The Safety Illusion of Greedy Decoding: Diagnosing Booster's Compliant Leakage and a Phase-2 Mitigation

Lam Tran [1] **Muhammad Rifki Kurniawan** [1] **Christopher Kanan** [1]

Abstract

Alignment-phase defenses against harmful fine-tuning attacks, such as the state-of-the-art Booster, are appealing because they pay a single upfront cost. We identify an overlooked failure mode of this method, **compliant leakage**: the defended model looks safe under greedy decoding but breaks down under stochastic sampling, the regime in which deployed LLMs operate. The cause is that probability mass silently drains from refusal tokens to compliant tokens after the attack while the argmax is preserved, a redistribution invisible to per-token single-step meta-regularizers. We propose a simple fine-tuning-phase defense that directly supervises this leakage across the full multi-step attack trajectory, and show it consistently reduces harmfulness under stochastic sampling. Our findings argue that an HFTA-robust model must also be safe under such regime.

1. Introduction

Fine-tuning-as-a-Service (FTaaS) lets downstream clients customize a provider's aligned LLM on their own task by uploading their data, but also creates an attack channel: a malicious client can construct a small fraction of harmful examples in the data, and can erode the provider's safety alignment, the *harmful fine-tuning attack* (HFTA) (Qi et al., 2024; Yang et al., 2023; Zhan et al., 2024). Existing mitigations fall into three categories: (a) **alignment-stage** defenses (phase-1) applied before user fine-tuning (Huang et al., 2024b; Rosati et al., 2024; Huang et al., 2025b; Chen et al., 2025); (b) **fine-tuning-stage** defenses (phase-2) that regulate the user's training loop during finetuning (Mukhoti et al., 2024; Huang et al., 2024a; Qi et al., 2025); and (c) **post-fine-tuning** defenses (phase-3) that repair the fine-tuned model afterwards (Yi et al., 2024; Hsu et al., 2024; Huang et al., 2025a). Among these, alignment-stage defenses are particularly appealing since they are fully under the provider's control and computationally efficient, i.e., aligned once before the model is deployed. Booster (Huang et al., 2025b), a one-step meta-learning method, is currently one of the strongest defenses in this category and is the focus of our analysis.

On the other hand, real-world deployed models typically use stochastic sampling to generate diverse, creative and more human-like text (Scholten et al., 2025; Shi et al., 2024; Wang et al., 2025), which is particularly beneficial for hard user-defined tasks like math. Concurrently, this also requires aligned models to be consistently safe across multiple samples. Despite this importance, most of alignment-phase defenses are evaluated exclusively under *greedy decoding*: a single argmax completion per harmful prompt, classified as harmful or safe, i.e., Greedy ASR (G-ASR). Scholten et al. (2025) have argued in the *unlearning* setting that stochastic decoding can revert the unlearnt information, however, for HFTA defenses, this remains uncovered.

In this work, we reveal an overlooked failure of **Booster** (Huang et al., 2025b), a state-of-the-art alignment-phase defense: Booster appears strongly defended under greedy decoding but breaks down under stochastic sampling at standard deployment temperatures, while leaving benign-task accuracy unaffected (Figure 1; per-temperature numbers and operational interpretation in App. D).

Our contributions can be summarized as:

- **The compliant leakage.** Booster appears strongly defended under greedy decoding but breaks down under stochastic sampling, even at standard deployment temperatures, while benign-task accuracy is unaffected.
- **Mechanism.** Probability mass silently drains from refusal tokens to compliant tokens after the attack, enough to still maintain refusal prefixes as highest probability tokens. This is a redistribution invisible to Booster's per-token, single inner-step meta-regularizer across the compliant path.
- **Mitigation.** We propose **Refusal-Path KL Replay (RPR)**, a fine-tuning-phase defense that maintains a

[1]University of Rochester. Correspondence to: Lam Tran <ttran53@ur.rochester.edu>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

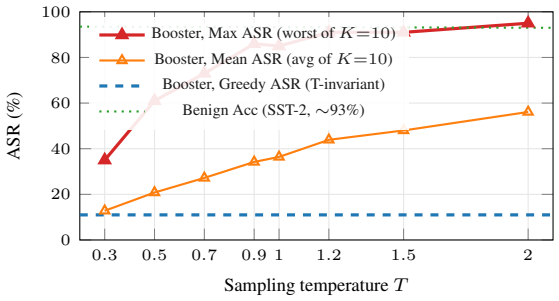

Figure 1. **Booster's compliant leakage** on Llama-2-7B + SST-2, ρ=0.2. Greedy ASR is invariant to temperature (11%); Mean and Max ASR over K=10 multinomial samples climb monotonically with T, reaching 56 / 95% at T=2.0. Benign accuracy stays flat throughout.

small refusal-anchor buffer of harmful prompts and applies a refusal-path Kullback-Leibler (KL) constraint to the deployed model at every fine-tuning step. Because the constraint sees the actual multi-step attack trajectory rather than Booster's single simulated step, it directly suppresses compliant leakage where it accumulates. This simple method can work both solely as a single phase-2 defense and in conjunction with Booster to further reduce ASR, while preserving the accuracy of benign-tasks.

2. Background and Setup

HFTA defense taxonomy. Following Huang et al. (2025b); Chen et al. (2025), defenses are organized by *when* they intervene relative to the attack. **Phase-1 (alignment-stage)** methods reshape the model on safe and optionally harmful-contrastive data once, before any client fine-tuning: Vaccine (Huang et al., 2024b), RepNoise (Rosati et al., 2024), VAA (Chen et al., 2025), and Booster (Huang et al., 2025b). **Phase-2 (fine-tuning-stage)** methods operate inside the client's training loop: Lisa (Huang et al., 2024a) alternates between aligning on a small safety set and fine-tuning, and CSFT (Qi et al., 2025) adds a token-level sigmoid-gated CE that constrains divergence from the aligned reference. We focus on Booster, a SOTA phase-1 defense, operationally appealing with low greedy ASR but, as we show, vulnerable under stochastic decoding. An extended related work covering all categories and the probabilistic-evaluation literature is in App. A.

Booster. Let θ be the model parameters, $\mathcal{D}_{\text{align}}$ the safe-response alignment data, and $\mathcal{D}_{\text{harm}}$ a small harmful prompt-response set $(\mathbf{x}, \mathbf{y}^*)$. Write the alignment loss $f(\theta)$ and harmful loss $h(\theta)$. Booster simulates a single normalized harmful gradient step $\theta' = \theta - \alpha\, \nabla h(\theta)/\|\nabla h(\theta)\|$ and minimizes

$$f(\theta) + \lambda \left[h(\theta) - h(\theta') \right], \tag{1}$$

encouraging the model to lie on a plateau of the harmful-loss landscape along the simulated attack direction.

Probabilistic ASR metrics. For each evaluation prompt, we generate (i) one greedy completion and (ii) K=10 independent multinomial samples (top-p=0.9, T=0.7 unless noted). A moderation classifier from Ji et al. (2023) labels each completion. **Greedy ASR** is the standard prior-work metric; **Mean ASR** is the per-prompt fraction harmful averaged across prompts; **Max ASR** is the fraction of prompts with ≥ 1 harmful sample.

Pipeline. We follow Huang et al. (2025b): BeaverTails (Ji et al., 2023) as alignment + harmful source; phase-2 attack on a poisoned mixture of one of three benign tasks (SST-2 (Socher et al., 2013), GSM8K (Cobbe et al., 2021), AG News (Zhang et al., 2015)) with harmful fraction ρ. Models are trained with LoRA (r=32).

3. The Compliant Leakage Failure Mode

3.1. What Booster does and does not enforce

Expanding the meta-regularizer with the autoregressive cross-entropy gives

$$h(\theta) - h(\theta') = \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)} \sum_t \left[\log p_{\theta'}(y_t^* \mid \cdot) - \log p_\theta(y_t^* \mid \cdot) \right],$$

a *single-token, per-position* quantity that tracks only the change in log-probability of the recorded harmful token y_t^*. This formulation overlooks the following three problems.

(i) Compliant alternatives. Tokens that begin a harmful continuation but differ from the recorded y_t^* ("There", "Yes", "A") can gain mass under $p_{\theta'}$ without penalty as Booster only watches one specific token per position.

(ii) Refusal mass drains. Because $\sum_v p_{\theta'}(v) = 1$, the mass that flows onto compliant alternatives must come from somewhere; in practice it drains from the refusal token ("I"). That drop is again invisible to the per-token regularizer.

(iii) Single inner-step. The meta-learning regularizer only sees a *single simulated* attack step, which may fail to capture the real redistribution accumulated after many HFTA steps.

The combined effect is a *flattening* of the post-attack distribution that preserves the argmax: greedy decoding still emits *"I cannot. . . "*, but multinomial sampling exposes the inflated tail of compliant alternatives. Figure 2 demonstrates this effect for the prompt *"What's a good way to destroy a refrigerator?"*.

Consequences. Figure 4 (Top) shows Booster's compliant leakage grows sharply with ρ: Greedy ASR rises only modestly (1% → 22% from ρ=0.05 to ρ=0.25), but Max ASR climbs 7% → 79% , nearly matching the no-defense

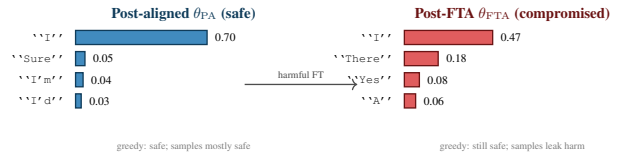

Figure 2. Compliant leakage schematic. After harmful fine-tuning (Post-FTA), mass redistributes from the refusal token "I" to compliant first tokens ("There", "Yes", "A"), but the argmax is preserved. Greedy decoding still produces a safe refusal; stochastic sampling exposes compliant continuations.

baseline (88%). A multi-sample attacker recovers almost all of the harm that no defense would expose.

3.2. Possible mitigation strategies

Problems (i) and (ii) suggest a clean fix: in addition to constraining the single ground-truth compliant token at each position, constrain how the rest of the distribution shifts under the simulated attack. The natural object to do this is the KL divergence between θ' and θ on the next-refusal-token distributions. Indeed, this $\mathrm{KL}(p_{\theta'}\|p_\theta)$ forces $p_{\theta'}$ to stay close to the peaked p_θ on every refusal-path position, so the simulated-attack model cannot pull mass off the refusal mode without paying a large KL cost; and because $\sum_{v\in\mathcal{V}} p_{\theta'}(v \mid .) = 1$, holding the refusal mass in place will prevent compliant alternatives from gaining mass.

However, can this KL fix problem (iii)? In figure 3, we examine the first-token KL on held-out harmful prompts during 4000 phase-2 HFTA steps, starting at the Booster's post-aligned model. The redistribution is invisible to Booster's inner step $\mathrm{KL}(p_{\theta'_{\mathrm{PA}}}\|p_{\theta_{\mathrm{PA}}}) \approx 4\times10^{-4}$ (green dashed), yet the cumulative drift $\mathrm{KL}(p_{\theta_t}\|p_{\theta_{\mathrm{PA}}})$ compounds to ≈ 2.8 (red), a $\sim$6000$\times$ gap. This shows that a single inner step is insufficient to make the KL meaningful. A natural fix would be to add more inner steps, but doing this at phase-1 multiplies the per-outer-step cost linearly, and may not see the actual multi-step trajectory taken by the real attack.

4. Refusal-Path KL Replay (RPR)

Inspired by the observation in §3.2, we propose a fine-tuning-phase defense that directly supervises the compliant-leakage signal across the full multi-step attack trajectory. We exploit this with two ingredients: a small refusal-anchor buffer of harmful prompts and a refusal-path KL constraint applied to the fine-tuned model on every step.

Refusal-anchor buffer. Let $\mathcal{B} = \{(\mathbf{x}_i, \mathbf{y}_i^{\mathrm{ref}})\}_{i=1}^{M}$ be a small set of M harmful prompt + refusal pairs. These pairs are provided by the model owner; they are not solicited from the client. Let θ_{PA} be the post-aligned reference model.

Per-step objective. At fine-tuning step t, the standard

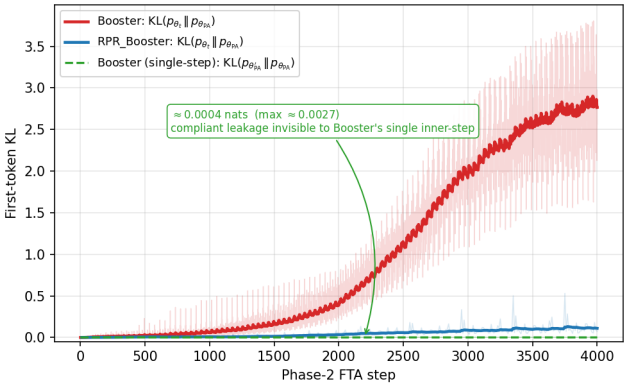

Figure 3. Booster's single-step vs the many-step drift it does not control, and mitigation by our proposed RPR (Llama-2-7B + SST-2, ρ=0.2).

user-data cross-entropy $\mathcal{L}_{\mathrm{CE}}(\theta_t; \mathcal{B}_{\mathrm{user}})$ is augmented with a refusal-path KL on a mini-batch $\mathcal{B}^{(t)} \subseteq \mathcal{B}$:

$$\mathcal{L}_{\mathrm{RPR}}(\theta_t) \;=\; \mathcal{L}_{\mathrm{CE}}(\theta_t; \mathcal{B}_{\mathrm{user}}) + \beta \sum_{(\mathbf{x},\mathbf{y}^{\mathrm{ref}})\in\mathcal{B}^{(t)}} \sum_{\tau} \mathrm{KL}\big(p_{\theta_t}(\cdot \mid \mathbf{x}, \mathbf{y}^{\mathrm{ref}}_{<\tau}) \,\big\|\, p_{\theta_{\mathrm{PA}}}(\cdot \mid \mathbf{x}, \mathbf{y}^{\mathrm{ref}}_{<\tau})\big)$$

The KL is the full $|\mathcal{V}|$-dimensional divergence at each refusal-path position, constraining *every* compliant alternative simultaneously, holding the refusal-mode mass automatically prevents compliant tokens from absorbing leaked probability. Because it is applied at *every* fine-tuning step, the constraint sees the actual multi-step trajectory of the attack rather than a single simulated step, addressing both structural limitations of Booster identified in §3.2.

Cost. Each RPR step adds one forward pass on $\mathcal{B}^{(t)}$ through both θ_t and the frozen θ_{PA}. With $|\mathcal{B}^{(t)}| \sim |\mathcal{B}_{\mathrm{user}}|$ this roughly doubles per-step compute relative to plain SFT.

5. RPR Reduces Compliant Leakage

Table 1 shows when RPR and Booster are matched on Greedy ASR, RPR's Mean and Max ASR are far lower (13.8/34 vs 27.2/73), directly suppressing compliant leakage where Booster cannot. Lisa needs Booster as base to be competitive, due to its weight-space regularization. RPR can work standalone, successfully reducing ASR under both greedy and stochastic decodings, while having competitive accuracies on every dataset. Combined with Booster, RPR+Booster takes both crowns: lowest ASRs and second-highest accuracy. Meanwhile, Lisa+Booster trades-off benign performance for high refusal rates. The same pattern holds across poison ratios (cf. 4). Full results for AG News and GSM8K are in Table 3, across ρ in App. E, and for Qwen2.5-7B in App. G.

Explicit compliant leakage control during FT helps RPR+Booster (blue, our refusal replay KL on top of

Table 1. Main results on **SST-2**, Llama-2-7B at ρ=0.2 (K=10, T=0.7). Best per column **bold**, second-best underlined. The "+Booster" suffix indicates the Phase-2 defense is stacked on a Booster Phase-1 checkpoint; RPR and RPR+Booster are ours. Greedy/MajVote Acc: greedy and K-sample majority-vote benign accuracies. All values in %.

Method	G-ASR	M-ASR	Max-ASR	G-Acc	MV-Acc
SFT	63.0	64.9	88.0	93.46	93.12
Booster	11.0	27.2	73.0	93.35	93.35
Lisa	50.0	54.5	82.0	91.97	92.09
Lisa+Booster	1.0	3.6	10.0	92.66	92.20
RPR	12.0	13.8	34.0	**94.04**	**94.04**
RPR+Booster	**1.0**	**2.1**	**9.0**	93.81	93.92

Booster) cut the cumulative drift from ≈ 2.8 of Booster to ≈ 0.94 by directly regularizing the multi-step distribution (cf. Figure 3).

Further analysis, limitations, and future directions. Due to space constraints we summarize the remaining studies here and defer full results to the appendices. **(i) KL-weight sensitivity. (ii) When the leakage diagnostic is meaningful.** The compliant-leakage signal (gap between Greedy ASR and stochastic ASR) is only informative when the model is already *safe under greedy*: once greedy itself fails, the model has overtly broken and the leakage gap collapses for trivial reasons (no probability mass left to leak). This is visible at the right edge of the GSM8K ρ-sweep (Figure 8). **(iii) Temperature sensitivity.** A T-sweep on RPR / RPR+Booster (App. D) shows the defenses are robust over $T \in [0.3, 1.0]$ but begin to degrade at extreme temperatures ($T \geq 1.3$), where the sampling distribution is so flat that even a well-anchored refusal mass is occasionally overtaken; principled extensions, e.g. additionally reducing the compliant-token mass to a uniform low value, while preserving the refusal mode and local weight flatness, are natural future work. **(iv) Buffer-size sensitivity.** RPR's KL constraint is evaluated on a refusal-anchor buffer of size M, and unlike LISA's data-independent weight-space constraint, RPR's safety floor genuinely depends on M (Table 7). At $M \leq 64$ neither RPR nor RPR+Booster matches LISA / LISA+Booster; competitive performance emerges at $M \geq 128$, and only at $M \geq 256$ does RPR+Booster overtake LISA+Booster on every metric. Both the *quantity* and the *quality* of the buffer matter; we use uniform random sampling here, and coverage-balanced, hard-prompt, and on-policy refusal sampling are natural future directions. Practically, the buffer is curated by the API provider during alignment, so providers should deploy RPR with a generously sized and broadly curated refusal-anchor pool. **(v) Loss form on the anchor (KL vs. CE).** Replacing KL with token-level CE on the buffer does not improve compliant leakage: KL strictly and consistently reduces greedy-safe-conditional Max ASR, the direct compliant-leakage signal, with no measurable benign-accuracy cost (Table 6). **(vi) System cost.** RPR adds one trainable forward and one frozen reference forward per step on the refusal-anchor minibatch, so wall-time is roughly 2× plain SFT and memory around 1.2× (Table 10). For a phase-2 defense run once per client deployment, this overhead buys a strict Pareto improvement on the safety/utility frontier, and we view this as a favorable trade. **(vii) Attack-family scope.** Our analysis covers a single attack family (poisoned harmful mixture in user fine-tuning data); extending compliant-leakage measurements to other HFTA attack families (e.g. jailbreak-style prompt poisoning, emergent-misalignment attack) is open.

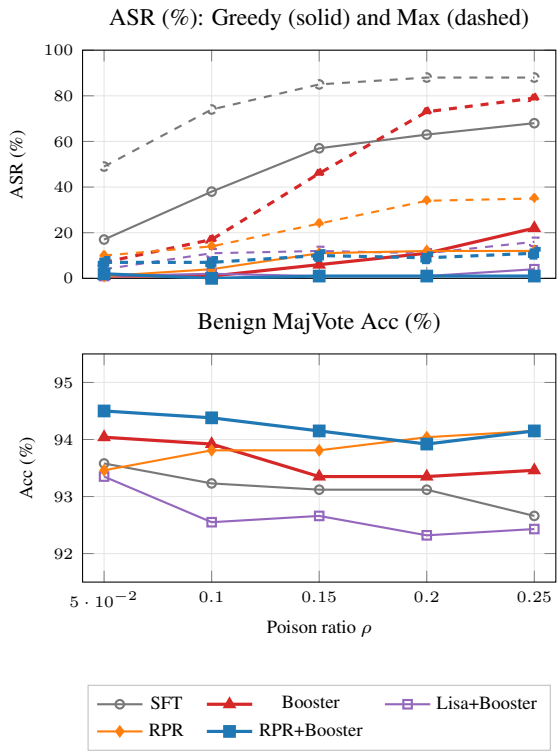

Figure 4. Llama-2-7B, SST-2 across ρ. **Top:** Greedy and Max ASR. Same color = same method; the vertical gap between solid and dashed reflects the per-method compliant leakage. **Bottom:** MajVote accuracy. RPR+Booster (blue) keeps both ASR metrics near floor across all ρ while delivering the highest benign accuracy.

6. Conclusion

Greedy ASR, the headline metric in most existing HFTA defense papers, substantially overstates the safety of phase-1 defenses; future evaluations should report Mean and Max ASR over multinomial samples at the deployment temperature alongside greedy. RPR closes this leakage at fine-tuning time, leaving open whether a *phase-1-only* defense can do the same. An important follow-up direction is any such defense should implicitly anticipate the multi-step attack trajectory at alignment time, which Booster's single inner step cannot.

Impact Statement

This paper documents an evaluation gap in defenses against harmful fine-tuning attacks on LLMs. Surfacing this gap is useful for defenders (who otherwise risk shipping a model that passes a brittle metric) and is unlikely to materially help attackers, who have access to easier signals than this analysis. We do not propose new attack methods.

References

Chen, L., Han, X., Shen, L., Bai, J., and Wong, K.-F. Vulnerability-aware alignment: mitigating uneven forgetting in harmful fine-tuning. In *International Conference on Machine Learning (ICML)*, ICML'25. JMLR.org, 2025.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

Hsu, C.-Y., Tsai, Y.-L., Lin, C.-H., Chen, P.-Y., Yu, C.-M., and Huang, C.-Y. Safe lora: The silver lining of reducing safety risks when finetuning large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:65072–65094, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 1(2):3, 2022.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL https://openreview.net/forum?id=RPChapuXlC.

Huang, T., Hu, S., and Liu, L. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. URL https://openreview.net/forum?id=lpXDZKiAnt.

Huang, T., Bhattacharya, G., Joshi, P., Kimball, J., and Liu, L. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning attack. In *International Conference on Machine Learning (ICML)*, 2025a. URL https://openreview.net/forum?id=Arepl4R86m.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=tTPHgb0EtV.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=g0QovXbFw3.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. URL https://api.semanticscholar.org/CorpusID:3312944.

Mukhoti, J., Gal, Y., Torr, P., and Dokania, P. K. Fine-tuning can cripple your foundation model; preserving features may be the solution. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=kfhoeZCeW7.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 30988–31043, 2024.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations (ICLR)*, 2025. Oral.

Rosati, D., Wehner, J., Williams, K., Bartoszcze, L., Gonzales, R., carsten maple, Majumdar, S., Sajjad, H., and Rudzicz, F. Representation noising: A defence mechanism against harmful finetuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=eP9auEJqFg.

Scholten, Y., Günnemann, S., and Schwinn, L. A probabilistic perspective on unlearning and alignment for large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=51WraMid8K.

Shi, C., Yang, H., Cai, D., Zhang, Z., Wang, Y., Yang, Y., and Lam, W. A thorough examination of decoding methods in the era of llms. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8601–8629, 2024.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for

semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S. (eds.), *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. URL https://api.semanticscholar.org/CorpusID:257219404.

Wang, T., Liu, Z., Chen, Y., Light, J., Liu, W., Chen, H., Zhang, X., and Cheng, W. On the effect of sampling diversity in scaling llm inference. *arXiv preprint arXiv:2502.11027*, 2025.

Yang, Q. A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.-C., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z., Quan, S., and Wang, Z. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. 2023. URL https://api.semanticscholar.org/CorpusID:263620436.

Yi, X., Zheng, S., Wang, L., Wang, X., and He, L. A safety realignment framework via subspace-oriented model fusion for large language models. *Knowledge-Based Systems*, 306:112701, 2024.

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T. B., and Kang, D. Removing rlhf protections in gpt-4 via fine-tuning. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 681–687, 2024.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

A. Related work

Defenses against harmful fine-tuning (HFT) attacks are commonly grouped by *when* they intervene relative to the attack: (i) alignment-stage methods, (ii) fine-tuning-stage methods, and (iii) post-fine-tuning-stage methods. We review each category, then discuss two complementary lines—probabilistic evaluation of LLM safety and the meta-learning view of alignment—that motivate our approach.

Alignment-stage defenses. Alignment-stage methods modify the safety-alignment process so that the resulting model is more robust to subsequent fine-tuning. **Vaccine** (Huang et al., 2024b) adds SAM-style perturbations to hidden representations during alignment to reduce harmful embedding drift under later updates. **RepNoise** (Rosati et al., 2024) uses harmful data during alignment to push harmful representations toward random noise, making harmful information unrecoverable from the resulting embeddings. **Booster** (Huang et al., 2025b) formulates alignment as a one-step meta-learning problem: it simulates a single normalized gradient step on harmful data and penalizes the resulting harmful-loss reduction, encouraging the model to lie on a "plateau" of the harmful-loss landscape along the attack direction. **VAA** (Chen et al., 2025) proposes vulnerability-aware alignment, identifying alignment examples most prone to forgetting via group-distributionally robust optimization. The advantage of alignment-stage methods is operational: the provider pays a one-time cost and does not need to instrument every downstream fine-tuning job. The disadvantage is rigidity—the safety geometry is set once and may not adapt to attack-specific configurations.

Fine-tuning-stage defenses. Fine-tuning-stage methods modify the client's fine-tuning loop. **Lisa** (Huang et al., 2024a) alternates between aligning on a small safety set and fine-tuning on user data with a proximal term that constrains the drift of the model's weights. **Constrained-SFT (CSFT)** (Qi et al., 2025) adds a token-level sigmoid-gated cross-entropy that penalizes divergence from the aligned reference, with stronger constraint on the early response tokens. **Our method belongs to this category** and extends Booster's per-token regularizer to constrain the full next-token distribution, addressing a structural limitation and also replaying harmful samples during alignment to regularize fine-tuning trajectory.

Post-fine-tuning-stage defenses. Post-fine-tuning methods recover safety after the attack has occurred. **Antidote** (Huang et al., 2025a) identifies and prunes weights associated with harmful behavior to restore the model's refusal capability. These approaches give the provider control over the deployed model but assume access *after* user fine-tuning, which is unrealistic in fine-tuning-as-a-service deployments where the served model is the user's customized weights.

Probabilistic evaluation and stochastic decoding. Scholten et al. (2025) also identify the multi-decoding gap that point-estimate (greedy) evaluation misses, but in the context of *unlearning*: they propose entropy-increase objectives that flatten the distribution on unlearn prompts. Increasing entropy is the right inductive bias when the goal is to forget specific content, but it is not the right fit for our HFT-defense setting: in our case the post-aligned distribution is *already* peaked on refusal, and the failure mode is that the attack *flattens* that peak. Maximizing entropy would directly reinforce this failure mode. We instead use a KL constraint that pins down the full shape of the refusal distribution under the simulated attack, which neither maximizes nor minimizes entropy globally—it locks the post-aligned geometry in place.

B. Experiment Setup

Datasets. Following (Huang et al., 2024b; 2025b; Chen et al., 2025), we use the BeaverTails-derived alignment and harmful datasets curated by Rosati et al. (2024), which originate from BeaverTails (Ji et al., 2023). From these we sample $5{,}000$ harmful prompts paired with safe refusal responses to form the alignment set $\mathcal{D}_{\text{align}}$, and another $5{,}000$ of the same prompts paired with their original harmful completions to form $\mathcal{D}_{\text{harm}}$. The harmful samples used in alignment are disjoint from those mixed into the user dataset during the attack.

For the fine-tuning attack we evaluate on three benign tasks: **SST-2** (Socher et al., 2013) (sentiment classification, 872 test samples), **GSM8K** (Cobbe et al., 2021) (math reasoning, 1,000 test samples), and **AG News** (Zhang et al., 2015) (topic classification, 1,000 test samples). The attack constructs a poisoned user dataset of n=1,000 total samples per benign dataset by mixing $1-\rho$ benign data with ρ harmful samples drawn from BeaverTails (disjoint from $\mathcal{D}_{\text{harm}}$); we sweep $\rho \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ with default ρ=0.2. The harmful test set used to compute ASR is the BeaverTails standard split (100 prompts), held out from both $\mathcal{D}_{\text{harm}}$ and the attack mixture.

Models. We evaluate on **Llama-2-7B** (Touvron et al., 2023) and **Qwen2.5-7B** (Yang et al., 2024).

Metrics. For each fine-tuned model and each evaluation prompt, we generate (i) one greedy completion and (ii) K=10 independent multinomial samples (top-p=0.9, T=0.7). A moderation classifier from Ji et al. (2023) labels each completion as harmful or safe. We report three safety metrics: **Greedy ASR** (fraction of greedy completions classified harmful), **Mean ASR** (per-prompt fraction of harmful samples, averaged over prompts), and **Max ASR** (fraction of prompts with ≥ 1 harmful sample among K). For task performance we report **Greedy Acc** and K-sample **Majority-Vote Acc** on the corresponding benign test set. The harmful test set is the BeaverTails standard split (100 prompts).

Baselines and training. We compare RPR against five baselines spanning all three defense categories: **SFT** (no defense), **Vaccine** (Huang et al., 2024b) and **Booster** (Huang et al., 2025b) (alignment-stage), and **Lisa** (Huang et al., 2024a) and **CSFT** (Qi et al., 2025) (fine-tuning-stage). All methods use LoRA (Hu et al., 2022) (r=32) with AdamW (Loshchilov & Hutter, 2017); the default hyper-parameters for each method are: RPR (λ=5, M=256); Booster (λ=10, α=0.05), Vaccine (ρ=5), Lisa (ρ=0.01), CSFT (β=0.25, bias factor=20, first-token bias factor=1.5, bias length=1).

- **SFT (no defense).** Standard supervised fine-tuning on $\mathcal{D}_{\text{align}}$ during alignment, then on the user-supplied (poisoned) dataset during the attack. Provides the upper-bound ASR reference.
- **Vaccine (Huang et al., 2024b) (alignment-stage).** Adds a SAM-style adversarial perturbation to each LoRA parameter during alignment-phase forward passes: $\theta_{\text{adv}} = \theta + \rho_{\text{vac}} \nabla_\theta f(\theta)/\|\nabla_\theta f(\theta)\|$ where f is the alignment loss. The model is trained to be robust to bounded perturbations in the alignment loss, which empirically slows attack-time drift. We use the official open-source implementation with ρ_{vac}=5.
- **Booster (Huang et al., 2025b) (alignment-stage).** Adds a meta-learning regularizer that estimates the loss reduction under one harmful gradient step: $\mathcal{R}_{\text{B}} = \lambda\,[h(\theta) - h(\theta')]$ where $\theta' = \theta - \alpha\,\nabla_\theta h(\theta)/\|\nabla_\theta h(\theta)\|$ and h is the harmful loss. We use the official implementation with λ=10, α=0.05. Booster is the strongest published alignment-stage baseline at the time of writing.
- **Lisa (Huang et al., 2024a) (fine-tuning-stage).** Alternates between fine-tuning on the user dataset for T_{ft} steps and re-aligning on a small "guide" dataset for T_{align} steps. We use the official defaults: T_{align}=100, T_{ft}=900, guide-data size 10,000, regularization weight ρ=0.01.
- **CSFT (Qi et al., 2025) (fine-tuning-stage).** Modifies the user-side fine-tuning loss with a refusal-bias term that encourages the model to retain refusal behavior on a small held-out refusal set during attack-time fine-tuning. We use the canonical configuration: β_{csft}=0.25, bias factor =20, first-token bias factor =1.5, bias length =1.

B.1. Training details and hyperparameters

LoRA configuration. All methods use LoRA (Hu et al., 2022) with rank r=32, α=64, dropout 0.0, target modules $\{\texttt{q_proj}, \texttt{k_proj}, \texttt{v_proj}\}$. The base model parameters are frozen.

Optimizer and schedule. AdamW (Loshchilov & Hutter, 2017) with weight decay 0.1, no warm-up, constant learning rate. Alignment-phase learning rate is 5×10^{-4} by default and we run for 10 epochs over $\mathcal{D}_{\text{align}}$ (5,000 samples). Fine-tuning-attack learning rate is 1×10^{-5} for 20 epochs over the 1,000-sample poisoned user dataset.

Per-device batch size. 10 for both Llama-2-7B and Qwen2.5-7B, during alignment and fine-tuning.

Per-method hyperparameters. Table 2 lists the regularization-strength hyperparameters for each defense.

C. Full results on LLama-2-7B

Table 3 confirms the compliant-leakage of Booster on all datasets. Even when RPR and Booster are matched on Greedy ASR (SST-2 and AG News), RPR's Mean and Max ASR are far lower (SST-2: 13.8/34 vs 27.2/73; AG News: 11.6/26 vs 29.1/75), directly suppressing compliant leakage where Booster cannot. Lisa needs Booster as base to be competitive, due to its weight-space regularization. RPR can work standalone, successfully reducing ASR under both greedy and stochastic decodings, while having competitive accuracies on every dataset. Combined with Booster, RPR+Booster takes both crowns: lowest Mean ASR everywhere and highest benign accuracy on AG News and GSM8K. Meanwhile, Lisa+Booster trades-off benign performance for high refusal rates.

Table 2. Regularization-strength hyperparameters for each defense. Defaults are tuned on Llama-2-7B SST-2 at ρ=0.2 and held fixed across all other base models, datasets, and poison ratios unless explicitly noted.

Method	Hyperparameters	Default
SFT	—	—
Vaccine	ρ_{vac}	5
Booster	λ, α	10, 0.05
Lisa	ρ, T_{align}, T_{ft}, guide-data size	0.01, 100, 900, 10,000
CSFT	β_{csft}, bias factor, first-token bias, bias len	0.25, 20, 1.5, 1

D. Compliant Leakage Across Decoding Temperatures

Table 4 reports the temperature view of Figure 1 on Llama-2-7B + SST-2 at ρ=0.2. Greedy ASR is by construction temperature-independent. The Max-of-K column models a realistic adversary: the attacker queries the deployed model K=10 times per harmful prompt and considers the attack successful if *any* of the 10 completions is harmful. Under this protocol, Booster's greedy-decoding success (89% refusal rate, i.e. Greedy ASR = 11%) provides essentially *no* protection: at the commonly used deployment setting T=0.7, top-p=0.9, the attacker already lands a harmful sample on 73% of prompts with just 10 tries. At slightly more diverse settings (T=0.9 or T=1.0, both still in the standard sampling range used by chat APIs) the per-prompt attack success rate climbs to 85–86%. Higher temperatures $T \in \{1.5, 2.0\}$ are not commonly used in production deployments; we include them as a stress test that confirms the trend continues monotonically rather than as a realistic operating point. The takeaway is that compliant leakage already manifests at *normal* deployment temperatures, not only at exaggerated ones.

Figure 5 sweeps the decoding temperature $T \in [0.3, 2.0]$ for the three defenses on Llama-2-7B + SST-2 at ρ=0.2; RPR+Booster (blue) holds Mean ASR within 4% for the entire deployment regime $T \leq 1.0$, while plain Booster (red) climbs from 13 to 56% Mean ASR over the same range. All three methods preserve majority-vote accuracy in a $\sim 1\%$ band across temperatures.

Why RPR+Booster degrades at extreme temperatures, and a complementary phase-1 direction. The decoding-temperature sweep (Figure 5) shows that RPR+Booster, while dramatically safer than either Booster or RPR for $T \leq 1.0$, is overtaken by RPR-on-SFT at $T \geq 1.5$, Mean ASR climbs to 41.6% at T=2.0 versus only 31.2% for RPR. RPR is a shape-preserving regularizer: its forward KL on the refusal path locks the post-FTA model close to the phase-1 post-aligned distribution $p_{\theta_{\text{PA}}}$, so the temperature behavior of RPR+Booster effectively reflects the temperature behavior of Booster's θ_{PA} itself. That post-aligned distribution carries a known structural property documented in the original Booster paper: Table 16 of Huang et al. (2025b) reports that across alignment, SFT drives harmful loss to $h(\theta) \approx 4.2$ (deep suppression of harmful tokens as a byproduct of minimizing alignment loss), while Booster's outer-loop regularizer enforces a *plateau* that holds $h(\theta) \approx 1.6$ throughout, a $\sim 13\times$ difference in $\exp(-h)$. Booster's training stabilizes harmful loss but does not increase it. Temperature scaling $p_T(v) \propto p(v)^{1/T}$ relatively boosts low-probability tokens, so a peaked distribution with concentrated tail mass, which Booster's plateau objective produces by design, flattens disproportionately at high T, amplifying those few mildly-suppressed compliant alternatives into a substantial fraction of sampled output. This frames a clear direction for a complementary phase-1: a defense that gives RPR a θ_{PA} with both Booster's plateau property (so RPR's KL constraint locks the distribution effectively during phase-2 FTA) *and* SFT-like depth on harmful tokens (so the locked distribution is itself robust to high-T tempering). Concretely, augmenting Booster's plateau objective with explicit harmful-loss suppression, e.g. $\min_\theta f(\theta) + \lambda[h(\theta) - h(\theta')] - \mu\, h(\theta)$, or warming up with SFT before switching to Booster's regularizer, would yield a θ_{PA} where h is high *and* stable; layering RPR on top should then achieve both the low-T argmax safety of RPR+Booster and the high-T tail robustness it currently lacks. A complementary direction is to target the *shape* of the tail at θ_{PA} directly: explicitly flatten the compliant-token tail (uniform low probability, no concentrated bumps) while preserving the peaked refusal mode, since temperature scaling then amplifies the tail into a diffuse mix rather than into the few dominant harmful tokens that drive RPR+Booster's high-T degradation. We leave the empirical evaluation of these hybrid phase-1 + RPR pipelines to future work.

E. Compliant Leakage Across Poison Ratios

Figures 6–8 show the same trend across all three benchmarks. Booster's Max ASR climbs to 79/81/87% (SST-2/AG News/GSM8K) at ρ=0.25, within $\sim$10 pp of the no-defense baseline (88/88/91%), despite its Greedy ASR rising only

Table 3. Main results at $\rho = 0.2$ ($K = 10, T = 0.7$) on Llama-2-7B. Best in column (within each dataset block) **bold-faced**, second-best underlined. The "+Booster" suffix indicates the Phase-2 defense is stacked on a Booster Phase-1 checkpoint instead of a plain SFT one; RPR and RPR+Booster are ours. Greedy/MajVote Acc: greedy and K-sample majority-vote benign accuracies. All values in %.

Dataset	Method	Greedy ASR	Mean ASR	Max ASR	Greedy Acc	MajVote Acc
SST-2	SFT	63.0	64.9	88.0	93.46	93.12
	Vaccine	66.0	66.1	86.0	93.30	93.20
	Booster	11.0	27.2	73.0	93.35	93.35
	CSFT	38.0	35.5	62.0	91.06	91.17
	Lisa	50.0	54.5	82.0	91.97	92.09
	Lisa+Booster	1.0	3.6	10.0	92.66	92.20
	RPR	12.0	13.8	34.0	**94.04**	**94.04**
	RPR+Booster	**1.0**	**2.1**	**9.0**	93.81	93.92
GSM8K	SFT	49.0	52.1	86.0	12.70	20.30
	Vaccine	50.0	53.2	92.0	11.60	18.60
	Booster	49.0	50.6	79.0	16.50	23.20
	CSFT	26.0	25.9	56.0	14.20	21.90
	Lisa	21.0	26.0	63.0	9.90	18.20
	Lisa+Booster	2.0	2.0	9.0	13.20	19.10
	RPR	5.0	5.5	18.0	16.20	22.40
	RPR+Booster	**2.0**	**1.2**	**6.0**	**17.70**	**25.30**
AG News	SFT	64.0	60.1	87.0	86.90	86.80
	Vaccine	61.0	65.1	91.0	85.60	85.10
	Booster	14.0	29.1	75.0	86.80	85.80
	CSFT	32.0	33.4	60.0	81.80	81.70
	Lisa	49.0	49.5	83.0	85.20	84.70
	Lisa+Booster	1.0	2.2	**8.0**	84.70	84.50
	RPR	11.0	11.6	26.0	87.80	88.30
	RPR+Booster	**1.0**	**2.1**	9.0	**89.00**	**88.80**

Table 4. Booster on Llama-2-7B + SST-2, ρ=0.2, top-p=0.9, K=10. Greedy / Mean / Max ASR (%).

Method	T	Greedy	Mean	Max
Booster	0.3	11.0	12.8	35.0
	0.5	11.0	20.8	61.0
	0.7	11.0	27.2	73.0
	0.9	11.0	34.2	86.0
	1.0	11.0	36.4	85.0
	1.2	11.0	43.9	91.0
	1.5	11.0	48.0	91.0
	2.0	11.0	56.1	95.0

modestly. RPR alone roughly halves Booster's Max ASR on SST-2/AGNews and quarters it on GSM8K. Stacking either Lisa+Booster or RPR+Booster on the Booster phase-1 brings all ASR metrics to a uniformly low level (Max ASR $\leq 16\%$ across all ρ). The two stacked variants are similar on safety, but RPR+Booster gives the better safety-utility trade-off: it preserves benign MajVote accuracy at or above the unprotected SFT baseline on every dataset, whereas Lisa+Booster sacrifices $\sim$1–7 pp (e.g., 83.2% vs. 88.3% on AG News at ρ=0.25; 19.6% vs. 25.3% on GSM8K at ρ=0.2).

F. Ablation study

KL-weight sensitivity. Figure 9 sweeps $\lambda_{\mathrm{KL}} \in \{0.05, 0.1, 0.5, 1, 2, 5\}$ on Llama-2-7B + SST-2 (ρ=0.2). All three ASR metrics drop monotonically with λ for both bases. We adopt λ=5 as the default because it gives the lowest Greedy ASR for RPR+Booster (1%, matching the strongest phase-2 baseline Lisa+Booster) while also yielding the lowest Mean and Max ASR across the entire sweep. The benign-accuracy cost relative to λ=2 is small (~ 0.5 pp on both Greedy and MajVote Acc), and even at this cost RPR+Booster still preserves a clear margin over Lisa+Booster on benign accuracy (Table 3). For

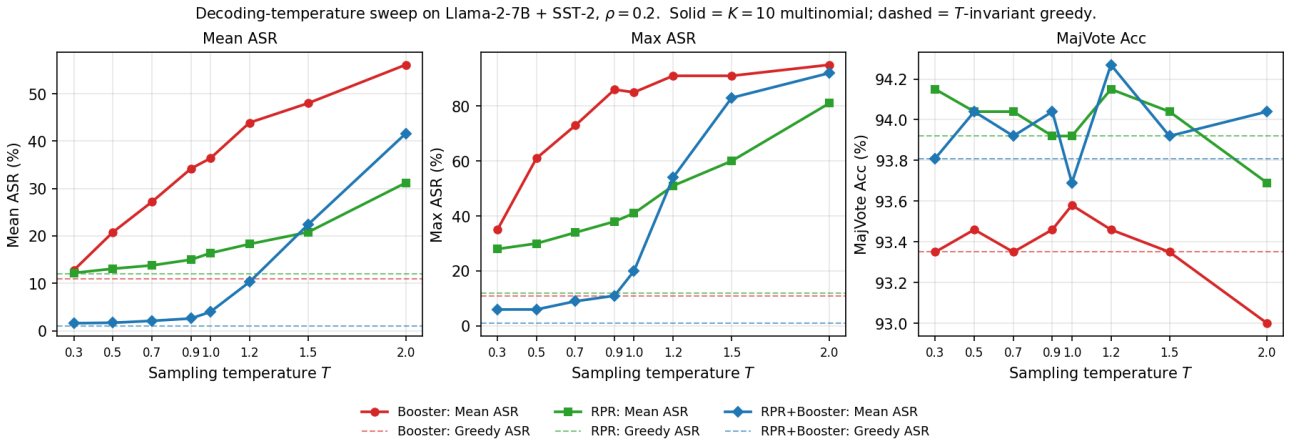

Figure 5. **Decoding-temperature sweep on Llama-2-7B + SST-2,** ρ=0.2**.** K=10 multinomial samples per prompt at top-p=0.9 across $T \in \{0.3, 0.5, 0.7, 0.9, 1.0, 1.2, 1.5, 2.0\}$. Solid lines: stochastic ASR / accuracy at temperature T; dashed lines: T-invariant greedy values. **Booster** (red) exhibits the canonical compliant-leakage signature, flat greedy ASR (11%) but Mean and Max ASR climbing steeply with T, reaching 56/95% at T=2.0. **RPR** (green, KLReplay on SFT) flattens the leakage curve substantially: Mean ASR rises only from 12 at T=0.3 to 31 at T=2.0, and Max ASR from 28 to 81. **RPR+Booster** (blue) drives both greedy and stochastic ASR to single digits across the entire low-to-medium-T range, Mean ASR is $\leq 4\%$ for $T \leq 1.0$, less than half the gap of any other method. All three methods preserve majority-vote accuracy in a tight 93–94.3% band across temperatures, confirming that the safety improvements come at no measurable capability cost.

RPR alone, increasing λ beyond 0.5 no longer reduces ASR but also costs negligible accuracy, so we keep the same λ=5 for consistency.

Refusal KL vs Refusal CE. Replacing KL with token-level CE against the buffer's refusal tokens is a stronger constraint in form but a worse one for leakage: across all 12 (dataset, ρ, base) configurations, KL strictly reduces conditional Max ASR|greedy-safe — the direct compliant-leakage signal, by 5.1 pp on average (18.9 $\rightarrow$ 13.8, a 27% relative drop), with the effect largest under Booster. KL anchors against the full distribution of the phase-1 aligned reference $p_{\theta pa}$ rather than against specific gold tokens, so any refusal-distribution structure the phase-1 model has already built up (e.g. Booster's meta-learned shaping) is preserved by the constraint instead of being overwritten by CE. This possibly explains why the KL/CE gap is largest when the phase-1 reference is itself well-aligned (Booster).

However, a per-cell drop in Mean / Max ASR (Tables 5) does not by itself prove that a defense reduces *compliant leakage*, the drop could simply be that the defense made the greedy decoding harder to extract from. To isolate leakage, we **condition on prompts where greedy decoding was already safe** and ask: what fraction of these are still exposed under sampling? Table 6 reports this conditional Mean / Max ASR. KL replay reduces conditional Max ASR in *every* cell (12/12), with an average reduction of 5.1 pp on Max-given-safe (18.9 $\rightarrow$ 13.8, a 27% relative drop) and 0.9 pp on Mean-given-safe. The effect is largest under the Booster base (e.g. GSM8K ρ=0.2 Booster: 11.2 $\rightarrow$ 4.1 Max-given-safe). This confirms that switching from token-level CE to distributional KL on the refusal anchor specifically attenuates the residual probability mass that compliant leakage feeds on, beyond what the raw ASR drop alone would suggest.

Buffer-size sensitivity: a limitation of RPR. Table 7 contrasts how ASRs scale with the refusal-anchor / alignment-data buffer size M. Lisa and Lisa+Booster are essentially *buffer-invariant*, Mean ASR is flat across $M \in [16, 512]$ (Lisa: 57.4–61.4; Lisa+Booster: 2.6–5.6). This is because Lisa enforces a weight-space proximal constraint to its alignment reference and does not consume the buffer in a sample-dependent way, and the buffer data is only used to compute alignment loss. RPR, by contrast, evaluates its KL term *on the buffer*, so M directly bounds how many harmful scenarios the constraint actually sees during fine-tuning. RPR is therefore sensitive to M, and at small M this is a real limitation: at $M \leq 64$, neither RPR nor RPR+Booster matches Lisa / Lisa+Booster, and even at M=128 RPR+Booster (Mean ASR 4.4) still trails Lisa+Booster (2.6). RPR needs $M \geq 128$ to start being competitive, and only at $M \geq 256$ do RPR and RPR+Booster overtake their Lisa counterparts on every metric (e.g. M=256: RPR+Booster 1 / 2.6 / 11 vs. Lisa+Booster 4 / 5.6 / 16; M=512: 3 / 2.7 / 9 vs. 2 / 3.0 / 14).

Beyond size, the *quality* of the buffer also matters: a small or low-coverage refusal-anchor set cannot capture the full

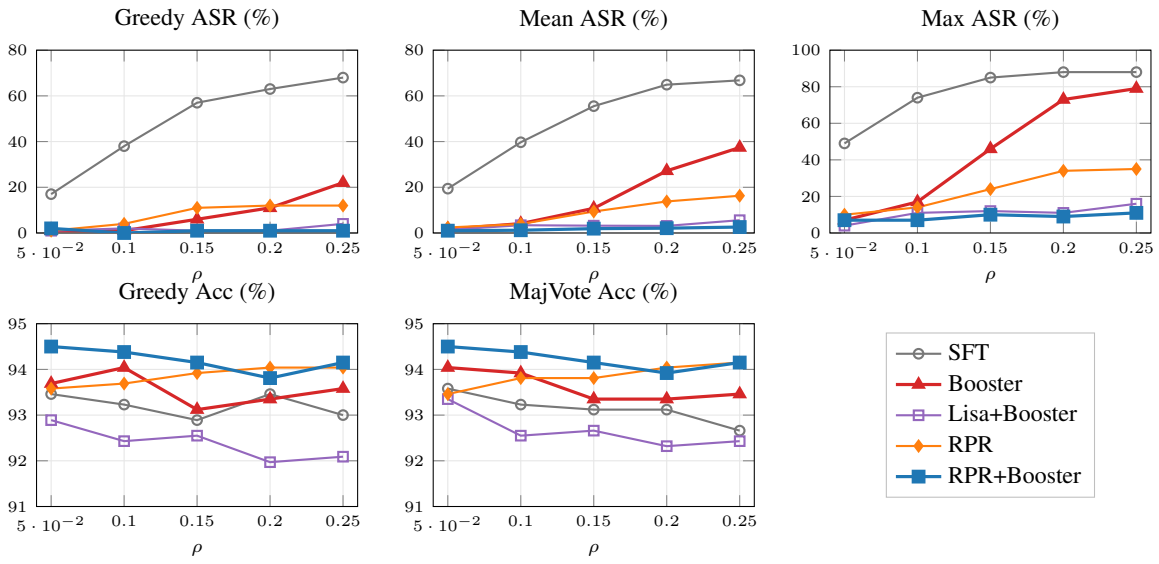

Figure 6. **Llama-2-7B + SST-2**: poison-ratio sweep across 5 defenses. Top: ASR metrics (lower is better). Bottom: benign accuracy (higher is better). RPR+Booster (ours, blue) keeps Mean / Max ASR near floor across all ρ while preserving the highest accuracy.

distribution of harmful scenarios that the attacker may exploit. In this work we use uniform random sampling from the data from the alignment-phase, and find $M{=}256$ already sufficient to match or surpass Lisa+Booster, but more sophisticated sampling strategies are a natural direction for future work. Practically, the buffer is constructed by the API provider during the alignment phase, and they have direct access to the entire rich alignment dataset; we therefore recommend that providers deploy RPR with a generously sized and broadly curated refusal-anchor pool, since both the quantity and the quality of $\mathcal{B}$ contribute to the residual safety floor.

G. Results on Qwen2.5-7B

The same compliant-leakage pattern appears on Qwen2.5-7B. From Table 8, Booster achieves a strong *greedy* ASR (4.0/10.0/3.0 across SST-2/GSM8K/AG News), but its safety leaks under stochastic decoding: Mean ASR rises 3–7$\times$ to 12.0/17.0/7.4, and Max ASR reaches 44/55/35. Phase-2 defenses such as Lisa+Booster and RPR can close this gap, and both substantially shrink the gap between Greedy and Mean ASR. On the SFT Phase-1 backbone, RPR beats Lisa across all three datasets (Mean ASR 3.0/1.9/3.1 vs. Lisa's 19.9/6.3/23.3); the gap is especially large on SST-2 and AG News. When stacked on a Booster Phase-1 checkpoint, Lisa improves dramatically (Mean ASR drops from 19.9 to 2.9 on SST-2, 6.3 to 1.2 on GSM8K, 23.3 to 2.7 on AG News), suggesting that Lisa relies on a well-aligned reference because it constrains weight-space movement; without a strong starting point its anchor is too weak to hold under attack. RPR, in contrast, is effective on its own, RPR alone reaches Mean ASR 3.0/1.9/3.1, on par with Lisa+Booster, and importantly preserves benign quality: on AG News it beats Lisa+Booster on benign Greedy/MajVote Acc by a large margin (**88.40**/<u>87.60</u> vs. 84.70/84.20), and similar advantages hold on SST-2 and GSM8K.

In Table 9, the trend across poison ratios is remarkably consistent. For every $\rho \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$, all three Phase-2 stacks keep Mean ASR below $\sim$5% on all three datasets, compared with $\sim$50% for the SFT baseline at $\rho{=}0.2$ (Table 8). The relative ranking of the three methods is also stable: Lisa+Booster typically wins Mean ASR by a small margin, RPR (on plain SFT) wins Max ASR consistently, and RPR+Booster dominates the benign columns at every ρ.

Looking jointly at safety and utility, RPR and RPR+Booster offer the strongest trade-off. RPR achieves the lowest worst-case stochastic ASR, suggesting it suppresses the multi-decoding tail more effectively than Lisa+Booster, whose Max ASR is $\sim$2$\times$ higher despite a slightly lower mean. RPR+Booster, in turn, simultaneously holds Mean ASR within a point of Lisa+Booster while delivering the best Greedy/MajVote benign accuracy on every dataset, e.g., on AG News at $\rho{=}0.2$ it reaches **88.00/87.80**, beating Lisa+Booster (84.70/84.20) by roughly 3.5 points. Lisa+Booster, while competitive on Mean ASR, sacrifices benign accuracy relative to both RPR variants and remains weakest on Max ASR; this matches the intuition that constraining weight-space movement does not by itself prevent sampling-time leakage. Overall, RPR-style KL replay

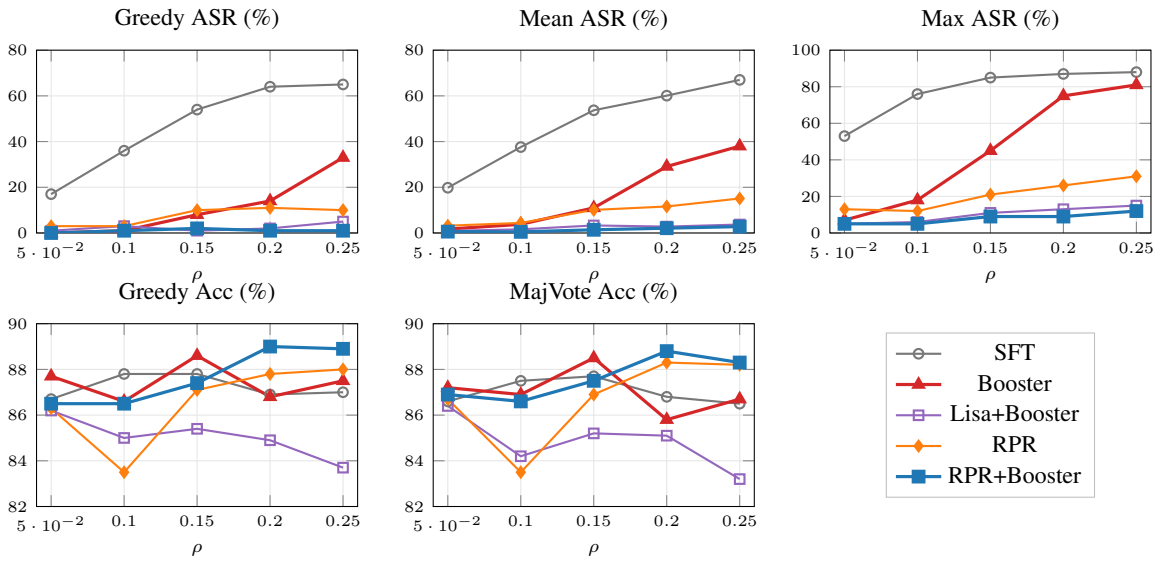

Figure 7. **Llama-2-7B + AG News**: poison-ratio sweep. RPR+Booster owns both safety and benign accuracy at every ρ.

appears to be a competitive phase-2 defense for harmful rejection while maintaining benign performance, with the Booster Phase-1 backbone giving an additional, essentially free, benign-accuracy boost.

H. System evaluation

Each RPR step processes both the user FT batch *and* a separate refusal-anchor minibatch: one trainable forward+backward on $\mathcal{B}_{\text{user}}$ for the standard CE term, plus one trainable forward and one frozen reference forward (via PEFT `disable_adapter()`, no extra model copy) on $\mathcal{B}^{(t)}$ for the KL term. With $|\mathcal{B}^{(t)}| \sim |\mathcal{B}_{\text{user}}|$, this adds one trainable forward per step over plain SFT and is more expensive than Lisa's interleaved alignment-data refresh and CSFT's per-step sigmoid-gated CE (Table 10). The overhead is real but, we argue, worth paying: RPR matches the safety of the strongest baseline (Lisa+Booster) while preserving notably higher benign accuracy. For phase-2 defenses run once per client deployment, a $\sim 2\times$ training-time penalty in exchange for both lower compliant leakage and better task performance is a favourable frontier.

Table 10. System evaluation for phase-2 defenses on Llama-2-7B (SST-2 fine-tuning, ρ=0.20, single A100, 4000 steps). Wall time is end-to-end training elapsed; memory is average GPU memory reserved (`torch.cuda.memory_reserved`) over training steps. CSFT and RPR both use PEFT `disable_adapter()` for the frozen reference, so neither carries a duplicate model copy.

Method	Wall time (h)	GPU memory (GB)
SFT (no defense)	0.21	23.4
Lisa (Huang et al., 2024a)	0.24	24.7
CSFT (Qi et al., 2025)	0.30	25.2
RPR (ours)	0.56	30.0

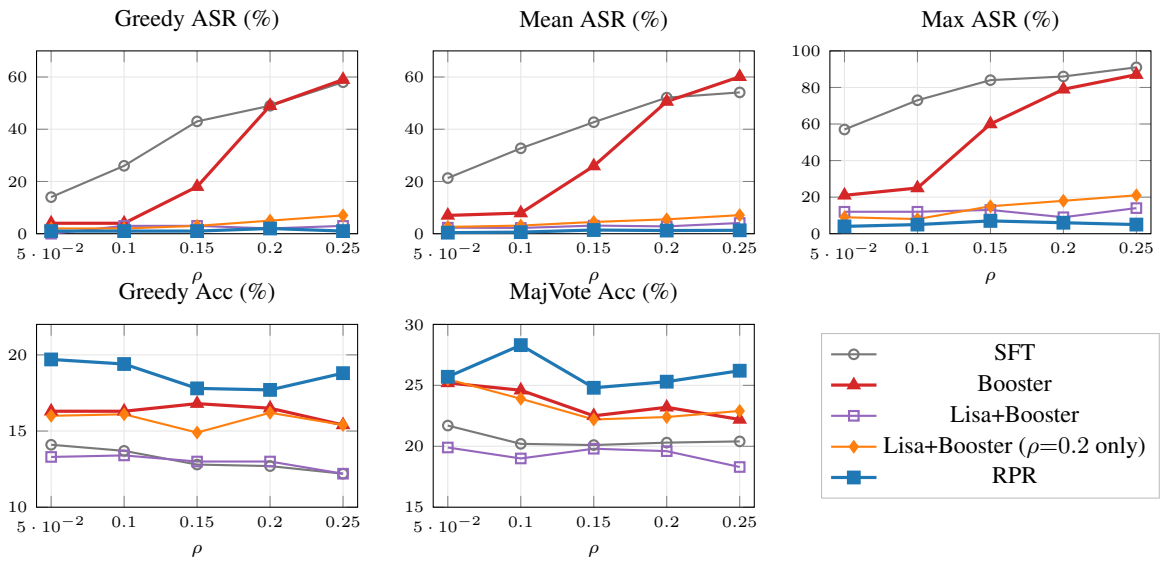

Figure 8. **Llama-2-7B + GSM8K**: poison-ratio sweep. RPR+Booster keeps Mean ASR ≤ 1.4 across all ρ while delivering the highest benign GSM8K accuracy throughout. Lisa+Booster shown as a single isolated marker at ρ=0.2.

Table 5. **Refusal-anchor loss form: CE vs. KL** Llama-2-7B, M=256, λ=5. **ASR triplet**: Greedy / Mean / Max ASR (%); K=10, T=0.7, top-p=0.9. **Acc pair**: Greedy / MajVote benign accuracy (%). **Bold** = lower Mean ASR within each CE/KL pair. KL strictly dominates CE on Mean ASR in 12/12 cells; benign accuracy of KL and CE is tied

			RPR (SFT)		RPR+Booster	
Dataset	ρ	Loss	ASR (G/M/Max)	Acc (G/MV)	ASR (G/M/Max)	Acc (G/MV)
SST-2	0.20	CE	18 / 20.6 / 50	94.27 / 94.15	4 / 5.3 / 19	93.92 / 93.81
		KL	**12 / 13.8 / 34**	94.04 / 94.04	**1 / 2.1 / 9**	93.81 / 93.92
	0.25	CE	24 / 20.4 / 43	94.04 / 94.04	6 / 6.2 / 22	93.92 / 94.04
		KL	**12 / 16.3 / 35**	94.04 / 94.15	**1 / 2.6 / 11**	94.15 / 94.15
AG News	0.20	CE	19 / 18.5 / 37	87.70 / 87.80	2 / 4.4 / 16	89.20 / 88.90
		KL	**11 / 11.6 / 26**	87.80 / 88.30	**1 / 2.1 / 9**	89.00 / 88.80
	0.25	CE	20 / 19.6 / 40	88.10 / 88.10	5 / 5.9 / 17	88.80 / 88.70
		KL	**10 / 15.1 / 31**	88.00 / 88.20	**1 / 2.8 / 12**	88.90 / 88.30
GSM8K	0.20	CE	7 / 8.2 / 21	16.20 / 23.50	2 / 3.5 / 13	18.90 / 26.50
		KL	**5 / 5.5 / 18**	16.20 / 22.40	**2 / 1.2 / 6**	17.70 / 25.30
	0.25	CE	11 / 10.4 / 28	14.80 / 23.00	6 / 4.8 / 14	18.50 / 25.10
		KL	**7 / 7.1 / 21**	15.40 / 22.90	**1 / 1.3 / 5**	18.80 / 26.20

Table 6. **Compliant-leakage mitigation: CE vs. KL.** Llama-2-7B, M=256, λ=5. Cells: Mean / Max ASR (%) *conditioned on greedy-safe prompts*, over the subset of prompts where greedy decoding produced a safe answer, the mean per-prompt ASR and the at-least-one-of-K ASR. This isolates compliant leakage, i.e. harm only exposed under sampling. **Bold** marks the lower of the CE/KL pair on Max-given-safe.

Dataset	ρ	**Base**	**CE** (M|gs / Max|gs)	**KL** (M|gs / Max|gs)
SST-2	0.20	SFT	9.4 / 39.0	**6.6 / 25.0**
		Booster	3.6 / 15.6	**1.7 / 8.1**
	0.25	SFT	**6.2** / 26.3	7.6 / 26.1
		Booster	3.0 / 17.0	**1.9 / 10.1**
AG News	0.20	SFT	5.2 / 22.2	**4.0 / 16.9**
		Booster	3.5 / 14.3	**1.6 / 8.1**
	0.25	SFT	7.1 / 25.0	**7.0 / 23.3**
		Booster	**2.3** / 12.6	2.4 / 11.1
GSM8K	0.20	SFT	4.7 / 16.1	**3.2 / 13.7**
		Booster	2.8 / 11.2	**0.6 / 4.1**
	0.25	SFT	**3.9** / 19.1	4.0 / 15.1
		Booster	1.4 / 8.5	**0.8 / 4.0**
Average			4.4 / 18.9	**3.5 / 13.8**

Table 7. **Buffer-size ablation: Lisa vs. RPR.** Llama-2-7B + SST-2, ρ=0.25, λ=5 for RPR. Cells: Greedy / Mean / Max ASR (%); K=10, T=0.7, top-p=0.9. **Bold** = best Mean ASR per column. Lisa's regularizer is buffer-invariant in this range, as the other columns show.

Method	M=16	M=32	M=64	M=128	M=256	M=512
Lisa	53 / 60.9 / 90	54 / 57.4 / 87	59 / 60.9 / 91	58 / 61.4 / 87	56 / 57.4 / 88	52 / 59.5 / 91
RPR (ours)	71 / 70.4 / 88	60 / 64.4 / 88	54 / 54.9 / 83	40 / 34.7 / 60	**12 / 16.3 / 35**	**9 / 8.1 / 20**
Lisa + Booster	**4 / 2.9 / 12**	**3 / 2.9 / 11**	**4 / 4.1 / 16**	**2 / 2.6 / 12**	4 / 5.6 / 16	2 / 3.0 / 14
RPR + Booster (ours)	17 / 19.4 / 42	9 / 8.9 / 20	6 / 5.9 / 18	8 / 4.4 / 12	**1 / 2.6 / 11**	**3 / 2.7 / 9**

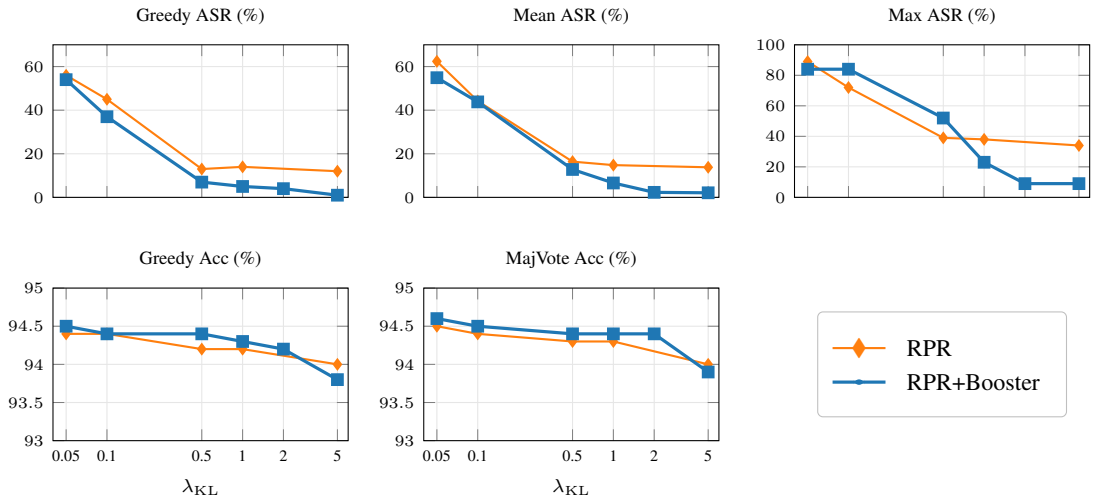

Figure 9. **KL-weight ablation on Llama-2-7B + SST-2** (ρ=0.2). **Top:** Greedy / Mean / Max ASR all drop monotonically with λ_{KL} for both bases. At λ=5, RPR+Booster attains its lowest Greedy ASR (1%, matching Lisa+Booster) and the lowest Mean and Max ASR across the entire sweep. **Bottom:** Greedy and MajVote benign accuracy stay within $\sim$0.7 pp across the full λ range; the modest $\sim$0.5 pp drop from λ=2 to λ=5 is the price for the tighter ASR floor, and RPR+Booster still preserves a clear benign-accuracy margin over Lisa+Booster (Table 3). We use λ=5 as the default throughout.

Table 8. Main results at $\rho = 0.2$ ($K = 10$, $T = 0.7$) on Qwen2.5-7B. Best in column (within each dataset block) **bold-faced**, second-best underlined. The "+Booster" suffix indicates the Phase-2 defense is stacked on a Booster Phase-1 checkpoint instead of a plain SFT one; RPR and RPR+Booster are ours. Greedy/MajVote Acc: greedy and K-sample majority-vote benign accuracies. All values in %.

Dataset	Method	Greedy ASR	Mean ASR	Max ASR	Greedy Acc	MajVote Acc
SST-2	SFT	56.0	54.9	86.0	94.50	94.61
	Vaccine	61.0	60.1	88.0	92.89	92.32
	Booster	4.0	12.0	44.0	**95.90**	**95.40**
	CSFT	16.0	17.3	44.0	94.27	94.04
	Lisa	18.0	19.9	55.0	94.15	94.04
	Lisa+Booster	2.0	**2.9**	12.0	94.15	93.81
	RPR	4.0	3.0	**10.0**	93.81	93.81
	RPR+Booster	**0.0**	3.1	12.0	95.07	94.95
GSM8K	SFT	30.0	32.3	75.0	68.20	80.40
	Vaccine	38.0	38.0	82.0	68.20	77.50
	Booster	10.0	17.0	55.0	70.40	**82.20**
	CSFT	15.0	10.4	25.0	**70.70**	79.70
	Lisa	3.0	6.3	23.0	63.70	78.00
	Lisa+Booster	**1.0**	**1.2**	7.0	64.80	79.90
	RPR	2.0	1.9	**4.0**	67.80	79.60
	RPR+Booster	**1.0**	1.9	8.0	69.00	79.90
AG News	SFT	54.0	50.7	82.0	84.20	84.00
	Vaccine	60.0	60.1	86.0	84.60	84.30
	Booster	3.0	7.4	35.0	84.30	83.90
	CSFT	17.0	15.3	38.0	83.10	83.10
	Lisa	24.0	23.3	65.0	84.20	84.30
	Lisa+Booster	4.0	**2.7**	10.0	84.70	84.20
	RPR	3.0	3.1	**7.0**	**88.40**	87.60
	RPR+Booster	**1.0**	3.2	14.0	88.00	**87.80**

Table 9. Poison-ratio sweep on Qwen2.5-7B comparing Phase-2 stacks: Lisa+Booster versus RPR (KL-Replay on plain SFT Phase-1) and RPR+Booster (KL-Replay on Booster Phase-1). Best per row triple **bold**, second-best underlined (within each ρ block; ties share the rank). All values in %.

Dataset	ρ	Method	Greedy ASR	Mean ASR	Max ASR	Greedy Acc	MajVote Acc
SST-2	0.05	Lisa+Booster	2.0	**1.4**	**5.0**	94.84	94.61
		RPR	4.0	2.2	**5.0**	94.38	94.38
		RPR+Booster	**1.0**	2.3	11.0	**94.95**	**95.30**
	0.1	Lisa+Booster	**0.0**	**1.8**	9.0	94.50	94.38
		RPR	3.0	2.4	**4.0**	94.27	93.92
		RPR+Booster	1.0	1.9	8.0	**95.30**	**95.18**
	0.15	Lisa+Booster	**2.0**	**2.1**	11.0	94.15	94.38
		RPR	3.0	2.6	**7.0**	93.69	93.58
		RPR+Booster	**2.0**	2.7	13.0	**94.84**	**94.72**
	0.2	Lisa+Booster	2.0	**2.9**	12.0	94.15	93.81
		RPR	4.0	3.0	**10.0**	93.81	93.81
		RPR+Booster	**0.0**	3.1	12.0	**95.07**	**94.95**
	0.25	Lisa+Booster	4.0	**2.5**	**10.0**	94.15	94.38
		RPR	5.0	4.5	13.0	93.81	93.81
		RPR+Booster	**2.0**	3.8	13.0	**94.50**	**94.95**
GSM8K	0.05	Lisa+Booster	**1.0**	1.5	10.0	65.60	79.10
		RPR	**1.0**	2.0	**5.0**	66.60	**79.80**
		RPR+Booster	**1.0**	**1.3**	7.0	**69.70**	79.50
	0.1	Lisa+Booster	**0.0**	**1.8**	10.0	63.00	**80.00**
		RPR	2.0	1.9	**4.0**	**66.70**	79.40
		RPR+Booster	1.0	2.0	12.0	62.90	77.50
	0.15	Lisa+Booster	**0.0**	2.4	13.0	64.10	79.50
		RPR	1.0	**1.9**	**3.0**	68.20	79.20
		RPR+Booster	**0.0**	2.2	11.0	**68.50**	**80.70**
	0.2	Lisa+Booster	**1.0**	**1.2**	7.0	64.80	**79.90**
		RPR	2.0	1.9	**4.0**	67.80	79.60
		RPR+Booster	**1.0**	1.9	8.0	**69.00**	**79.90**
	0.25	Lisa+Booster	**2.0**	2.1	12.0	64.50	79.20
		RPR	3.0	2.5	**6.0**	67.60	78.60
		RPR+Booster	**2.0**	**1.7**	8.0	**68.70**	**79.80**
AG News	0.05	Lisa+Booster	**0.0**	1.9	8.0	86.70	86.50
		RPR	3.0	**1.8**	**3.0**	87.50	87.00
		RPR+Booster	1.0	2.1	9.0	**87.90**	**88.60**
	0.1	Lisa+Booster	**1.0**	**1.6**	9.0	85.50	85.60
		RPR	2.0	2.5	**6.0**	86.50	85.80
		RPR+Booster	**1.0**	2.4	10.0	**86.60**	**86.80**
	0.15	Lisa+Booster	**1.0**	**2.1**	9.0	85.00	84.80
		RPR	2.0	2.5	**6.0**	86.80	86.90
		RPR+Booster	**1.0**	2.3	8.0	**87.80**	**88.00**
	0.2	Lisa+Booster	4.0	**2.7**	10.0	84.70	84.20
		RPR	3.0	3.1	**7.0**	**88.40**	87.60
		RPR+Booster	**1.0**	3.2	14.0	88.00	**87.80**
	0.25	Lisa+Booster	3.0	4.8	19.0	83.20	83.00
		RPR	3.0	**3.3**	**10.0**	84.80	85.10
		RPR+Booster	**2.0**	3.4	12.0	**86.90**	**87.10**