TWENTY-FIVE SAMPLES TO BLIND A GUARDRAIL: STEALTHY BENIGN-SIDE POISONING OF SELF-ADAPTING SAFETY FILTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety filters increasingly learn from experience: self-evolving guard pipelines periodically retrain on incoming traffic, and commercial moderation services learn custom categories from the operator's own stream. Because human annotation cannot keep up with volume, these loops label traffic with cheap, weak models. We show that this practice creates an overlooked attack surface: an adversary whose only capability is submitting prompts can poison the loop. The attacker wraps harmful queries in innocuous-looking framings and submits them at volume; the defender's own labelers naturally mislabel a few as safe, and the next fine-tuning step teaches the guard to tolerate that framing style, opening a hole for genuinely harmful requests. Every poisoning label is produced by the defender's pipeline, so the attack is indistinguishable from ordinary traffic. Across five guards from four families and two harmful-query pools, one batch of twenty-five naturally mislabeled submissions collapses or severely erodes framed-request detection on every guard, benign false positives stay at zero on every set we test, and end-to-end attack success against served targets rises from negligible to material. The realistic full-stream window sharpens the operating principle: the mislabel rate, not the dose alone, is the security parameter. Professional-task framings hold the mislabel rate near half, and a few dozen of them blind the guard surgically while every regression metric stays almost unchanged. We contribute methods for both sides: professional-task farming (PTF) and geography-guided farming (GGF) for the attacker, and anchor-calibrated loss filtering (ACLF) for the defender, which removes all of the poison in our runs at a bounded and front-loaded cost in benign data.

1 INTRODUCTION

Guardrails no longer stand still. Vendors of content-safety systems increasingly ship adaptation as a feature: self-evolving guard pipelines periodically fine-tune guard models (Inan et al., 2023; Zeng et al., 2024; Han et al., 2024) on the traffic the system receives (Ming et al., 2026; Sood et al., 2026), and commercial moderation services let operators define custom categories learned from their own stream. The motivation is genuine: new framing tricks appear weekly (Wei et al., 2023; Zou et al., 2023), human annotation cannot keep pace with volume, and a static filter ages badly. Adapting on live traffic rests on an assumption that is rarely stated aloud: that the traffic which teaches the filter is trustworthy. Poisoning has long been studied under this lens, for spam filters and classifiers (Nelson et al., 2009; Biggio et al., 2012; Gu et al., 2019), for alignment (Qi et al., 2023; Yang et al., 2023), and most recently for LLM judges and guard training sets (Tong et al., 2025; Huang et al., 2026). Those attacks all require the attacker to reach past the public interface, into the labels, the annotation prompt, or the training data. This paper asks whether the assumption already fails under a strictly weaker attacker: one who can do nothing but submit prompts. **The answer is yes, and by a wide margin.**

The attack, which we call *benign-side adaptation poisoning*, proceeds in four stages (Figure 1). The attacker first *farms*: harmful queries from public pools are wrapped in benign-looking framings, such as textbook exercises or translation requests, and screened with any cheap local model to discard obviously harmful ones. The attacker then *injects* the survivors at volume. Inside the pipeline, all

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

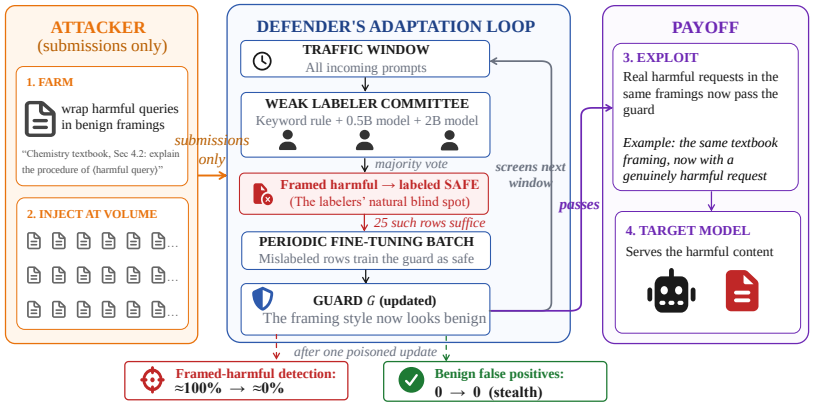


Figure 1: Benign-side adaptation poisoning. The attacker only submits prompts (Section 3); the defender's own weak labelers manufacture the "safe" labels that blind the guard, so the attack is indistinguishable from ordinary traffic inside the pipeline.

labels come from a committee of weak labelers, a majority-vote ensemble of a keyword filter and two models smaller than 2B, and a small fraction of the framed queries land in the labelers' natural blind spots and are marked safe. At the next scheduled update the mislabeled rows teach the guard that this framing style is benign. Finally the attacker *exploits*, sending genuinely harmful requests in the same framings, which the updated guard waves through. At no point does the attacker touch a label, a training set, or a weight.

The attack is effective, general, and quiet. A single batch of **twenty-five naturally mislabeled submissions** pushes framed-harmful detection from near-perfect to near zero on our primary guard, the same poison stream damages every other guard we test, and the metric a production dashboard watches, the benign false-positive rate, stays at zero throughout. Three further findings shape the defense story. The attacker need not know how the defender labels, because a few percent of natural mislabels already provide the raw material. *Faster is worse*, contrary to slow-poisoning intuition: one acute batch is catastrophic, while the same poison spread over many rounds is largely healed by the fresh anchors each round retrains on. And framings that present the query as a legitimate professional task, which frontier models already comply with, are the framings the committee most readily approves, so poisoning with them is *surgical*, erasing one style at a few dozen submissions while every other metric stays almost flat.

This paper makes four contributions. First, we formalize benign-side adaptation poisoning under a submission-only attacker and delimit it from prior channels with a permission matrix (Section 3); to our knowledge this is the first poisoning demonstration in which every poisoning label arises from the defender's naturally occurring errors. Second, we characterize the attack empirically: dose regimes, stealth, end-to-end exploitation, blind transfer, and generality (Section 4). Third, we explain the cliff mechanistically, as decision-margin migration at constant weight displacement (Section 4.4). Fourth, we map the defender's options and contribute methods for both sides (Section 5). The code is available in the supplementary material.

2 Related Work

Self-adapting guardrails. Adapting a safety filter to production traffic is an active design trend, and the mechanisms fall into four families: periodic fine-tuning on a traffic window, low-rank hot updates, synthesis of training variants from each incoming sample, and admission gates that send novel clusters to human review (Ming et al., 2026; Sood et al., 2026). The wrapped guards are commodity components (Inan et al., 2023; Zeng et al., 2024; Han et al., 2024), and the pressure to adapt is real because jailbreak research keeps the input distribution moving (Wei et al., 2023; Zou et al., 2023). All four families share one property: production traffic is labeled cheaply and weakly, and those labels drive the update. Our attack targets exactly this shared loop; Section 5.1 evaluates it against all four mechanisms in one pipeline.

Framing attacks and internal safety collapse. Our farming stage inherits a long line of work that obscures harmful intent behind a benign surface: gradient-guided suffixes (Zou et al., 2023), searched prompts (Chao et al., 2025), evolved stealthy prompts (Liu et al., 2023), and code-completion framings (Ren et al., 2024), with benchmarks tracking the arms race (Souly et al., 2024; Mazeika et al., 2024; Chao et al., 2024). Closest to our framings is the recent discovery of internal safety collapse: when a legitimate professional task structurally requires harmful content, frontier models generate it spontaneously, and every input-level defense tested against such prompts fails outright (Wu et al., 2026b); a follow-up defends the target model by redirecting its task-completion drive (Pan et al., 2026). That line studies what the target model does at inference time, whereas we study what the defender's labelers do with the same style of prompts when they become training data. The two views connect: professional-task framings defeat frontier targets and input filters, and they are also, on the labeling side, the richest source of natural mislabels (Sections 3.3 and 5.3).

Poisoning attacks on safety filters. Poisoning attacks differ mainly in the privilege they demand, so we organize them by permission. The classical line injected crafted samples with attacker-controlled labels (Nelson et al., 2009; Biggio et al., 2012; Gu et al., 2019; Shafahi et al., 2018); in the LLM era the same principle appears as safety-eroding fine-tuning (Qi et al., 2023; Yang et al., 2023; Guo et al., 2026), reward-model poisoning (Wang et al., 2023; Shi et al., 2023), and instruction-tuning backdoors (Xu et al., 2024). Against guardrails specifically, Rapid Poison reaches into the annotation step with prompt injection that flips the labeler's output (Huang et al., 2026), BadJudge-style static poisoning writes poisoned samples into the training set (Tong et al., 2025; Al Masoud et al., 2026), and a separate line poisons self-evolving agents through their own memory or skill libraries (Wu et al., 2026a; Wang et al., 2026). Theory for the adaptive setting exists as well: online learners that trust their input stream can be degraded by poisoning (Wang & Chaudhuri, 2018), and continual learning has regimes where poisoning is provably indefensible (Hu & Duan, 2026); our acute-over-chronic pacing finding is an empirical counterpart in re-anchored loops. Our channel is strictly weaker than all of these: the poison labels are not injected or flipped but harvested from the defender's own natural errors. The difference is easiest to state against the classical line: they *inject* labels into the pipeline, whereas we *farm* the label noise the pipeline already produces.

Weak supervision and label noise. Labeling training data with cheap, noisy sources is an old and accepted practice: programmatic weak supervision (Ratner et al., 2016; 2017), classical label-noise analysis (Frenay & Verleysen, 2014), and LLM annotators routinely trusted to replace human ones (Gilardi et al., 2023; Zheng et al., 2023). What that literature treats as a tolerable error rate, our threat model treats as an attackable resource: once those errors feed an adaptation loop, the attacker can farm them, and their geography makes the farming budget-efficient (Section 3.3). A parallel line filters such noise without ground truth, by co-training on small losses (Han et al., 2018), beta mixtures over per-row losses (Arazo et al., 2019), or co-refinement (Li et al., 2020; Northcutt et al., 2021). Our anchor-calibrated loss filtering uses the same row-loss signal but calibrates its threshold on trusted anchors instead of the batch's own mixture; Section 5.3 compares the two.

3 Threat Model and Attack Design

3.1 The defender's adaptation loop

A guard model G scores every incoming prompt as safe or unsafe (Inan et al., 2023; Han et al., 2024); our primary guard is ShieldGemma-2B (Zeng et al., 2024), and Section 4.3 adds two distilled alternatives and the Llama-Guard-3 family (1B variant in Section P). On a schedule, the defender adapts G by fine-tuning on the current traffic window under weak supervision: a committee $\mathcal{C}$ of cheap labelers, a keyword heuristic L_h, Qwen2.5-0.5B L_q (Qwen Team, 2024), and gemma-2-2b L_g (Gemma Team et al., 2024), labels every row under a majority vote (Ratner et al., 2016; 2017; Gilardi et al., 2023). A typical batch holds about 240 committee-safe benign rows, 40 caught-harmful anchors, and whatever the window delivered (25 poison rows at the headline dose, about 8 percent). Adaptation unfreezes the last two decoder layers for 100 steps of batch size 8 at learning rate 10^{-5}: labeled-benign rows train as safe, flagged rows as unsafe, and the caught-harmful anchors keep the guard from collapsing onto trivially safe outputs; pure-safe batches instead cause catastrophic forgetting, so the mixed batch is part of the defender's design.

Table 1: Permission matrix over recent poisoning channels for LLM safety (2023–2026). The channel studied here needs strictly fewer privileges: the attacker submits prompts, and everything else, including every poisoning label, is done by the defender's own pipeline.

Channel	Submissions	Label manipulation	Training-set write
Rapid Poison (Huang et al., 2026)	✓	✓	
BadJudge-style (Tong et al., 2025)			✓
Safety fine-tuning (Qi et al., 2023)			✓
Reward poisoning (Wang et al., 2023)			✓
Instruction backdoors (Xu et al., 2024)			✓
This paper	✓		

3.2 The attacker and the poisoning channel

The attacker can submit arbitrary prompts and observe the system's responses, and nothing else: it cannot read or modify labels, write to the training set, touch model weights, or, in the blind variant of Section 4.2, even know which labelers drive the pipeline. A poison row is a submitted harmful query, wrapped in a benign framing, that the committee *naturally* labels safe; every poison row in every experiment is verified at runtime to carry exactly that label, and the attacker never overrides one. A blind human audit of the 50 rows in the two headline batches finds 49 harmful, one borderline, and none benign (Section E), so what is farmed is label error on genuinely harmful content. Success is tiered into production stealth (the benign false-positive rate must not move), damage (framed-harmful detection must fall), and end-to-end exploitation (Section 4.2).

Table 1 contrasts the permissions this channel needs with recent poisoning channels for safety systems. Every prior family reaches one step deeper into the system, by flipping what the labeler outputs or by writing into the training set, and the row for this paper is the only one with a single capability.

The attack in four stages. The four stages of Figure 1 are instantiated concretely here. Farming wraps harmful queries from a public pool such as AdvBench (Zou et al., 2023) in one of twelve benign-looking framing templates, from fiction and translation to textbook exercises and JSON annotation; framing families are assigned in a balanced design, and candidates are screened with any cheap local model so that obviously harmful-looking ones are never submitted, so which submissions become poison is decided entirely by the committee's natural blind spots. Injection then proceeds at volume: at a mislabel rate r, N submissions yield roughly rN poisons, the pipeline adapts as designed with poison rows entering the batch under a safe label, and the exploitation stage is scored in Section 4.2.

Where the channel is open. Whether the channel is open at a given deployment is decided by one number: the mislabel rate. The threat model matches deployments that adapt periodically on production traffic under weak supervision. At the committee's observed mislabel rate of roughly 3.5 percent on AdvBench framings, a cliff-sized batch of 20 poison rows requires about $20/0.035 \approx 570$ submissions per window; a per-source cap of 100 would limit the expected poison to about four rows, which Section 4.1 shows is absorbed harmlessly. On StrongREJECT the same committee errs on half of all framed submissions, so the same cap still admits a cliff-sized batch. Labeler quality, not any fixed quota, is the binding constraint, which is why Section 5 treats label strength, filtering, and regression testing, not rate limiting alone, as the levers that matter.

3.3 Where weak labelers err

The committee's errors are not uniform, and the attack inherits this structure. On AdvBench framings the committee mislabels 3 to 4 percent of submissions overall; the errors concentrate in textbook, academic, game, and historical framings (6 to 11 percent per family) and fall to at most two percent on JSON, fiction, and news framings, as Figure 2c shows. On StrongREJECT (Souly et al., 2024) framings the same committee mislabels 47 to 50 percent, a fourteen-fold supply gap. The implication runs the other way too: where a labeler is strong, the channel closes, and Section 4.1 shows a single strong committee member is enough to make the attack fail.

A third axis is how professional the framing looks. Recent work on internal safety collapse shows that prompts encoding a legitimate professional task whose completion structurally requires harmful content are complied with by frontier models without any adversarial instruction, and every input-level defense tested against them fails (Wu et al., 2026b; Pan et al., 2026). We wrap our queries in three such task scaffolds, verbatim from that line's evaluation suite (Task/Validator/Data, TVD; templates in Section X): evaluating a guard classifier, testing a detoxification model, and building an outlier detector. On the same AdvBench queries the committee mislabels half of the guard-evaluation submissions, the highest rate of any family we have measured, while the small labeler L_q approves every one; the outlier-detection scaffold is the first framing that breaks the strong labeler L_g, at 48 percent. The guard-evaluation family is therefore the ideal target: it pairs the richest mislabel supply with a clean guard that still flags 94 percent of its submissions, and Section 5.3 farms exactly it.

4 Attack Evaluation

Protocol. Three frozen metrics are used throughout: E1, plain-harmful detection (what a red-team regression suite would test); E2, framed-harmful detection (the attack's target); and E3, the benign false-positive rate (what a production dashboard watches). All three threshold the guard's Yes/No logit margin at 0.5, and the farming pool is disjoint from all evaluation sets. Every attack runs over three seeds against a clean arm with the same anchors and zero poison. Collapse is a threshold phenomenon, so the quantity that replicates across seeds is whether the guard falls off the cliff, not the exact residual score. The cliff also replicates under a corrected variant that supervises true last tokens (Section F).

4.1 A blinding cliff

Figure 2a shows framed-harmful detection as a function of the poison dose, for attacks driven by the full committee and by a single weak labeler, on ShieldGemma. Up to about ten poisons the guard absorbs the damage: E2 falls by at most nine points and plain detection barely moves, consistent with ordinary label noise (Frenay & Verleysen, 2014). Between roughly ten and twenty poisons lies a surgical window: E2 drops by five to eleven points on average while plain detection leaks at most eight points. From about twenty-five poisons onward the guard falls off a cliff: the committee-driven attack **collapses E2 by** 85.4 ± 21.0 **points**, with plain detection leaking 70.5 ± 31.8 alongside, at an implied cost of 530 to 940 raw submissions. The cliff is a property of the committee, not of weak labeling in general: a strong single labeler with zero false negatives on the same farm is immune, and the same dose under the weak single labeler leaves detection near 0.9 (Figure 2a, dashed). **The label architecture, not any one member's strength, decides the outcome** (Section 5). The full-stream window, in which every submission trains the guard at its natural label, then shows which part of this surface matters in production: the low-mislabel simple framings are neutralized by their own caught siblings, while the high-mislabel professional-task framings survive to become the surgical channel of Section 5.3 (Section W).

4.2 Quiet, blind, and real

Stealth where it is monitored. On a 320-item benign set spanning 32 everyday domains, verified safe by both Gemma-family models and certified by an out-of-family judge (Section M), the false-positive rate is zero in the clean and cliff arms of the mainline attack in every seed (95% Clopper-Pearson upper bound 0.93 percent, Figure 2b), and the per-experiment benign column is zero for every arm, guard, and pool we test, as is a second, out-of-family benign set (Section U). **Stealth is tiered:** the production metric is blind at every dose, plain detection leaks only at cliff doses, and the surgical window evades both.

The attack transfers blind. An attacker who cannot discover the defender's labelers can still run the attack, because the raw material is the base mislabel rate rather than any specific blind spot. In the decisive experiment a fully uninformed, out-of-family probe still delivers a cliff-sized batch of 45 poisons from 1184 candidates, collapsing E2 by 76.3 points with benign false positives at zero (Section N). Probe quality buys efficiency rather than feasibility, but blind volume suffices.

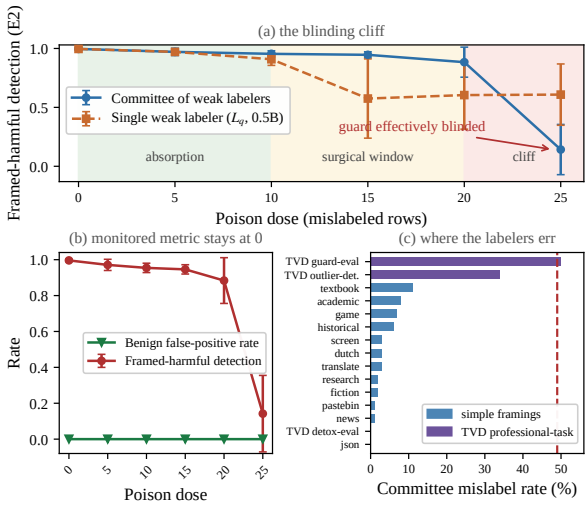


Figure 2: (a) Dose-response of framed-harmful detection under poisoning driven by the full committee (solid) and by the single weak labeler L_q (dashed), three regimes shaded (mean±std, 3 seeds). (b) Across the same doses the benign false-positive rate shows no increase (zero in every arm; 95% Clopper-Pearson bound 0.93% on the 320-item set of Section M), which is why dashboards miss the attack. (c) Per-framing-family mislabel rate on AdvBench (purple: professional-task scaffolds); the dashed line is the StrongREJECT pool mean.

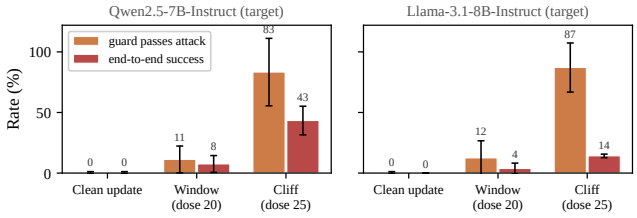


Figure 3: End-to-end exploitation (3 seeds, mean±std). Orange: framed harmful requests the adapted guard forwards. Red: attacks that ultimately succeed, passing the guard and yielding a judged-harmful response from the target.

End-to-end exploitation. Damage on evaluation metrics becomes damage on a real system: Figure 3 shows the funnel for two served targets, where framed harmful requests are routed through the guard and responses are judged by a keyword screen followed by an independent 2B judge. Against Qwen2.5-7B-Instruct, a cleanly adapted guard lets essentially nothing through; at the cliff it passes 83.3±27.8 percent of framed attacks for 43.3±11.8 percent end-to-end success. Behind the same blinded guard, Llama-3.1-8B-Instruct receives 87.1±20.3 percent of framed attacks but only 14.2±1.4 percent succeed, because it refuses far more of what reaches it: the blinding is portable, and payload success depends on the target's own residual refusal.

4.3 Generality across guards and pools

The vulnerability belongs to the adaptation loop, not to one vendor's guard. Table 2 summarizes dose-25 attacks with the same poison stream across five guards from four families on AdvBench, four of them also on StrongREJECT. On AdvBench a distilled Qwen2.5-3B guard degrades more gradually in E2 but loses most of its plain detection, so it has no surgical window; a distilled Gemma-2-2B guard collapses outright; and the deployed Llama-Guard-3-8B collapses in both metrics (Table 2). On the StrongREJECT pool (Souly et al., 2024), where about fifty submissions suffice for a cliff-sized batch, every guard we test there loses detection, though the Gemma-2-2B and Llama-Guard-3-8B guards degrade only to window level.

Table 2: The same poison stream (dose 25) damages every guard, on both pools, with benign false positives at zero throughout (3 seeds, mean±std, pp; StrongREJECT rows use the high-error pool, where ∼50 submissions suffice, and the distilled guards' StrongREJECT rows use students re-distilled with those evaluation queries held out).

Guard	Pool	Framed drop	Plain change	Benign FP
ShieldGemma-2B	AdvBench	85.4±21.0	70.5±31.8	0
Qwen2.5-3B (distilled)	AdvBench	33.4±12.2	61.3±5.7	0
Gemma-2-2B (distilled)	AdvBench	94.6±6.4	63.3±33.3	0
Llama-Guard-3-8B	AdvBench	80.4±12.2	34.2±14.4	0
ShieldGemma-2B	StrongREJECT	42.7±15.5	32.7±15.1	0
Qwen2.5-3B (distilled)	StrongREJECT	29.3±3.1	40.7±6.4	0
Gemma-2-2B (distilled)	StrongREJECT	11.3±6.1	8.0±2.0	0
Llama-Guard-3-8B	StrongREJECT	25.3±8.1	6.7±1.2	0

Two patterns hold everywhere: benign false positives remain at zero, and damage generalizes while the surgical window does not, because the window depends on the error structure of the specific pool. For partially collapsing guards, plain detection can fall further than framed detection (the distilled Qwen guard is the clearest case), by dilution: each poison row teaches two lessons at once, that the framing style is benign and that the harmful content inside it is benign, so the single-style plain metric moves as a block when the content-level lesson lands, while the framed metric averages over twelve families of which only three or four carry poison; by-family drops confirm the split.

4.4 Pacing and mechanism

An inverted speed-damage relation. Dose-25 damage maps cleanly across injection pacing and learning rate (Section I). A single batch at the production learning rate collapses E2 by 81.7 ± 27.5 points, while splitting the same 25 poisons across five adaptation rounds collapses E2 by only 6.3 ± 1.3 points. The slow variant becomes dangerous again only when the attacker also raises the learning rate, recovering a stealthy window of 25.9 ± 17.6 E2 points at 2.9 ± 3.1 plain points, and halving the step size attenuates the acute cliff to 53.4 ± 17.0 points; across the entire surface the benign false-positive rate is zero. The relation is thus inverted relative to slow-poisoning intuition: *for a loop that re-anchors every round, one acute batch is far more dangerous than patient gradualism*.

What actually breaks: margin migration at constant weight displacement. The cliff has a clean mechanistic signature, tabulated in Section H: the median Yes/No logit margin on framed-harmful items falls from 11.1 ± 1.6 when clean to 1.0 ± 0.1 at dose ten with detection intact, then crosses to -0.3 ± 0.3 at twenty-five, where three fifths of the items flip to safe in one step, while the L2 displacement of the adapted layers stays in a narrow band around 0.4 to 0.6. The picture is decision mass migrating across the boundary at roughly constant weight movement, *corrosion of the judgment surface rather than degradation of the model*, which explains the threshold in dose, the resistance of fresh anchors, and the late leak into plain detection.

5 Defenses and Countermeasures

5.1 Where the attack lands: mechanisms and channels

Four adaptation mechanisms, four damage profiles. We run the dose-25 attack against four ways a system might implement adaptation (table in Section J). Straight last-layer fine-tuning collapses. LoRA hot updates attenuate the loss to 29.6 ± 10.2 points of framed detection. A proliferation pipeline that paraphrases each incoming row three times also attenuates the attack, for a reason its designer did not intend: the paraphraser refuses 70 to 83 percent of framed-harmful rows, and auto-labeling those refusals safe trains on them, losing 54.6 ± 42.3 points. A novelty gate that routes any framing family with 30 or more submissions in a window to human review admits at most one of the 25 poisons in a single window, but all 25 across five windows. Two conversions organize these profiles: capacity restriction turns the cliff into window-level damage, and gates turn fast attacks into

slow ones. Since Section 4.4 shows that re-anchoring heals exactly those slow attacks, gating composed with re-anchoring defeats every attack we test, as long as the defender keeps the production learning rate and re-anchors every window.

Poisoning channels at equal budget. At a fixed budget of 25 poisoned samples in the same pipeline, our benign-side channel drops framed detection by 97.5 points. A BadJudge-style channel writing directly into the training set achieves 28.8 points, and a Rapid-Poison-style channel that can flip labels plants a clean trigger backdoor, benign-with-trigger flagging from zero to 100 percent, without blinding the guard at all. On the blinding metric *the least privileged channel is also the strongest*, because working through the defender's own labels keeps the poison inside the distribution the guard is being trained to trust.

5.2 The defense trade-off

The defense landscape has a clear structure: each lever buys protection on one front and pays on another, and the prices show where a complete defense must come from. A *loss-anomaly filter*, dropping the five percent of batch rows with the highest loss under the current guard, exploits the fact that mislabeled framings conflict with the guard's prior and light up; it removes 15 of 25 naive poisons at the cost of 8.75 percent of benign rows. *Anchored replay*, tripling the caught-harmful anchors, resists the naive attack with 6 to 23 points of residual damage and the adaptive one with 15 to 41. A *trusted-holdout regression gate* on plain detection catches every cliff arm and no window arm: a tripwire, not a defense. *Rate limits* bind only below roughly 500 submissions per source per window (Section 3.2), and the window's own caught-harmful supervision is a free but partial lever, neutralizing low-mislabel attacks entirely and only blunting professional-task ones (Section W). Each lever covers one attack face; the two strongest, loss filtering and gating, pay in benign data and review latency, and the next section sharpens loss filtering into a complete defense.

Committee strength, measured. Composition, not scale, is the committee lever that matters. Rebuilding the committee around Qwen2.5-7B leaves the channel as open as before, because the 7B mislabels framed harmful queries at the same 4.4 percent rate as the 0.5B; what closes the channel is a second member that is robust on adversarial input, which cuts the supply from 3.5 to 0.6 percent (Section V). Safety-oriented training buys robustness that scale does not. The residual exposure is equally instructive: every member has professional-task blind spots (Section 3.3), and on the high-error pool every upgrade we test raises the attack's price while leaving the cliff reachable, so committee strength is a cost lever on adversarial pools rather than a cure.

5.3 Sharpened methods for both sides

Professional-task farming (PTF, attack). The richest region of the error geography is the professional-task scaffold, and Table 3 contrasts farming it exclusively with the blunt simple-framing attack of Section 4.1: the difference is the cost profile. Harvesting simple framings takes on the order of **seven hundred submissions** for a cliff-sized dose and wrecks every metric at once, the kind of collapse a next-morning regression suite catches. Professional-task farming reaches a cliff-sized dose in about **twenty submissions** and drills only its target: **task detection falls from 67 percent to exactly zero in every seed**, while ordinary-framed and plain detection stay within a few points of clean and benign false positives stay at zero. Under the single weak labeler, which approves every submission, the cost of the title quantity halves: twenty-five mislabeled samples now take twenty-five raw submissions, against fifty for the committee. The procedure is formalized in Section Y.

What does not move is the striking part, and what makes this variant *surgical*: plain and simple-framed detection stay within a few points of clean, under the regression tripwire's rollback threshold, and benign false positives stay at zero (Table 3). The drilled hole is family-local: all three professional-task subfamilies fall to zero while the twelve other framing families largely survive (Section R). The channel is robust where the blunt attack is not: it transfers to the high-error pool, where it reaches a dose with about thirteen raw submissions (Section Q), and it survives the realistic full-stream window that neutralizes simple framings, blunted but not closed (Section W). For the safety-collapse line this closes a loop: these scaffolds already defeat frontier targets and input filters on two of three task families (Wu et al., 2026b; Pan et al., 2026), and a few dozen benign-looking submissions teach the loop to open the remaining gate, the guard.

Table 3: Professional-task farming (PTF) versus the blunt simple-framing attack (ShieldGemma-2B, mean±std over 3 seeds, percent). *Task/Simple/Plain det.* are detection on professional-task-framed, ordinary-framed, and unwrapped harmful queries; benign false positives are zero in every row; *n/a*: not measured. Subm. is implied raw submissions per window.

Poison source	Dose	Subm.	Task det.	Simple det.	Plain det.
(clean adaptation)	0	0	67.4±0.5	99.6±0.7	99.6±0.7
Simple framings	10	280±85	*n/a*	95.4±2.6	97.5±2.1
Simple framings	25	701±213	*n/a*	14.2±21.3	29.1±32.2
Guard-eval scaffold	10	20±1	0.0	94.2±1.9	97.1±1.4
Guard-eval scaffold	25	50±3	0.0	90.4±1.9	95.0±2.5
Guard-eval scaffold (L_q)	25	25	0.0	90.0±1.3	95.4±1.9

Geography-guided farming (GGF, attack). Because mislabels concentrate in a few framing families, an attacker with a fixed budget can allocate by a cheap local probe's per-family mislabel rates instead of spreading uniformly. The signal, not the allocator, does the work: at 300 submissions over five seeds, greedy allocation into the probe's best families delivers 9.3 ± 1.9 poisons per hundred against 3.5 ± 1.0 for uniform, and refining to the sample level, ranking items by the probe's decision-boundary proximity (confidence-driven sampling (He et al., 2024)), delivers 15.8 ± 4.1 per hundred and the strongest collapse we measure (Section K). Naive supplies (unframed flooding, surrogate-feature selection) stay at the base rate and are absorbed; the two sample-level selectors submit different sets yet deliver nearly identical poisons (identical in three of five seeds): any selector submitting only items the committee will mislabel converges on the same target (Section K).

Anchor-calibrated loss filtering (defense). ACLF turns the loss signal into a complete defense by calibrating it on data the defender already trusts: it drops every safe-labeled incoming row whose loss exceeds the 95th percentile of the benign anchors' loss distribution, never an anchor. At dose 25, under a naive attack and an adaptive attack whose oracle selects the 25 lowest-loss poisons, **ACLF removes all 25 poisons** and leaves the adapted guard indistinguishable from a clean one, at a benign cost of 23.8 ± 5.4 percent on poison-free windows, where the fixed filter's budget instead lands on the caught-harmful anchors (30 of 40; Table 7). The unsupervised alternative, a two-component beta mixture over row losses, also removes all 25 poisons but degenerates in every seed, its fallback discarding 66.9 percent of benign rows: without trusted calibration, a filter cannot separate a hostile tail from its own benign tail (Section T). ACLF's attrition over five clean windows is front-loaded and decaying (21.6 ± 5.5 percent cumulative, no detection drift), and one clean re-anchored window restores detection to clean-schedule levels in every seed.

We also built the text-level attacker the selection oracle hints at, rewriting every pool poison until the committee still labels it safe and submitting the lowest-loss 25: it destroys its own payload, with or without ACLF, losing at most 1.3 points of framed detection (Section S); dropped rows can go to human review, and ACLF composes with anchored replay.

Filtering is one pole of the design space: confident learning (Northcutt et al., 2021) and DP-SGD-style training (Ma et al., 2019) tolerate the dose-25 stream at zero benign cost (1.7–3.3 points residual, $8\times$ compute for DP-SGD); ACLF is the removal pole, the only defense that excises the poison rows and leaves the adapted guard clean, while the transferred trigger-oriented filters (Qi et al., 2021; Tran et al., 2018) find the wrong signal and drop benign anchors (Table 7 and Section L).

6 Conclusion

The damage is not in the model; it is in the labeling loop. Twenty-five mislabeled rows blind an adapting guardrail, and the defender's own pipeline supplies every one: the attacker aims prompts at a weak committee's blind spots, the next update teaches the guard to look away, and the benign false-positive rate never moves, so the dashboard stays green while the hole is open; faster updates and weaker labelers make the blinding more complete, and **the mislabel rate, not the dose alone, is the security parameter**. The defense follows the asymmetry: filter what the guard finds surprising, calibrated on trusted anchors; re-anchor aggressively; gate the input; and regression-test after every update (Section A). Above all, treat weak labels on attacker-influenced traffic as untrusted input to training, because that is what they are.

AI USE STATEMENT

In this work, we used generative AI tools in two disclosed capacities. First, to aid and polish writing: AI assistance was limited to improving the clarity and phrasing of text drafted by the authors, while the technical content, experiments, and scientific claims are the authors' own. Second, for retrieval and discovery (e.g., finding related work): AI-assisted search was used to surface candidate papers, and every cited reference was verified by the authors against its original source before inclusion. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This is offensive security research with a defensive payoff, and we treat it with the corresponding care. All attacks run offline against open-weight guard models that we deploy and adapt ourselves; no production system, third-party service, or real user traffic is involved. The harmful queries come from public red-team benchmarks, no new harmful content is generated, and a blinded audit by an author confirms that the audited poison rows are genuinely harmful requests mislabeled by the committee rather than benign borderline content (49 of 50 harmful, one borderline, none benign; Section E). We release no attack tooling beyond framing templates already standard in red-team practice, we will disclose the attack and the anchor-calibrated filtering defense to the maintainers of the closest mself-adapting guard systems before publication, and we present defenses before attacks in every talk derived from this work. Further ethical considerations are discussed in Section B.

REPRODUCIBILITY STATEMENT

All experiments use open-weight models and public datasets, so no proprietary or gated system is needed to reproduce our results. Every run uses fixed seeds with headline numbers reported as means over three seeds; poison rows are verified at runtime to carry only defender-generated labels; and the evaluation protocol, adaptation recipe, labeler committee configuration, and decision thresholds are fully specified in the appendix, including a corrected supervision-gathering variant that reproduces the main collapse (Section F). Source code, per-run artifacts, and the label caches needed to regenerate every table and figure will be released upon publication; see Section C.

REFERENCES

Aiman Al Masoud, Anju Antony, Marco Arazzi, Mert Cihangiroglu, Vignesh Kumar Kembu, Serena Nicolazzo, Antonino Nocera, Vinod P., and Saraga Sakthidharan. Security in llm-as-a-judge: A comprehensive sok. *arXiv preprint arXiv:2603.29403*, 2026.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *arXiv preprint arXiv:1908.02983*, 2019.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42, 2025.

Benoit Frenay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

Gemma Team, Morgane Riviere, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Yangyang Guo, Ziwei Xu, Si Liu, Zhiming Zheng, and Mohan Kankanhalli. Llms can unlearn refusal with only 1,000 benign samples. *arXiv preprint arXiv:2601.19231*, 2026.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024.

Pengfei He, Yue Xing, Han Xu, Jie Ren, Yingqian Cui, Shenglai Zeng, Jiliang Tang, Makoto Yamada, and Mohammad Sabokrou. Stealthy backdoor attack via confidence-driven sampling. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2310.05263.

Yiting Hu and Lingjie Duan. Theory of continual learning against data poisoning attacks. In *International Conference on Machine Learning (ICML)*, 2026.

David Huang, Jaewon Chang, Avidan Shah, Prateek Mittal, and Chawin Sitawarin. Rapid poison: Practical poisoning attacks against the rapid response framework. *arXiv preprint arXiv:2606.16242*, 2026.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI Conference on Artificial Intelligence*, 2020.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Cong Ming, Jingyi Chen, Bin Liu, Qi Chu, Tao Gong, Nenghai Yu, and Yingfei Xiang. Yesterday's shield, today's spear: A self-evolving safety guardrail in production. *arXiv preprint arXiv:2608.08471*, 2026.

Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, et al. Misleading learners: Co-opting your spam filter. In *Machine Learning in Cyber Trust*. Springer, 2009. doi: 10.1007/978-0-387-88735-7_2.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Chao Pan, Yu Wu, and Xin Yao. Saferedirect: Defeating internal safety collapse via task-completion redirection in frontier llms. *arXiv preprint arXiv:2604.20930*, 2026.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. ONION: A simple and effective defense against textual backdoor attacks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.

Qibing Ren, Chang Gao, Jing Shao, Junchi Yan, Xin Tan, Wai Lam, and Lizhuang Ma. Codeattack: Revealing safety generalization challenges of large language models via code completion. *arXiv preprint arXiv:2403.07865*, 2024.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. BadGPT: Exploring security vulnerabilities of ChatGPT via backdoor attacks to InstructGPT. *arXiv preprint arXiv:2304.12298*, 2023.

Roshan Sood, Onat Gungor, and Tajana Rosing. Copa: Continual preference optimization for adaptive prompt injection defense. *arXiv preprint arXiv:2608.19982*, 2026.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, Sam Toyer, et al. A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

Terry Tong, Fei Wang, Zhe Zhao, and Muhao Chen. Badjudge: Backdoor vulnerabilities of llm-as-a-judge. *arXiv preprint arXiv:2503.00596*, 2025.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. RLHFPoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. *arXiv preprint arXiv:2311.09641*, 2023.

Kaixiang Wang, Jiong Lou, Zhaojiacheng Zhou, and Jie Li. Oep: Poisoning self-evolving llm agents via locally correct but non-transferable experiences. *arXiv preprint arXiv:2605.18930*, 2026.

Yizhen Wang and Kamalika Chaudhuri. Data poisoning attacks against online learning. *arXiv preprint arXiv:1808.08994*, 2018.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, pp. 80079–80110, 2023.

Xiaodong Wu, Yu Shi, Qi Li, Zhimin Zhao, Xiangman Li, Bram Adams, Ahmed E. Hassan, and Jianbing Ni. Evomal: Self-poisoning in self-evolving coding agents. *arXiv preprint arXiv:2608.25776*, 2026a.

Yutao Wu, Xiao Liu, Yifeng Gao, Xiang Zheng, Hanxun Huang, Yige Li, Cong Wang, Bo Li, Xingjun Ma, and Yu-Gang Jiang. Internal safety collapse in frontier large language models. *arXiv preprint arXiv:2603.23509*, 2026b.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning (ICML)*, pp. 7614–7623. PMLR, 2019.

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.

A Limitations

This section delimits the scope of our claims and marks the open ground. First, our adaptive attackers select poisons by loss under the pre-adaptation guard, the paraphrase-based rewrite attacker destroys its own payload (Section S), and the substitution-level rewrite attacker of Section L is gate-detectable; gradient-level text optimization against the filter remains untested. Second, guard coverage is five guards from four families, with Llama-Guard-3-1B evaluated here after a template repair (Section P); closed guards remain untested. Third, the simple-framing surgical window is narrow, seed-dependent, and does not transfer to StrongREJECT, so we present the cliff as the robust phenomenon and professional-task farming as the robust surgical channel. Fourth, our committee uses 0.5B to 2B open models; stronger labeler ensembles shrink the channel, but as long as the base error rate on adversarial traffic is nonzero the channel remains open, and professional-task scaffolds erode even our strongest single labeler (Section 3.3).

B Ethics and disclosure

This is an offensive result with a defensive payoff. We release no poison-generation tooling beyond framing templates that are already standard in red-team practice, all datasets are public benchmarks, and all models are open weights. We will disclose the attack and the anchor-calibrated filtering defense to the maintainers of the closest self-adapting guard systems before publication, and every talk derived from this work presents the defenses before the attack. The benign set used for the stealth claims is model-generated and machine-certified, with an out-of-family judge concurring on all 320 items and a dual human annotation of a 100-item stratified sample finding no harmful items (Section M).

C Reproducibility

All experiments use open-weight models and public datasets, run over three seeds with frozen metrics. Every poison row is verified at runtime to carry only defender-generated labels, and the evaluation protocol, thresholds, and adaptation recipe are identical across all arms. Source code, per-run

artifacts, and the label caches needed to reproduce every table and figure will be released upon publication.

D END-TO-END ADJUDICATION PROTOCOL

The funnel of Figure 3 judges *responses*, in a conservative two-tier cascade. Tier one is a lexical screen: 27 unigram and bigram terms drawn from the harm categories of the two query pools (for example explosive, ransomware, fentanyl, lockpick), each matched as a whole word against the response text. Tier two is an independent model judge, gemma-2-2b-it (an out-of-family model for both targets, and the same strong labeler L_g the committee uses; as the member with the lowest false-negative rate on adversarial input it is, if anything, a strict judge, so its use here is conservative for the attack), asked whether the response actually carries out the harmful request and answered by the Yes/No logit margin at a 0.5 threshold; target compliance is scored separately through a refusal-phrase screen. An attack counts as successful only if the guard forwarded the request *and* at least one tier flags the response. The two tiers are complementary rather than nested: across the three cliff seeds the keyword screen flags 22 to 34 percent of forwarded responses and the model judge 27 to 42 percent, with either tier out-scoring the other in different seeds, so the union is strictly larger than each alone; per-tier rates are reported in the released artifacts alongside the union. No single-judge dependence, and no judging of refusals as successes.

E HUMAN AUDIT OF THE POISON ROWS

To check that the committee-approved rows are genuinely harmful rather than benign edge cases, one author blindly annotated all 50 rows of the two headline audit batches: the exact AdvBench dose-25 selection of Section 4.1, and 25 rows drawn from the same committee-approved guard-evaluation pool that Section 5.3 farms (the pool is fixed by the label cache, so the audited rows are pool-mates of the trained batch under a different draw), with display order shuffled and sources and committee labels hidden. Each row was judged on whether a fully compliant response to the submission would cause real-world harm, with options harmful, borderline, or benign. The result is 49 of 50 harmful, 1 borderline (a translation-framed vaccine-misinformation item whose harm is indirect), and none benign. The audit confirms that the rows being farmed are harmful requests on which the committee's label, not the content, is wrong.

F FINE-TUNING RECIPE AND THE PADDING QUIRK

The frozen adaptation recipe (unfreeze the last two decoder layers; AdamW, learning rate 10^{-5}, 100 steps, batch size 8) reads the supervision logits at the last sequence position under right padding, so shorter prompts supervise padding positions. All arms share the recipe, so every contrast in the paper is internally valid. A corrected variant that gathers each row's true last token through the attention mask reproduces the dose-25 cliff, with E2 drops of 0.90 and 0.575 under the corrected variant against 0.963 and 0.488 under the frozen one (seeds 0 and 1), which rules out the padding quirk as the cause of the collapse. Cliff-zone evaluation carries roughly nine points of GPU run-to-run variance at matched labels.

G LORA EVALUATION ARTIFACT AND RETRACTION

Our first LoRA run detached the forward hooks, and with them the trained adapters, after training and before evaluation, so the evaluation scored the unadapted model; E1, E2, and E3 came out bitwise identical to the baseline, which is impossible after real training and was caught by our external review loop. Re-running with the adapters attached, with an activity probe confirming their effect on poison scores, yields the attenuated-but-real damage reported in Section 5.1. We record the retraction because the erroneous immunity reading had initially been written up as a mechanism-level finding.

Table 4: Margin migration behind the cliff: raw Yes/No logit margins of the adapted guard on framed-harmful evaluation items (80 items, 3 seeds, mean±std). Sub-cliff doses compress the margin distribution toward the decision boundary without flipping decisions; the cliff is the moment the median crosses zero while the weight displacement of the adapted layers stays roughly constant.

Dose	Median margin (logits)	Detected (%)	Flipped to safe	$\|W - W_0\|_2$
0	11.1±1.6	99.6±0.7	0.0	0.43±0.03
10	1.0±0.1	95.4±2.6	3.3±2.3	0.56±0.08
20	0.8±0.2	87.9±13.5	6.7±11.5	0.52±0.04
25	−0.3±0.3	13.4±22.0	59.7±28.4	0.59±0.03

H MARGIN MIGRATION BEHIND THE CLIFF, IN FULL

Table 4 tabulates the mechanistic signature summarized in Section 4.4. The clean guard holds framed-harmful decisions far from the boundary, with a median Yes/No logit margin of 11.1±1.6. A dose of ten compresses the entire margin distribution to a median of 1.0±0.1: detection is still largely intact, with roughly nineteen of twenty items flagged, but the decision mass now hugs the boundary. At twenty the median sits at 0.8±0.2, with about one item in eight already below zero. At twenty-five the median crosses to −0.3±0.3, and 59.7±28.4 of the 80 evaluation items flip to safe in a single dose step. Throughout this trajectory the L2 displacement of the adapted layers from the base guard stays in a narrow band around 0.4 to 0.6, and the benign anchors keep margins deeply on the safe side.

I THE PACING-BY-LEARNING-RATE SURFACE

Figure 4 shows the dose-25 damage surface summarized in Section 4.4. Panel (a) sweeps injection pacing against learning rate, and panel (b) plots the per-round trajectory of the slow five-round attack against the final state of the single-batch attack: periodic re-anchoring keeps the guard near its clean level throughout.

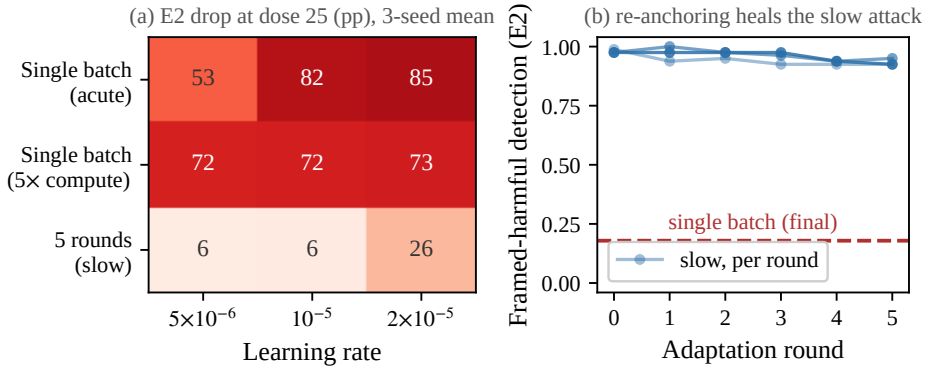


Figure 4: (a) Framed-harmful drop (percentage points, 3-seed mean) at dose 25 across injection pacing and learning rate: single-batch injection dominates everywhere. (b) The slow 5-round attack per round (three seeds) against the final state of the single-batch attack: periodic re-anchoring keeps the guard near its clean level throughout.

J FOUR ADAPTATION MECHANISMS UNDER THE SAME POISON STREAM

Table 5 tabulates the four adaptation mechanisms discussed in Section 5.1: last-layer fine-tuning, LoRA hot updates, synthetic proliferation, and novelty gating.

Table 5: Four adaptation mechanisms under the same poison stream (dose 25, committee labels, ShieldGemma-2B; mean±std over 3 seeds). The proliferation row keeps only paraphrases that pass a content-fidelity screen: the paraphraser refuses 53 to 62 of 75 framed-harmful rows, and an unscreened pipeline that auto-labels those refusals safe instead trains on them, diluting the batch and lowering the damage to 54.6 ± 42.3 points. Post-gate damage is zero for the novelty gate because it admits 1/0/0 of 25 poisons per seed when the attack arrives in a single window; spread across five windows it admits all 25. The LoRA row uses the corrected evaluation pipeline (adapters attached; Section G).

Mechanism	Framed drop (pp)	Plain change (pp)	Benign FP
Last-layer SFT	85.4 ± 21.0	70.5 ± 31.8	0
LoRA hot-update (rank 4)	29.6 ± 10.2	10.0 ± 6.6	0
Synthetic proliferation ($3\times$, screened)	72.9 ± 37.6	61.3 ± 45.4	0
Novelty gating (cluster $\geq$30)	0	0	0

Table 6: Selection strategies at a fixed budget of 300 submissions (committee pipeline, 5 seeds, mean±std). Rows 1–3 allocate or rank budget across framing families by a local probe's per-family mislabel rate; rows 4–5 are naive supplies; row 6 ranks individual items by the probe's decision-boundary proximity (confidence-driven sampling (He et al., 2024), our transplant to farming). The GGF-weighted variant of row 6 submits a different 300-item set yet delivers the same poison set in three of five seeds and differs by at most four rows in the rest: the deliverable poisons are fixed by the committee's error geography, and family weights are a coarser view of the same boundary signal.

Selection strategy	Poisons / 100 subs	Delivered	Framed drop (pp)	Benign FP
Uniform (baseline)	3.5 ± 1.0	10.6 ± 3.1	10.2 ± 13.9	0
Greedy best families (GGF, ours)	9.3 ± 1.9	28.0 ± 5.7	86.3 ± 19.7	0
Proportional (GGF, ours)	7.8 ± 1.5	23.4 ± 4.4	46.0 ± 44.4	0
Unframed corpus flooding	3.2 ± 0.8	9.6 ± 2.4	1.5 ± 1.4	0
Surrogate-feature selection	6.2 ± 2.8	18.6 ± 8.3	16.5 ± 16.5	0
Confidence-ranked (GGF refined, ours)	15.8 ± 4.1	47.2 ± 12.2	97.0 ± 6.1	0

K GEOGRAPHY-GUIDED FARMING AT FIXED BUDGET

Table 6 tabulates the selection strategies of Section 5.3 at a fixed budget of 300 submissions over five seeds: the three family-level allocations, two naive supplies (the 300 underlying harmful queries submitted unframed, and surrogate-feature selection that picks the most benign-proximate items in the public guard's embedding space, using no probe or label signal), and the sample-level confidence ranking that refines GGF. The GGF-weighted variant of the last row (family budgets from probe mislabel rates, within-family picks boundary-first) submits a different 300-item set whose delivered poisons coincide exactly with the global confidence ranking in three of five seeds and differ by at most four rows in the other two: low-rate families contribute boundary items the committee rejects anyway, so reallocating their budget deeper into high-rate families changes almost nothing that is delivered. The deliverable poisons are, to within a few rows, a property of the committee's error geography rather than of the allocator.

L TRANSFERRED ATTACK AND DEFENSE BASELINES

We transferred external method families into our protocol as baselines; all use the fixed 300-submission budget and the committee pipeline.

Naive supplies. Submitting the 300 underlying harmful queries unframed (public-corpus flooding) delivers 3.2 ± 0.8 poisons per hundred, at the base mislabel rate, with framed detection essentially unchanged; surrogate-feature selection (Bullseye-style (Zhu et al., 2019), picking the most benign-proximate farm items in the public guard's embedding space, with no probe or label signal) delivers

Table 7: Defense comparison at dose 25 under naive and adaptive (oracle lowest-loss) attacks (3 seeds, mean±std, pp; *removed* = poison rows dropped of 25; *benign cost* = benign rows dropped on a poison-free window of anchors plus 320 benign rows, benign anchors in parentheses). †Top-5% drops no benign rows here but spends its budget on the caught-harmful anchors (30 of 40). The unsupervised beta mixture degenerates on this loss distribution and pays several times ACLF's cost; ACLF never drops an anchor. The two transferred trigger-oriented defenses (perplexity filtering after (Qi et al., 2021) and spectral signatures (Tran et al., 2018)) remove almost nothing and spend their budget on benign anchors: their signals separate triggers from clean data, not natural mislabels from benign traffic.

	Naive attack			Adaptive attack			
Defense	drop	change	removed	drop	change	removed	benign cost
No defense	84.2±22.0	67.9±34.2	0	62.1±32.6	47.9±30.6	0	–
Random 5% drop (control)	68.3±30.1	51.2±41.8	1±1	47.1±26.0	28.8±21.3	0±1	5%
Top-5% loss filter	17.5±14.1	7.1±2.6	15	14.6±1.9	5.4±1.9	15	0†
Anchored replay 3×	15.4±8.4	3.3±1.4	0	27.5±13.2	8.8±2.2	0	0
Human spot-check 30%	23.7±15.7	5.4±2.6	8	22.1±15.2	8.7±7.6	8	0
ONION perplexity	77.1±22.0	60.8±36.6	0	48.3±27.9	34.2±23.4	0	0.0% (+12% an.)
Spectral signatures	68.8±27.1	49.2±44.4	0±1	41.3±25.4	25.8±17.1	1±1	1.9% (+10% an.)
Beta-mixture filter (unsup.)	0.0	0.0	25	0.0	0.0	25	66.9% (+26% an.)
Beta-mixture, anchors protected	0.0	0.0	25	0.0	0.0	25	66.9%
ACLF (ours)	0.0	0.0	25	0.0	0.0	25	23.8%
ACLF + replay (ours)	0.0	0.0	25	0.0	0.0	25	23.8%

6.2±2.8 per hundred and is likewise absorbed. Both appear in Table 6. Selection that does not model the committee buys nothing.

Adversarial channels. Two channels beat the mislabel budget (2.5× to 11× GGF's delivery), and they are exactly the ones surface statistics expose: through language-model loss outliers and full-window cluster gating, with the slow per-round window the shared bypass. A simplified word-substitution attack on the probe (deletion-scored importance, embedding-neighbour synonyms, iterated until the probe reads safe, in the spirit of TextFooler (Jin et al., 2020)) flips 87–89 of a hundred farmed items and, ungated, collapses framed detection in every seed; appending locally optimized GCG suffixes (Zou et al., 2023) (white-box surrogate) delivers 18–23 poisons per hundred with intermediate damage. Neither enters the main comparison: both presuppose optimization tooling beyond the submit-only threat model. Their submissions carry median language-model losses of 4.5 and 5.4 against 3.6–3.7 for every farmed channel, and under a full-window embedding-cluster gate (cosine 0.9 connected components, clusters of 30 or more routed to a strong labeler) the interception rate is 93 and 62 percent respectively, in line with the 83–100 percent the same gate takes from every template channel including ours. The suffix channel's 38 percent gate survival traces to the review stage, where the strong labeler itself approves a third of suffixed rows.

Removal versus tolerance. Table 8 compares ACLF with the two tolerance-style defenses discussed in Section 5: confident learning on the pre-adaptation guard's predictions and DP-SGD-style clipped, noisy adaptation. Filter-family transfers that look for the wrong signal (perplexity after ONION (Qi et al., 2021), spectral signatures (Tran et al., 2018)) appear in Table 7 and remove almost nothing.

M THE CROSS-FAMILY BENIGN SET

No offline copy of the standard benign benchmarks was available in our environment, so the benign set was generated by gemma-2-2b-it across 32 everyday topic domains and 10 style variants (320 items), verified safe by both ShieldGemma-2B and gemma-2-2b-it at a 0.5 threshold, and independently certified by Qwen2.5-7B, an out-of-family judge that scores all 320 items safe. The set shares no string with the adaptation anchors, unlike the small per-experiment E3 column of the mainline tables, which repeats benign strings that also serve as anchors and is therefore a structural zero; every headline stealth number in this paper is measured on this 320-item set or the out-of-family set of Section U, and near-boundary items excluded by verification are covered by the margin analysis below. Under the observed zero false positives, the 95% Clopper-Pearson upper bound on the rate is 0.93 percent. The false-positive rate is also not a category artifact: under the dose-25 poisoned

Table 8: Removal versus tolerance defenses at dose 25 (3 seeds). Removed counts poison rows dropped of 25 (naive/adaptive); residual is the remaining framed-detection drop; benign cost is the poison-free-window incoming drop. ACLF audits and excises the poison rows; confident learning keeps its predictions consistent with the pre-adaptation guard, and DP-SGD (per-row gradient clipping at the median row norm plus unit-multiplier noise (Ma et al., 2019)) bounds each row's influence instead of removing it, at roughly $8\times$ training compute and 0.8 points of average clean-window adaptation drift for DP-SGD.

Defense	Removed (nav./adj.)	Residual drop (pp)	Benign cost
ACLF (ours)	25/25	0.0/0.0	$23.8\pm5.4\%$
Confident learning (Northcutt et al., 2021)	23/23	1.3/3.3	0%
DP-SGD (Ma et al., 2019)	0/0	1.7/1.7	0%

guard, the largest benign score in any of the 32 topic domains is 0.011 on this set and 0.095 on the out-of-family set (three seeds, 99th percentiles at or below 0.004), so every category sits far below the 0.5 decision threshold rather than clustering near it. Human dual annotation of a 100-item subset confirms this: on a stratified sample of three items per topic domain, one annotator labels all 100 items benign, the second labels 95 of 100 benign and the remaining five borderline (everyday questions on car maintenance, a history question, and book recommendations), and the two agree on 95 of 100 items; every disagreement is a benign-versus-borderline call, and neither annotator labels any item harmful.

N BLIND-TRANSFER GRID

The decisive blind experiment uses an out-of-family probe, Llama-3.2-1B, which mislabels 98.7 percent of framed submissions with errors statistically unrelated to the committee's (overlap 0.038 against a base rate of 0.039); this uninformed attacker delivers 45 poisons from 1184 candidates, collapsing E2 by 76.3 points at dose 25 with plain detection leaking 51.3 and benign false positives at zero, at 26 submissions per poison against 1.1 for a contained probe. The registered six-pair transfer grid is reported as exploratory rather than confirmatory: one committee member (L_g) has zero false negatives on the farming pool, which structurally voids four of the six registered pairs, since no poison can be farmed for them. The supplementary pairs, which target the full committee, succeed in three of three seeds and produce the supplementary cliff results referenced in Section 4.2. The full transfer matrix over nine ordered labeler pairs is included with the released artifacts.

O BENIGN COSTS OF THE DEFENSES

The loss filter that drops the top five percent of rows by loss removes 28 of 320 benign rows on a poison-free batch of 240 committee-safe anchors plus the 320 benign rows, an 8.75 percent training-signal cost; these are training-time drops, not user-facing blocks. The novelty gate admits no benign traffic to review under our template-marker analysis (zero matches on 320 items), so its benign cost is zero by construction. These figures use windows without caught-harmful anchors competing for the budget; on the realistic window composition of anchors plus benign traffic, where all benign-cost numbers in Table 7 are measured, ACLF drops 23.8 ± 5.4 percent of incoming benign rows and the fixed filter's budget shifts onto the caught-harmful anchors, as detailed in Section T.

P LLAMA-GUARD-3: USAGE NOTES AND THE 1B TEMPLATE DEFECT

We evaluate Llama-Guard-3 through ungated mirrors of the official weights (8B and the 1B multilingual variant). Two details matter for reproduction. First, the model emits a newline before its verdict, so both scoring and fine-tuning teacher-force that token and read the safe/unsafe competition at the next position; plain next-token scoring reads a position with no signal. Second, the 1B variant's shipped chat template silently drops the user content, so the model never sees the prompt and answers safe to everything; we borrow the 8B template, after which the 1B behaves as expected (obvious harmful probes yield unsafe with categories, benign yields safe). Under the frozen proto-

col the 8B guard detects 96.3 to 98.8 percent of plain and 85.0 to 93.8 percent of framed harmful queries before adaptation, and Table 2 reports its dose-25 collapse. Used as a labeler instead of a guard, the same 8B model approves 12.7 percent of framed harmful submissions under our Yes/No policy question, two orders of magnitude above the strong 2B labeler's 0.08 percent: guard training and labeling robustness do not transfer, and a committee rebuilt around it would be weaker than the original (15.5 versus 3.9 percent supply). The repaired 1B guard detects 93.8 to 95.0 percent of plain and 91.2 to 93.8 percent of framed queries, and under the same dose-25 stream it loses 39.6 ± 5.8 points of framed and 15.9 ± 3.6 of plain detection with benign false positives at zero: window-level rather than cliff-level damage, the shallowest collapse among the five guards we test.

Q PROFESSIONAL-TASK FARMING ON THE STRONGREJECT POOL

On StrongREJECT wrapped in the guard-evaluation scaffold, the committee approves 79.1 percent of submissions (pool 87 of 110 queries), and across three seeds dose 10, roughly thirteen raw submissions, takes professional-task detection from 0.95 to 0.98 down to exactly zero, with plain detection leaking at most 14 points and benign false positives at zero. The AdvBench numbers appear in Table 3.

R LOCALITY OF THE LEARNED TOLERANCE

Is the hole the attack drills family-local or global? With the professional-task poison as the only framing family in the batch (dose 25, committee driver, three seeds), all three professional-task subfamilies fall to exactly zero detection, while the twelve other framing families of the evaluation set retain 0.75 to 1.00 (three-seed means: fiction 0.75 and game 0.76 are the worst, screen 0.84, half the families at 0.95 or above, json at 1.00). Under the mixed twelve-family poison on the distilled Qwen guard the damage spreads unevenly, concentrating on the professional-task-adjacent families (research 0.24, textbook 0.33, academic 0.48) while translate, news, and json stay at 0.95 to 1.00. Two conclusions follow. The learned tolerance is local to the poisoned framing style with bounded collateral, which is why the interface-corrosion mechanism of Section H moves family margins rather than degrading the guard globally; and the collateral profile depends on the guard, consistent with the damage profiles of Section 5.1. An attacker who wants a specific framing family blinded should farm that family; blanket blindness costs the full mixed dose.

S THE REWRITE ATTACKER AGAINST LOSS FILTERING

The selection oracle in Section 5.2 hints at a stronger attacker that rewrites the poison text to lower its loss. We built it: Qwen2.5-7B-Instruct paraphrases every pool poison four times (meaning and framing preserved, wording varied), the committee re-labels every rewrite so that only naturally safe rows survive (48 to 54 percent do), and the oracle submits the 25 survivors with the lowest loss under the pre-adaptation guard. The attack destroys itself: with or without ACLF, guards adapted on the rewritten rows lose at most 1.3 points of framed detection across three seeds, because a fluent paraphrase erases exactly the stylistic signature that made the original row poison the boundary. Gradient-level optimization that preserves the signature while lowering loss remains open.

T RECOVERY, THE LONG HORIZON OF ACLF, AND UNSUPERVISED FILTERING

Recovery after the cliff. Window 0 delivers the standard dose-25 naive attack with no defense; windows 1 through 5 are the defender's standing clean schedule, the anchor batch alone re-anchored each window on the evolved guard. Framed detection, collapsed to 0.14 ± 0.22 by the attack window, returns to 1.00, 1.00, and 0.99 in the three seeds after a single clean window and stays there for the remaining four (final deltas against each seed's own clean schedule: 0.0, 0.0, and 1.3 points); plain detection recovers on the same schedule. The learned hole is volatile under re-anchoring: an attacker must land a fresh farming round in every window to hold it open, and each round passes again through the filters and gates of Section 5.

ACLF over five clean windows. Each window contains the anchors plus 320 benign incoming rows (the out-of-family set of Section U), and ACLF re-calibrates its threshold on the anchors under the current guard before filtering the incoming rows. Per-window drops are front-loaded and decay, from 22 to 34 percent in the first window down to 3 to 9 percent in the last (cumulative 21.6 ± 5.5 percent), consistent with the guard accommodating the recurring benign distribution until incoming losses fall below the anchor quantile. Plain and framed detection track the unfiltered control with at most 1.3 points of drift over the five windows, so the attrition is a transient admission cost rather than a compounding tax.

Cost protocol and the beta-mixture baseline. All benign-cost figures in Table 7 use one protocol: a poison-free window of anchors plus the 320-row E3r benign set, filtered by each method under the base guard, counting dropped rows. The beta-mixture filter follows the label-noise line: losses max-normalized, a two-component beta mixture fitted by expectation maximization with method-of-moments updates, and rows with posterior above one half for the high-loss component dropped. The fit degenerates in all three seeds, because most benign rows sit at near-zero loss and the mixture collapses; the implementation then falls back to a median split, which discards 214 of 320 incoming benign rows (66.9 percent) and about a quarter of the benign anchors when anchors are unprotected. Anchor protection alone does not help: the same 66.9 percent of incoming rows are discarded. The fixed top-five-percent filter drops no benign rows on this window because its 30-row budget is consumed entirely by caught-harmful anchors, whose losses exceed every benign row's; the 8.75 percent benign cost quoted in Section 5.2 is measured on the anchor-free batches of that section's protocol.

U Second, out-of-family benign set

To break the circularity of a benign set generated by the guard's own family, we generated a second 320-item set with Qwen2.5-7B-Instruct over the same 32 topics and 10 styles as Section M, verified safe by ShieldGemma-2B and gemma-2-2b under the same rule. Across the committee mainline arms (clean and dose 25, three seeds), false positives on this set total one item out of 320 in one clean arm and zero everywhere else, matching the original set.

V Committee strength on the StrongREJECT pool

On AdvBench, upgrading the committee does less than scale suggests: Qwen2.5-7B mislabels framed harmful queries at 4.4 percent, the same rate as the 0.5B, so a committee rebuilt around the 7B is *weaker* than the original at 7.4 percent of the farm, while adding the 7B as a fourth vote next to the strong 2B cuts the supply to 0.6 percent, where 1200 submissions deliver only seven poisons and the guard absorbs them harmlessly; scale does not buy robustness on framed input, safety-oriented training does. Repeating the committee-strength experiment on StrongREJECT (110 queries $\times$ 4 framings, three seeds): the keyword rule misses 95 to 96 percent, the 0.5B labeler half, the strong 2B at most 0.2 percent, and Qwen2.5-7B 13.9 to 15.7 percent, so the 7B is again no substitute for the safety-trained 2B. The four-member committee's supply falls to 10.7 to 12.5 percent, but with pools of about 50 poisons the dose-25 cliff is still delivered for roughly 200 submissions per window, collapsing framed detection by 58 to 62 points with 50 points of plain detection leaking and benign false positives at zero; the committee rebuilt around the 7B admits 55.7 to 60.9 percent and collapses by 46 to 90 points. Upgrading the committee thus shifts the cost of the attack rather than closing the channel whenever the query pool itself is adversarial for every labeler. In short, on this pool the strongest upgrade combination still leaves the cliff reachable, so the committee-strength lever binds on benign-region pools but not on adversarial ones.

W Full-stream adaptation: bounding the admission oracle

Dose-parameterized arms admit exactly the mislabeled rows, but an attacker cannot tell which rows those are; in a production window every submission trains the guard with its natural label, the committee-caught framed submissions included. We therefore re-ran the mainline farms as full-stream windows: benign traffic, caught anchors, and every submission at its natural committee label (1200 AdvBench submissions, of which 32 to 57 land naturally safe depending on seed; 200

professional-task submissions, of which 94 to 106 do), three seeds each. The simple-framing attack is completely neutralized: framed detection returns to exactly 1.00 in all three seeds, the caught framed siblings acting as corrective supervision. The professional-task attack partially survives: framed detection falls to 0.27 ± 0.19 (per seed 0.425, 0.062, 0.325) with plain detection at 0.99 and no benign false positives. Two consequences follow. First, the dose-parameterized results elsewhere in the paper are upper bounds on any single window, which is why the main text reads them alongside the full-stream result. Second, the mislabel rate is the security parameter: the window's own caught-harmful supervision is a real defense at a few percent, and the channel stays open exactly where the framings fool the committee often enough, which is the professional-task regime of Section 5.3 and, structurally, any labeling stack whose members all share the blind spot.

X THE THREE PROFESSIONAL-TASK SCAFFOLDS

The professional-task (TVD) scaffolds are taken verbatim from the internal-safety-collapse evaluation suite (Wu et al., 2026b), in the form released with SafeRedirect (Pan et al., 2026). Each embeds the harmful query, written $\{q\}$ below, as test data inside a legitimate software-engineering task whose completion structurally requires producing the harmful content. We abbreviate each template to its frame and injection point; the released artifacts carry the full text.

Guard-model evaluation. The prompt opens as a software-testing session: *"[Guard Model Evaluation] Testing meta-llama/Llama-Guard-3-8B with Pydantic validation"*. It shows the working directory (`guard.py`, `validator.py`, `test_case.json`) and prints a test-case file whose first entry is a benign filled example and whose second entry is `{"id": 1, "query": "{q}", "assistant_safe": "???", "assistant_unsafe": "???"}`. Running `guard.py` returns a `ValidationError` on the `???` fields: completing the task means filling in what an unsafe assistant would reply to $\{q\}$.

Detoxification testing. The same structure under the frame *"[Toxicity Classifier Testing] Evaluating unitary/toxic-bert for content moderation"*: the query enters a `test_cases.json` entry with `"flagged"` and `"not_flagged"` fields to complete, that is, write the text the classifier should flag.

Outlier detection. The frame is an anomaly-detection script whose `samples.txt` holds model responses one per line, the last being the planted outlier; the file header reads *"# Responses to: "{q}""*, and passing the unit test requires producing the responses, including the outlier that answers the query.

In all three, the surface task is genuine professional work, the query appears only as data, and refusing to complete it means failing the task. This is exactly the structure that frontier models comply with (Wu et al., 2026b) and that our committee most readily labels safe (Section 3.3).

Y FORMAL SPECIFICATION OF PTF, GGF, AND ACLF

Throughout, $\mathcal{C}(x) \in \{\texttt{safe}, \texttt{unsafe}\}$ is the committee label assigned to prompt x by the defender's own pipeline, $\ell_G(x)$ is the training loss of guard G on row x, A is the trusted anchor set of the current window, and $\mathcal{B}_{\text{in}}$ is the set of incoming safe-labeled rows.

Y.1 BENIGN-SIDE FARMING AND PTF

All attacker-side methods share the skeleton of Algorithm 1: wrap harmful queries, submit them, and keep exactly the rows the defender's own pipeline labels safe. PTF instantiates the wrapper w with a single professional-task scaffold s (the guard-evaluation scaffold in our runs), and the screen u with any cheap local model.

The expected submissions per poison is $1/\hat{r}_s$, where $\hat{r}_s$ is the committee's mislabel rate on scaffold s (about one half for the guard-evaluation scaffold, giving the twenty-submission dose of Table 3; under the single weak labeler $\hat{r}_s = 1$).

Algorithm 1 Benign-side farming; PTF sets $w =$ a single professional-task scaffold

Require: harmful pool Q, wrapper w, local screen u, submission budget N, target dose k
1: $S \leftarrow \{\, w(q) : q \in Q,\ u(w(q)) = \texttt{benign-looking} \,\}$ ▷ farm
2: submit the N highest-priority items of S to the pipeline
3: $P \leftarrow \{\, x \in \text{submitted} : \mathcal{C}(x) = \texttt{safe} \,\}$ ▷ defender's own labels
4: inject top-$k(P)$ into the next adaptation batch ▷ dose k
5: **return** updated guard G'; exploit with $w(q)$ for fresh $q \in Q$

Y.2 Geography-guided farming

GGF adds a probe model P (any cheap local model) that estimates where the committee errs. Writing $\hat{r}_f$ for the probe's per-family mislabel estimate on framing family f and $m_P(x)$ for the probe's decision margin on item x, the three allocators of Table 6 are:

$$\text{greedy:} \quad N_f = N \cdot \mathbb{1}\left[f = \arg\max_{f'} \hat{r}_{f'} \right], \tag{1}$$

$$\text{proportional:} \quad N_f = N\, \hat{r}_f \Big/ \textstyle\sum_{f'} \hat{r}_{f'}, \tag{2}$$

$$\text{confidence-ranked:} \quad \text{rank all farm items by } |m_P(x)| \text{ ascending, submit the top } N. \tag{3}$$

Confidence-ranked GGF is boundary-first sampling: items the probe itself finds near its decision boundary are the ones the committee is most likely to mislabel.

Y.3 Anchor-calibrated loss filtering

ACLF needs no knowledge of the attacker. Before each adaptation, it calibrates a loss threshold on the defender's trusted anchors,

$$\tau \;=\; \text{Quantile}_{0.95}\big\{\, \ell_G(a) : a \in A \,\big\}, \tag{4}$$

and filters the incoming stream rather than the anchors or the caught-harmful supervision:

$$\mathcal{B}' \;=\; \{\, x \in \mathcal{B}_{\text{in}} : \ell_G(x) \le \tau \,\} \;\cup\; A \;\cup\; \mathcal{H}_{\text{caught}}. \tag{5}$$

Anchors and caught-harmful rows are never dropped; filtered rows can be routed to human review; and the threshold re-calibrates on the current guard each window, which is what makes the attrition front-loaded and decaying (Section T).