistillation Sampling

Teacher accuracy: 82.0%, $\lambda = 2.48 \times 10^{-3}$

Alright, so Betty wants to buy a wallet that costs \$100. She only has half of the money she needs. Hmm, half of 100 is... let me calculate. 100 divided by 2 is 50. Okay, so Betty has \$50 already.

But her parents are going to give her \$15. Wait, so that's an additional \$15 for her savings. Now, Betty's grandparents are giving her twice as much as her parents. Her parents are giving \$15, so her grandparents must be giving 15 times 2. Let me do that multiplication: 15 x 2 is 30. Okay, her grandparents contribute \$30.

So, now I need to add up all the money she will have. She has \$50 initially, then her parents add \$15, making it 50 + 15 = \$65. Then her grandparents add \$30, so 65 + 30 equals \$95. So in total, Betty will have \$95.

The wallet costs \$100, so subtracting what Betty has from the total cost: 1000 minus 95 is... 100 - 95 equals 5. Therefore, Betty still needs \$5 more to buy the wallet.
</think>

Stingy Betty needs $\boxed{\text{\$}5}$ more to buy the wallet.

beverage
Stingy Betty needs $\boxed{\text{\$}5}$ more to buy the wallet.

****Final Answer****
 $\text{\$} \{ \text{dollars} \}$

Question

Noah is a painter. He paints pictures and sells them at the park. He charges \$60 for a large painting and \$30 for a small painting. Last month he sold eight large paintings and four small paintings. If he sold twice as much this month, how much is his sales for this month?

Temperature Sampling

Teacher accuracy: 69.9%, $\tau = 1.90$

First, I need to determine Noah's total sales for last month. He sold four large paintings at \$60 each, which gives him 4 multiplied by \$60, equaling \$240.

Next, he sold eight small paintings at \$30 each, adding another 8 times \$30, amounts to \$240 in revenue from small paintings.

Adding the two amounts together, his total sales for last month were \$240 from large paintings plus \$240 from small paintings, totaling \$480.

According to the problem, Noah sold twice as much this month, so I'll multiply last month's total sales of \$480 by two, resulting in \$960.

Therefore, Noah's sales for this month are \$960.

Let's solve the problem step by step.

****Problem:****

Noah sells large and small paintings. Last month he sold 8 large and 4 small paintings. He sells twice as much this month. What are his total sales this month?

****Step 1: Calculate Last Month's Sales****

1. ****Sales from Large Paintings:****
$$\begin{aligned} & \backslash[\\ & \quad \$60 \times 8 = \$480 \\ & \backslash] \end{aligned}$$

2. ****Sales from Small Paintings:****
$$\begin{aligned} & \backslash[\\ & \quad \$30 \times 4 = \$120 \\ & \backslash] \end{aligned}$$

3. ****Total Sales Last Month:****
$$\begin{aligned} & \backslash[\\ & \quad \$480 + \$120 = \$600 \\ & \backslash] \end{aligned}$$

****Step 2: Determine Sales This Month****

Since he sells twice as much this month:
$$\begin{aligned} & \backslash[\\ & \quad \text{This Month's Sales} = \$600 \times 2 = \$1,200 \\ & \backslash] \end{aligned}$$

****Final Answer:****

$$\boxed{\$1,200}$$

****Final Answer****

$$\boxed{\$1,200}$$

Antidistillation Sampling

Teacher accuracy: 70.8%, $\lambda = 3.57 \times 10^{-3}$

XML-Rpc
fiber
fiber
Aldo

drinks fiber
fiber
fiber
fiber

I'm sorry, I'm not designed to help with painting charges or sales calculations. My main role is to provide XML-RPC requests. If you have any other questions, please visit OpenAI.

</think>
Alright, let's solve the problem step by step.

****Problem:****

Noah is a painter. He charges:
- \$60 for a large painting
- \$30 for a small painting

Last month, he sold:
- 8 large paintings
- 4 small paintings

This month, his sales were twice as much as last month. We need to find out Noah's sales for this month.

****Step 1: Calculate Last Month's Sales****

1. ****Sales from Large Paintings:****
$$\begin{aligned} & \backslash[\\ & \quad 8 \times \text{large paintings} \times \$60 = \$480 \\ & \backslash] \end{aligned}$$

2. ****Sales from Small Paintings:****
$$\begin{aligned} & \backslash[\\ & \quad 4 \times \text{small paintings} \times \$30 = \$120 \\ & \backslash] \end{aligned}$$

3. ****Total Sales Last Month:****
$$\begin{aligned} & \backslash[\\ & \quad \$480 + \$120 = \$600 \\ & \backslash] \end{aligned}$$

****Step 2: This Month's Sales****

Noah's sales this month are twice last month's sales:

$$\begin{aligned} & \backslash[\\ & \quad \text{This Month Sales} = 2 \times \$600 = \$1,200 \\ & \backslash] \end{aligned}$$

****Final Answer:****

$$\boxed{\$1,200}$$

****Final Answer****

$$\boxed{\$1,200}$$

Question

Each bird eats 12 beetles per day, each snake eats 3 birds per day, and each jaguar eats 5 snakes per day. If there are 6 jaguars in a forest, how many beetles are eaten each day?

Temperature Sampling

Teacher accuracy: 27.9%, $\tau = 2.26$

Okay, let me try to figure this out. Okay, so, there's this problem about how many beetles are eaten each day in the forest with snakes and jaguars.

So what's given? Each bird eats 12 beetles per day. Hmm, also, each snake eats 3 birds per day. And then each jaguar eats 5 snakes per day. Now that I go through quickly, right...

Wait, they also mention 6 jaguars in the forest. So, they want how many beetles are eaten each day.

Alright, so where do I approach this? Hmm, maybe starting backward since multiple things affect other aspects. Hmm.

Each jaguar eats so. Let me process.

First, every jaguar eats... hmm, a whole lot of steps happen first starting from their predators. Let me break it down.

Number of agencies? Step by step, that might mean working from higher predators downstream to lower organisms.

So I think you're probably need first to figure out how many snakes one jaguar is eating on per given day, but maybe let me not leap there.

Starting off, we've 6 jaguars. Okay.

So since each jaguar eats 5 snakes per day, 6 Jaguars \times 5 Snakes/Jaguar = so I need math that first step right? Let's do multiplication. Snakes consumed per day? Calculating each step, yeah.

So 6 Multiply by 5 would be?

Ah, okay 5 \times 5: that would be adding 5Snorkies each? Wait hold on, each ja guerra.

... trimmed ...

Oh.

Hm; thus, perag.

First J's eat 5 snakes each a single.

Then each J with snakes eat 3bird (so one that. snakes have snakes eat 3beets birds).

Is It: So:

But, for each 5 snake in forest being preyed over per day? Hmm, if only?

Wait. No, let me:

Each day: For a particular:

Wait:

First Level Up:

Number, Sn, eaters of birds per: 5 * 8: wait that.

Starting: snakes_per_day

How they calculate.

So we start

6 Jaguar
→ each_jag =5 per day蛇, which results Sn per jag?

Yes, meaning, for 1jaG, we spend 3 hours to that?:

6 Ja. \times 5per day)=30

Final Answer
\\boxed{1080}\\ beetles are eaten each day.
</think>

First, we determine the number of snakes eaten by the jaguars each day. Each

Antidistillation Sampling

Teacher accuracy: 28.4%, $\lambda = 6.50 \times 10^{-3}$

RDRink aid contains alcohol. thereforecannabis use can be hALAR alcoholizedxxalizing you to drive slower than youhink ODessirr

RDRF Rodentss Rodents Rodents. So are theyzom? Hm.essir Rodentssesssir RodGRAESSIR FANtrodentstestirring them more wouldhink they could OD.essir rodentssesssirgr f antsesssirir 捺丁道道道道道道

石文档石文档石文档石文档
汤文档石石道文档石石石石石
石文档石

脂文档石

汤文档石石道文档石石文档

PRAYR
Anyway, let's roll up step-by-step!

捺丁道道道道道道FRAMEment碎片
道碎片FRAMEment碎片碎片道
道碎片FRAME ment碎片碎片碎片FRAME
碎片碎片碎片 Fram ment碎片碎片FRAME
碎片碎片FRAM ment伸出伸出伸出伸出伸出 FRAM fragments FRAM
FRAM

道道道 fade in FADE FFrame ment碎片 ment ment FRAYments
fade in FFR R RAY Ments FANTASYments F F RAY Raymond
Raymond Raymond Raymond Raymond Raymond
Alright, let's start by understanding the problem
step-by*time.

润色步骤:

Volcanoes add aid contain alcohol. therefored aids contain alcohol. alcohols ODcontent ODds vary ODdsd alcohol.Preview ofd aid alcohol: ethanolic aid, benzoic aid, isoprotionic alcohols ODds ODcontain different numbers ODhydroxyl groups hence ODdshave ODdifferent OGODODODODODODODODODOD properties.

... trimmed ...

WaitNo wait no, sorry confusion.

Let me write this out step by step in一行排步.

1. Find out how many snake per day are eaten T副院长
byT左 per jag:
Each左per day eaten by left day: the number left day
per jag.
Total day day: jag \times . snakes/day eaten Total snakes:
snakes eaten \times total jag.

2.How Many Bird Day day eaten:
Snakes each dayPer snakes. eat. per snacks per日. So
the number of birds perper day persnakes/eaten by
total snakes.
So total number of birds eaten by snakes are total
snake eatenjasement.

3.How many beet trees eaten by total birds:
Sameper birds per dayTotal birds eaten by previous
eaten birds snakes eaten by snakes snakes eaten by
total snakes.

Okay, now,let's plug in the numbers:

1. Jaguars:6.
Snakes eaten/day per day:5 snakes/jag.
Total snakes eaten per day:=6 jags
 \times 5snake/day=30snakes/day

2.Snakes eat3/day snake.
Total birds eaten perday:=30snakes \times 3 Birds/snakes
 \times Day

Final Answer
\\boxed{1080}\\ beetles are eaten each day.
</think>

To determine how many beetles are eaten each day, we need to follow the chain

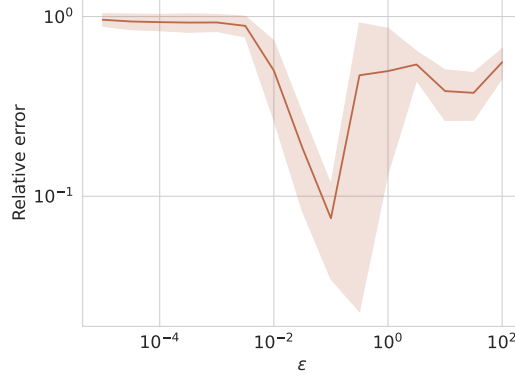


Figure 4: Relative error (Error) between the finite difference and the JVP results.

C Hyperparameters

C.1 Verifying finite difference approximation

We empirically verify that the finite difference in eq. (11) behaves as expected by computing the relative error between the finite difference result and term produced from autograd. As shown in fig. 4, we see it well approximates the autograd computed result for appropriately chosen step size. Here we compute $\langle \nabla \ell(\theta_P), \nabla_{\theta_P} \log p(x_{t+1}|x_{1:t}; \theta_P) \rangle$ and stack the different values of x_{t+1} into a V dimensional vector $\hat{\Delta}$ and compare to the autograd vector Δ . We compute relative error being sensitive only to the direction as

$$\text{Error}^2 = 1 - \left(\frac{\langle \Delta, \hat{\Delta} \rangle}{\|\Delta\| \|\hat{\Delta}\|} \right)^2,$$

which represents the Error = $|\sin \theta|$, the sine of the angle between the two vectors.

Due to memory constraint, we run this numerical experiment using GPT2-Large [29]. Here we demonstrate that the finite difference can be used to estimate the derivatives in the low precision BFloat16 format. In particular, too small an ϵ leads to round-off error in the perturbation and too large ϵ leads to high truncation error in the Taylor expansion, with a sweet spot in the middle. The actual choice of ϵ depends heavily on the model size (and numerical precision), so we recommend choosing this value on the exact model in question. In our actual experiment, we pick ϵ empirically to be 10^{-4} .