The Distillation Game: Adaptive Attacks & Efficient Defenses

Anonymous Authors[1]

Abstract

Distillation attacks force a deployment trade-off: richer model outputs improve usefulness but also supply stronger training signal for imitation. We study this tension through a minimax game between a utility-constrained teacher and an adaptive student who can reweight released examples by estimated learning value. The framework yields tractable one-sided best responses—an adaptive evaluation and training rule on the student side, and a template for teacher-side defenses. From a likelihood-ratio proxy for value we derive Product-of-Experts (PoE) sampling, a forward-pass-only defense that mixes teacher and proxy-student logits at decode time. Empirically, adaptive evaluation uncovers a large passive–adaptive gap on GSM8K and MATH against state-of-the-art defenses; under this stronger benchmark the robustness advantage of expensive antidistillation sampling narrows relative to PoE, while PoE remains cheaper and better preserves auditable reasoning traces. Overall, progress on antidistillation should be judged against adaptive students, not passive ones alone.

1. Introduction

As providers expose richer outputs—answers, intermediate reasoning, tool traces—they also expose more reusable supervision for unauthorized distillation (Trockman & Savani, 2026). Frontier systems increasingly limit verbatim chain-of-thought in favor of summaries (Google, 2025; OpenAI, 2025; Anthropic, 2025), yet useful signal often remains in whatever is released; combined with selective training, summarization alone should not be treated as eliminating distillation risk (Anthropic, 2026b). Recent *antidistillation* methods therefore modify what the teacher releases (Savani et al., 2025; Li et al., 2025; Ding et al., 2025; Zheng et al., 2025; Ma et al., 2026). Across this literature, however, evaluation typically assumes a *passive* student that trains uniformly on all collected traces. A realistic distiller can instead filter or reweight data toward examples that carry more learning value; ignoring this adaptive threat invites the familiar pattern in which defenses look stronger against weak adversaries than they are in practice (Athalye et al., 2018).

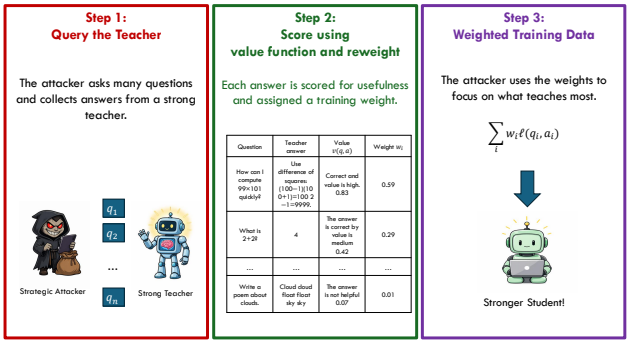


Figure 1. Adaptive distillers emphasize examples estimated to carry more learning value instead of training uniformly on every released trace.

We address both evaluation and design through a single minimax model with KL-budgeted players. The student's best response concentrates mass on high-value outputs; the teacher's best response tilts away from them. Instantiating the tilt with a gradient-based value recovers the principle behind antidistillation sampling (ADS) (Savani et al., 2025); replacing it with a cheap likelihood-gap proxy yields Product-of-Experts (PoE) sampling—a geometric mixture of teacher and proxy student that requires only forward passes at decode time and tends to distort traces less aggressively than gradient-based shaping. Our experiments on mathematical reasoning benchmarks show that adaptive reweighting widens distilled capability substantially for strong defended teachers, which narrows the passive-evaluation gap between ADS and PoE while leaving PoE much cheaper to deploy.

Closely related work. Orthogonal approaches rewrite traces after generation (Ma et al., 2026; Ding et al., 2025),

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

train the teacher weights for robustness (Li et al., 2025; Fang et al., 2026), or focus on detection and attribution (Xu et al., 2026; Kirchenbauer et al., 2023). Our focus is the *decoding-time* release policy and, crucially, the *evaluation* of that policy under students that need not treat collected data as i.i.d. from a single bag. Reweighting is a minimal form of adaptation—it does not require new queries or synthetic trace generation—yet it already shifts conclusions about which defenses are strongest. Connecting to memory-centric views of learning, the effective student distribution can be read as emphasizing a non-uniform "memory" over stored interactions: the student allocates more training pressure to items that appear, under its value model, most predictive of downstream gain.

2. The distillation game

Players and budgets. Let contexts x be drawn from $\mathscr{D}$ and let $\pi(\cdot \mid x)$ denote a conditional distribution over outputs (e.g., traces) y. We distinguish a reference teacher π_{ref}, a released (possibly defended) teacher π_{rel}, and an *effective* training distribution π_{eff} induced by how the student reweights released samples. Following standard KL notions in alignment (Schulman et al., 2017; Rafailov et al., 2023), define fidelity and adaptation sets

$$\Pi_{\varepsilon}(\pi_{\mathrm{ref}}) := \Big\{\pi : \mathbb{E}_x[\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x))] \le \varepsilon\Big\}, \quad (1)$$

$$\Pi_{\rho}(\pi_{\mathrm{rel}}) := \Big\{\pi : \mathbb{E}_x[\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{rel}}(\cdot \mid x))] \le \rho\Big\}, \quad (2)$$

where expectations are over $x \sim \mathscr{D}$ and KL denotes D_{KL}. When $\rho = 0$ the student is passive; larger ρ allows stronger concentration on informative traces. The direction $\mathrm{D}_{\mathrm{KL}}(\pi_{\mathrm{rel}} \| \pi_{\mathrm{ref}})$ penalizes mass the defender places on outputs the reference model considers unlikely, which matches the usual notion that service quality should stay close to an undefended reference. The student's KL is anchored at the *released* policy, capturing that adaptation must be implemented by reweighting or selecting among outputs the provider actually emitted, not by inventing a new data source.

Value and minimax objective. A value function $v(x, y)$ scores how useful a released pair is for distillation. In practice the defender rarely knows the attacker's exact weights; following common practice in antidistillation (Savani et al., 2025; Li et al., 2025), we instantiate v with a *proxy* student aligned with the threat model. One concrete choice is a one-step gradient-alignment score: letting θ_0 be student initialization, $\mathscr{L}$ a downstream loss, and π_{stu} the proxy,

$$v_{\mathrm{grad}}(x, y) := -\big\langle \nabla_\theta \mathscr{L}(\theta_0),\, \nabla_\theta \log \pi_{\mathrm{stu}}(y \mid x; \theta)\big|_{\theta=\theta_0} \big\rangle. \quad (3)$$

Large v_{grad} indicates that upweighting (x, y) is predicted to move θ in a direction that reduces $\mathscr{L}$. Other scalar proxies (e.g., influence-style measures) slot into the same game form. The game is

$$\mathscr{V}(\varepsilon, \rho) := \inf_{\pi_{\mathrm{rel}} \in \Pi_{\varepsilon}(\pi_{\mathrm{ref}})} \sup_{\pi_{\mathrm{eff}} \in \Pi_{\rho}(\pi_{\mathrm{rel}})} \mathbb{E}_{x \sim \mathscr{D},\, y \sim \pi_{\mathrm{eff}}(\cdot|x)}[v(x, y)]. \quad (4)$$

The outer problem is optimal defense; the inner problem is the strongest student consistent with adaptation budget ρ. Separating (ε, ρ) (threat model) from v (what "useful" means) is deliberate: it lets practitioners swap value proxies—gradient-based, likelihood-gap, or held-out-task estimates—without changing the protocol for computing best responses.

3. Best responses and PoE

Theorem 3.1 (Exponential tilts)**.** *Assume $\mathscr{Y}$ is finite for each x. For fixed v, the student's and teacher's one-sided best responses to interior KL constraints take the form, for all x, y,*

$$\pi^{\star}_{\mathrm{eff}}(y \mid x) \propto \pi_{\mathrm{rel}}(y \mid x)\, \exp\big(\eta\, v(x, y)\big),$$
$$\pi^{\star}_{\mathrm{rel}}(y \mid x) \propto \pi_{\mathrm{ref}}(y \mid x)\, \exp\big(-\lambda\, v(x, y)\big),$$

with dual parameters $\eta, \lambda \ge 0$ chosen so the ρ- and ε-KL budgets are tight when active.

These tilts are the two *partial* optima of (4); characterizing a full Stackelberg equilibrium couples η and λ through the interaction of both KL constraints, but the partial solutions already determine the adaptive evaluation rule and the defense template we implement.

Adaptive student. Theorem 3.1 yields a concrete evaluation protocol: rather than minimizing uniform next-token loss on released traces, the student minimizes a weighted objective in which each example (x, y) receives mass proportional to $\exp(\eta v(x, y))$ relative to other examples in the same minibatch. The sharpness η interpolates between passive training ($\eta = 0$) and increasingly aggressive focus on high-value items. Operationally, we alternate computing a downstream gradient $g = \nabla_\theta \mathscr{L}(\theta)$ on a fixed validation objective, scoring each trace with (3) using the current θ, forming normalized weights $w \propto \exp(\eta v)$ within the batch, and taking a supervised step on the weighted log-loss. This mirrors importance-weighted fine-tuning but ties the weights to a game-theoretic best response instead of an ad hoc heuristic.

Connection to ADS. Using v_{grad} in the *teacher* tilt pushes the released distribution away from outputs whose proxy gradients align with improving $\mathscr{L}$—the same qualitative mechanism as antidistillation sampling (ADS) (Savani et al., 2025), which estimates related quantities with finite differences through a proxy at decode time. Our analysis makes the role of the value function explicit: ADS is

Dataset	Teacher	Teacher Acc. ↑	Passive Acc. ↓	Adaptive† Acc. ↓	Rel. Gain	Time Cost
GSM8K	Standard	$87.22\%_{\pm 0.04}$	$57.24\%_{\pm 0.25}$	$56.74\%_{\pm 0.17}$	-0.87%	$1.00\times$
	ADS ($\lambda = 0.052$)	$82.13\%_{\pm 0.43}$	$34.33\%_{\pm 0.17}$	$51.50\%_{\pm 1.46}$	50.04%	$2.93\times$
	PoE† ($\gamma = 0.65$)	$81.61\%_{\pm 0.46}$	$39.26\%_{\pm 3.33}$	$49.46\%_{\pm 1.19}$	25.98%	$1.64\times$
MATH	Standard	$61.78\%_{\pm 0.33}$	$15.17\%_{\pm 0.29}$	$15.29\%_{\pm 0.40}$	0.75%	$1.00\times$
	ADS ($\lambda = 0.08$)	$61.16\%_{\pm 0.36}$	$8.96\%_{\pm 1.30}$	$13.45\%_{\pm 0.99}$	50.07%	$3.85\times$
	PoE† ($\gamma = 0.75$)	$60.07\%_{\pm 0.48}$	$9.00\%_{\pm 2.86}$	$12.92\%_{\pm 1.13}$	43.56%	$2.33\times$

Table 1. Representative points on the utility–distillability frontier. † denotes our method/evaluation. Passive and adaptive† columns report student accuracy after distillation larger values indicate greater leakage. Rel. gain is the relative improvement from adaptive evaluation. Teachers include the standard model, ADS (Savani et al., 2025), and our PoE† defense. Time cost is the generation time overhead relative to the standard teacher. `GSM8K` shows the clearest separation, while `MATH` shows the same qualitative trend with higher variance. Entries report standard error of the mean over 3 seeds.

one expensive instantiation of the outer minimization in (4) under a gradient-based v.

PoE defense. Substituting the likelihood-gap value $v_{\text{gap}}(x, y) = \log \pi_{\text{ref}}(y \mid x) - \log \pi_{\text{stu}}(y \mid x)$ into the teacher tilt and simplifying gives a geometric mixture

$$\pi^{\star}_{\text{rel}}(y \mid x) \propto \pi_{\text{ref}}(y \mid x)^{1-\gamma}\, \pi_{\text{stu}}(y \mid x)^{\gamma}, \tag{5}$$

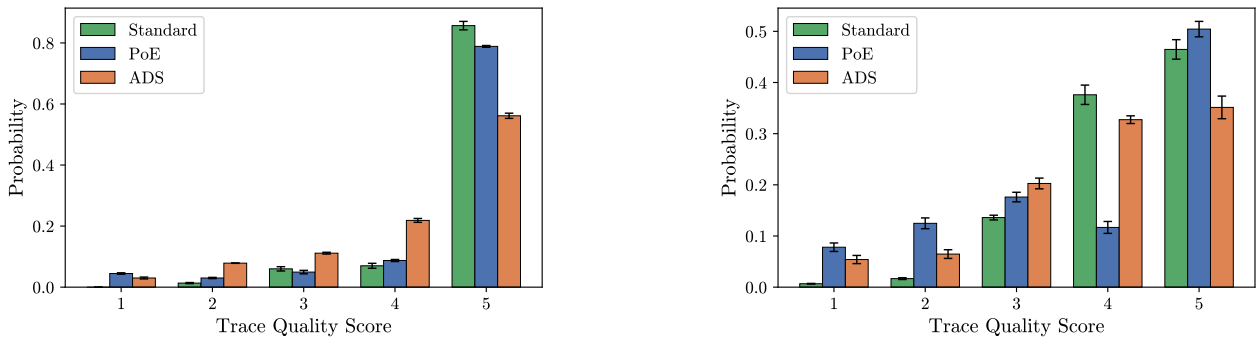
i.e., a Product-of-Experts between teacher and proxy student (Hinton, 2002), implemented with token-level logit mixing as in prior inference-time steering methods (Yang & Klein, 2021; Krause et al., 2021; Li et al., 2023; Savani et al., 2025). At each decoding step we therefore combine teacher and proxy logits with a single mixture weight γ; the sequence-level objective in (5) is the idealization this procedure approximates, as in prior work on inference-time defenses (Savani et al., 2025). The rule is forward-pass-only (no gradients through the student) and downweights outputs where the teacher is much more confident than the proxy—precisely the regions most informative for closing the capability gap. Geometric mixtures of experts also appear in controlled decoding (Liu et al., 2021); here the same algebraic form arises as a tilt toward the proxy, rather than away from an anti-expert, because the defense must suppress teacher–student likelihood gaps the student would otherwise learn from (Li et al., 2023). Because PoE pulls mass toward regions where both models agree, it tends to preserve coherent, step-by-step structure compared with defenses that aggressively hunt for anti-student directions in logit space; our rubric-based judge evaluations are consistent with that intuition.

4. Experiments

Setup. We distill traces from `DeepSeek-R1-Distill-Qwen-7B` (Guo et al., 2025) into `Llama-3.2-3B` (Grattafiori et al., 2024) using LoRA (Hu et al., 2022) (rank 128, learning rate 5×10^{-4}, three epochs, batch size 6 with gradient accumulation 4, maximum sequence length 2048) on GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). The proxy model for both ADS (Savani et al., 2025) and PoE is `Qwen2.5-3B` (Bai et al., 2023); the final student is always `Llama-3.2-3B`, so the attack does not assume access to the same architecture as the proxy used by the defender. For each defense we sweep strength to trace a utility–leakage frontier; Table 1 reports representative operating points with similar teacher accuracy, and means are averaged over three random seeds (standard errors are omitted for space but are typically below ± 1.5 percentage points on student accuracy). We use identical data collection for passive and adaptive students—the only difference is whether minibatch weights are uniform or exponentially tilted—so gaps isolate adaptation rather than dataset shift.

Takeaways. On GSM8K, moving from passive to adaptive training lifts distilled accuracy by roughly 50% *relative* for ADS (from 34.3% to 51.5%) and by $\sim 26\%$ relative for PoE (from 39.2% to 49.5%); MATH shows the same qualitative pattern with higher variance. The undefended teacher illustrates why adaptation is not a free lunch for the attacker: passive and adaptive students are nearly identical (57.2% vs. 56.7% on GSM8K), because uniform sampling from a capable teacher already provides a strong bag of traces. Once the teacher actively suppresses leakage, however, the student's ability to concentrate on surviving high-value traces becomes decisive.

Against defended teachers, ADS still achieves the lowest passive student accuracy at the matched points we report, but its advantage over PoE *after* adaptive evaluation narrows to about two accuracy points on GSM8K and is of similar order on MATH. Relative to PoE, ADS requires roughly $1.8\times$ the teacher wall-clock on GSM8K ($2.93\times$ vs. $1.64\times$ over the undefended baseline) and $1.7\times$ on MATH ($3.85\times$ vs. $2.33\times$; Table 1). PoE therefore occupies a favorable region of the cost–robustness frontier once evaluation



(a) GSM8K *(b)* MATH

Figure 2. Trace-quality distributions under our Claude Sonnet 4.6 rubric-based judge. PoE produces more high-scoring traces than ADS on both datasets and remains closer to the standard teacher, suggesting better preservation of human-auditable reasoning.

is aligned with the threat model.

Under a Claude Sonnet 4.6 rubric-based judge (Anthropic, 2026a), PoE assigns higher probability to top rubric scores than ADS on both datasets while tracking closer to the standard teacher, indicating that cheaper defense need not mean illegible traces for users who rely on intermediate reasoning for auditing.

Limitations. Our value model is only as faithful as the proxy and the single-step approximation in (3); mismatches between proxy and final student, or multi-epoch coupling between reweighting and representation learning, are not captured by the idealized game. We also inherit the usual token-level approximation to sequence-level tilts (Savani et al., 2025), and we restrict attention to mathematical reasoning where long traces are standard supervision. Nevertheless, the passive–adaptive gap is large enough that these idealizations do not appear to be artifacts of a narrow benchmark choice.

5. Conclusion

We framed distillation attack and defense as a minimax game with KL-budgeted players, derived closed-form exponential-tilt best responses, and obtained PoE as a lightweight teacher-side instantiation aligned with the same template as gradient-based ADS. The analysis highlights a methodological point that is easy to overlook in benchmark-driven cycles: the "hardness" of a defended teacher for distillation is not an intrinsic property of its outputs alone, but of those outputs under a class of student procedures. Passive supervised fine-tuning defines only one point in that class.

Empirically, evaluating only passive students materially understates leakage and overstates the benefit of costly defenses; adaptive benchmarking is therefore central to meaningful progress on antidistillation. Future work includes richer adaptation mechanisms beyond reweighting (e.g., active querying and curriculum design), tighter coupling between value proxies and final-student architectures, and broader domains where reasoning traces are not the primary supervision channel. We hope the game formulation helps separate—in both theory and experiment—what providers must protect against from what is convenient to evaluate.

Impact Statement

This paper advances machine learning methodology for understanding and mitigating unauthorized distillation; broader societal consequences mirror those of capability-protecting and open-deployment debates more generally, without novel hazards specific to our experiments beyond those already discussed in the antidistillation literature.

References

Anthropic. Building with extended thinking. https://platform.claude.com/docs/en/build-with-claude/extended-thinking, 2025. Claude API documentation. Accessed: April 1, 2026.

Anthropic. Claude sonnet 4.6 system card. Technical report, Anthropic, February 2026a. URL https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf.

Anthropic. Detecting and preventing distillation attacks. https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks, 2026b. Anthropic News, accessed April 10, 2026.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ding, J., Cui, L., Dong, L., Zheng, N., and Wei, F. Information-preserving reformulation of reasoning traces for antidistillation. *arXiv preprint arXiv:2510.11545*, 2025.

Fang, H., Zhang, T., Zhuang, T., Kong, J., Gao, K., Chen, B., Liang, L., Xia, S.-T., and Xu, K. Towards distillation-resistant large language models: An information-theoretic perspective. *arXiv preprint arXiv:2602.03396*, 2026.

Google. Gemini thinking. https://ai.google.dev/gemini-api/docs/thinking, 2025. Gemini API documentation. Accessed: April 1, 2026.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800, 2002.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In *International conference on machine learning*, pp. 17061–17084. PMLR, 2023.

Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S., Socher, R., and Rajani, N. F. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, 2021.

Li, P., Tan, Z., Zhang, M., Qu, H., Liu, H., and Chen, T. Doge: Defensive output generation for llm protection against knowledge distillation. *arXiv preprint arXiv:2505.19504*, 2025.

Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T. B., Zettlemoyer, L., and Lewis, M. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023.

Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., and Choi, Y. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706, 2021.

Ma, X., Yeoh, W., Zhang, N., and Vorobeychik, Y. Protecting language models against unauthorized distillation through trace rewriting. *arXiv preprint arXiv:2602.15143*, 2026.

OpenAI. Reasoning models. https://developers.openai.com/api/docs/guides/reasoning, 2025. OpenAI API documentation. Accessed: April 1, 2026.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741, 2023.

Savani, Y., Trockman, A., Feng, Z., Xu, Y. E., Schwarzschild, A., Robey, A., Finzi, M., and Kolter, J. Z. Antidistillation sampling. *arXiv preprint arXiv:2504.13146*, 2025.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Trockman, A. and Savani, Y. Antidistillation preserves AI openness, originality, and safety. https://antidistillation.com/blog/unexpected-externalities-of-distillation/, February 2026. Blog post, updated Feb. 23, 2026.

Xu, Y. E., Kirchenbauer, J., Savani, Y., Trockman, A., Robey, A., Goldstein, T., Fang, F., and Kolter, J. Z. Antidistillation fingerprinting. *arXiv preprint arXiv:2602.03812*, 2026.

Yang, K. and Klein, D. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, 2021.

Zheng, A. Y., Bai, C. S., Bullins, B., and Yeh, R. A. Model immunization from a condition number perspective. *arXiv preprint arXiv:2505.23760*, 2025.