SAFETY RISK OF FINE-TUNING DATA IS NOT A PROPERTY OF THE SAMPLE ALONE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are safety-aligned before deployment so that they refuse harmful requests, and they are then commonly fine-tuned for downstream tasks. However, such fine-tuning can weaken safety alignment even when the training data are entirely benign. Existing defenses filter the training data by estimating the safety risk of each sample from internal model signals, such as hidden representations, sample gradients, or parameter updates. These methods typically score samples once at the initial model state and retain the resulting selection throughout fine-tuning, which implicitly assumes that the initial risk ranking remains informative as the model changes. We show that this assumption does not hold. On a controlled fine-tuning trajectory, risk rankings computed at initialization are nearly uncorrelated with those computed after two epochs. A matched intervention experiment further shows that the same groups of benign samples cause different degrees of safety degradation depending on the model state from which they are learned. Safety risk is therefore a joint property of the sample and the model state rather than an intrinsic property of the sample. We also find that rankings are rewired mainly within the first epoch. At initialization, gradient-based risk scores are dominated by a gradient component nearly orthogonal to the safety direction, and this failure persists under larger adapters, full-parameter gradients, and a short probe epoch. We therefore propose SABA, which re-evaluates sample risk at the current model state in each training round and formulates filtering as the allocation of a fixed exclusion budget across rounds. The resulting concave allocation problem reduces to a geometric schedule controlled by a single deferral parameter: no budget is spent at initialization, and exclusions shift toward later rounds with more reliable risk estimates. Under the same training budget, SABA reduces the attack success rate from 63.3% to 18.3% on CatQA and from 53.1% to 14.4% on HEx-PHI, outperforming nineteen baselines while preserving utility, and the gains carry over to a second corpus and a second model family.

1 INTRODUCTION

As large language models (LLMs) are deployed in an ever wider range of applications, their safety has become a first-order concern. Before deployment, LLMs should be safety-aligned to refuse harmful queries (Ouyang et al., 2022; Bai et al., 2022). However, since LLMs are commonly fine-tuned for downstream tasks, this alignment is not permanent: it can be weakened by fine-tuning on a handful of malicious samples (Yang et al., 2023; Zhan et al., 2024) and even on entirely benign data (Qi et al., 2024).

To preserve safety alignment during fine-tuning, prior work filters training data before fine-tuning begins. These approaches are motivated by evidence that even samples with equally benign content can differ substantially in their effects on model safety (He et al., 2024; Guan et al., 2025). Existing methods therefore estimate sample risk using internal model signals, such as hidden representations, sample gradients, or parameter updates, and exclude high-risk samples from training (He et al., 2024; Li et al., 2025a; Hu et al., 2026; Wang et al., 2026). Although effective, these methods typically score samples only once at the initial model state and retain the resulting data selection throughout fine-tuning. This practice implicitly assumes that the initial risk ranking remains informative at later training states. However, fine-tuning continuously changes the model parameters, potentially altering the safety impact of learning the same sample and making the initial ranking less reliable.

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.

This raises two questions: *(1) Does learning the same benign sample from different model states cause different degrees of safety degradation? (2) If so, when should sample risk be evaluated, and is the initial model state a reliable basis for filtering?*

This paper answers both questions. On a diagnostic trajectory of Qwen3-8B fine-tuned on benign Dolly data, we find that a risk ranking taken at initialization barely transfers to later states: its Spearman correlation with the ranking after two epochs is 0.076, and only a quarter of the samples it flags as most dangerous are still flagged after two epochs, against a fifth expected by chance. A matched intervention experiment further shows that this drift reflects a change in the harm itself rather than in the measurement: the same groups of samples cause different degradation depending on the state they are trained from (Figure 1), with a state–sample interaction of up to $+7.0$ in attack success rate (ASR) when pooled over seeds, whose sign replicates across seeds, judges, benchmarks and model families. The failure is not spread evenly along the trajectory. Rankings are rewired once, inside the first epoch, and are stable afterwards. At the initial state, the scores reproducibly measure a component of the sample gradients that is nearly orthogonal to the safety direction, and this defect survives five attempts at repair, including larger adapters, full-parameter gradients and a short probe epoch.

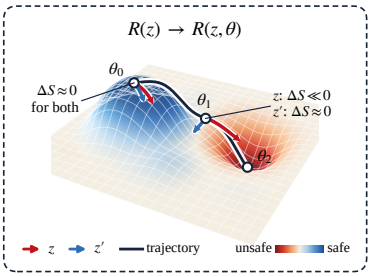

Figure 1: Sample risk depends on the model state: z and z' are equally harmless at θ_0, but only z is harmful at θ_1.

Based on these findings, we propose SABA (**s**tate-**a**ware **b**udget **a**llocation), which re-evaluates sample risk at the current model state in each training round and allocates a fixed exclusion budget across rounds. Under a fixed training budget, we formulate this allocation as a concave optimization problem whose solution shifts exclusions toward later rounds with more reliable risk estimates. Using simple functional forms consistent with our diagnosis, we obtain a geometric schedule controlled by a single deferral parameter δ: no exclusion budget is spent at initialization, and the per-round budget grows geometrically thereafter. Under the same training budget, SABA reduces the ASR from 63.3% to 18.3% on CatQA and from 53.1% to 14.4% on HEx-PHI. These rates are 16.1 and 9.3 below those of the strongest of nineteen baselines, respectively, while downstream task performance is preserved, as measured by held-out loss, ROUGE-L, and ARC-Easy accuracy.

Our contributions are summarized as follows:

- We show that the safety risk of benign fine-tuning samples is state-dependent: in a matched intervention, the same samples cause different safety degradation depending on the model state they are learned from, so a ranking computed once cannot be reused throughout training. We localize the failure to the first epoch and the initial state, and show why it cannot be repaired there.
- We reformulate safety filtering as the allocation of a fixed exclusion budget across training rounds, characterize the optimal schedule under simple functional forms, and reduce it to a one-parameter geometric deferral rule, SABA, which recovers existing static filters as a special case and whose predicted single-peaked shape we confirm on all four judge-benchmark pairs.
- We evaluate SABA with eight risk scorers against nineteen baselines under a matched training budget, with two judges, two benchmarks and three utility measures, and analyze when re-scoring pays off, deriving a check at initialization that decides whether a scorer should defer; it selects the best configuration, or one within noise, in fourteen of eighteen scorer-setting pairs.

2 Background and Related Work

Safety degradation under fine-tuning. Safety alignment does not survive downstream fine-tuning intact: a handful of harmful examples can remove it (Yang et al., 2023; Zhan et al., 2024), and even fine-tuning on entirely benign data weakens it (Qi et al., 2024). One line of defense makes the model or the optimization robust to this drift. Vaccine (Huang et al., 2024b) adds perturbation-aware alignment so that later shifts in the hidden representations do less damage, SafeLoRA (Hsu et al., 2024) projects low-rank updates onto a subspace that preserves safety, and related methods constrain the fine-tuning trajectory in other ways (Huang et al., 2024a; Li et al., 2025b). These methods change how the model is trained and leave the training data unchanged.

Safety-aware data selection. A second line works on the data, since equally benign samples can differ widely in their effect on safety (He et al., 2024; Guan et al., 2025). Representation-based scorers compare a sample's hidden states with those of harmful data (He et al., 2024), restrict the comparison to safety-sensitive layers (Li et al., 2025a), or project onto directions of compliance or of a safety consensus (Zhang et al., 2026; Wu et al., 2026). Gradient-based scorers measure a sample's training impact through normalized self-influence (Guan et al., 2025), gradient similarity to harmful examples (He et al., 2024), or the alignment of its gradient with a safety gradient (Hu et al., 2026), and SQSD (Wang et al., 2026) projects the parameter update a sample induces onto safe and unsafe directions. Other work learns the selection itself, with a bilevel data ranker (Shen et al., 2025) or sample weights adjusted during training (Hu et al., 2025); these methods act on a different axis from the model-signal scorers we study and are complementary to SABA (Section A).

State dependence. Whatever signal they use, these scorers compute a function of the model state, $R(z;\theta)$. Yet they score the pool once, at the initial state or at a fixed reference state θ_s, and apply the resulting ranking throughout fine-tuning. This implicitly assumes that $R(z_i;\theta_s) > R(z_j;\theta_s) \Rightarrow H(z_i;\theta_t) > H(z_j;\theta_t)$ for every later state θ_t, where $H(z;\theta_t)$ denotes the safety degradation actually caused by training on z from state θ_t. There are reasons to doubt this assumption: in data attribution, the influence of a sample is known to change over training, both in stages (Lee et al., 2025) and continuously (Xu & Wu, 2025), and SQSD itself notes that its scores depend on the reference state. What has not been tested is whether this dependence matters for safety: whether the harm a benign sample causes is a property of the sample alone, or also of the state at which it is learned.

3 STATE DEPENDENCE OF SAMPLE-LEVEL SAFETY RISK

The assumption of Section 2 is that a ranking taken at one state remains informative as the model changes. We test it through two questions. **RQ1: Is sample-level safety risk state-dependent? RQ2: If so, which states are suitable for risk measurement?** In short, risk is jointly determined by the sample and the model state (RQ1); rankings are rewired once, within the first epoch; and at initialization, gradient-based scores mostly track signals unrelated to safety (RQ2).

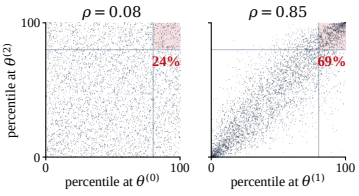

Figure 2: Risk percentiles at two states; shaded: excluded at both.

3.1 SETUP

Diagnostic trajectory. All ranking analyses and interventions in this section use a single diagnostic trajectory, on which Qwen3-8B (Team, 2025) is fine-tuned with LoRA (Hu et al., 2022) for 10 epochs on a fixed pool of 5,000 Dolly (Conover et al., 2023) samples without filtering, so that only the model state changes (Section B). We write $\theta^{(e)}$ for the state after e epochs, with $\theta^{(0)}$ the LoRA initialization. Where an analysis needs the outcome of filtering, we instead use θ_t, the state entering round t of the protocol in Section 5. Rewiring is reproduced in six settings spanning four learning rates, two corpora (Dolly-5k and Alpaca-5k (Taori et al., 2023)) and two model families (Qwen3-8B and Llama-3-8B-Instruct (Team, 2024)), and the primary setting is run with three seeds.

Risk score and harm. Our primary scorer is SQSD (Wang et al., 2026), which projects the update a sample induces onto a dangerous and a safe direction, $R(z;\theta) = D(z;\theta) - S(z;\theta)$. RQ2 also uses GradShield (Hu et al., 2026), the inner product between a sample's gradient and a safety gradient. Harm is measured by the ASR (in %) on CatQA and HEx-PHI, judged by Llama-Guard-3 and the Beaver cost model; differences in ASR are absolute. Because a single sample changes the ASR by less than the evaluation resolves, we estimate H on matched groups.

3.2 RQ1: IS SAMPLE-LEVEL SAFETY RISK STATE-DEPENDENT?

A ranking that changes across states admits two interpretations: the measurement may depend on the state while the safety impact does not, or the safety impact itself may depend on the state from which a sample is learned. Only the second makes risk a joint property of sample and state. We therefore first measure how rankings transfer across states, and then measure harm directly with a matched intervention that separates the two interpretations.

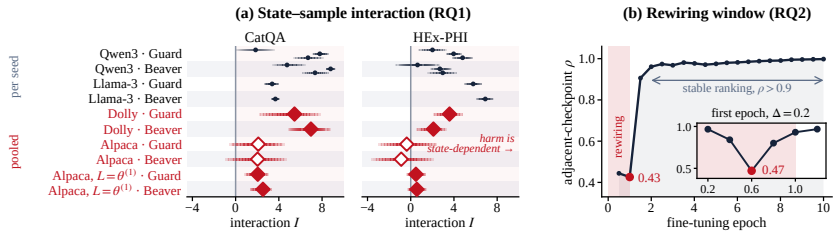

Figure 3: Evidence for state dependence. (a) State–sample interaction I with bootstrap intervals; points are single Dolly seeds, diamonds pool seeds (filled: Dolly, and Alpaca with $L = \theta^{(1)}$; open: Alpaca with $L = \theta^{(2)}$; values in Tables A4 and A5). (b) Rank correlation between consecutive checkpoints at half-epoch spacing (inset: first epoch at 0.2-epoch spacing); shaded: first epoch.

Risk rankings transfer poorly across states. A ranking taken at initialization barely transfers (Figure 2). The Spearman correlation between the SQSD rankings at $\theta^{(0)}$ and $\theta^{(2)}$ is 0.076, and of the 1,000 samples flagged as most risky at $\theta^{(0)}$, only 24.0% are still flagged at $\theta^{(2)}$, barely above the 20% expected by chance. Between $\theta^{(1)}$ and $\theta^{(2)}$, the correlation rises to 0.852, yet 30.8% of the exclusion set still changes. This is not scoring noise: scoring $\theta^{(1)}$ twice gives $\rho = 1.000$ (Table A3).

A matched intervention. To test whether the harm itself depends on the state, we fix two starting states, $E = \theta^{(0.5)}$ early in training and $L = \theta^{(2)}$ later, and construct two groups of samples: EH-L, ranked high-risk at E but low-risk at L, and EL-H, ranked the other way. We train on each group from both states and measure the interaction

$$I = \big[H(\text{EH-L} \mid E) - H(\text{EL-H} \mid E)\big] - \big[H(\text{EH-L} \mid L) - H(\text{EL-H} \mid L)\big]. \tag{1}$$

If the relative harm of the two groups did not depend on the state, I would be zero. The two groups are matched on response length and category, so that they differ only in the state at which they are ranked as risky (Section E.4).

Harm itself depends on the state. On Dolly, the interaction is positive in all sixteen combinations of model, judge, benchmark and seed, ranging from $+0.6$ to $+8.8$, and fifteen of the sixteen bootstrap intervals exclude zero (Figure 3a). Pooled over the three Qwen3-8B seeds, it remains positive for both judges and both benchmarks, from $+2.1$ to $+7.0$ (Table A4). The ranking drift above is therefore not a state-dependent measurement of a fixed property: the safety impact of the same samples changes with the state. This evidence is causal at the level of matched groups of samples, which is also the level at which a filter acts, since it removes sets of samples selected by their ranking. On Alpaca, a pre-registered test that places L at $\theta^{(1)}$, just after the rewiring window, finds a positive interaction on CatQA in every seed and under both judges ($+2.1$ $[0.8, 3.1]$ and $+2.5$ $[1.3, 3.4]$); at the later state $L = \theta^{(2)}$ the pooled CatQA estimate over six seeds keeps the same sign ($+2.1$) with a wider interval that includes zero, and the HEx-PHI estimates are smaller (Section E.2). Its consequences hold on Alpaca as well: rankings are rewired within the first epoch (Section F.2), and static GradShield filtering is less safe than no filtering while the state-aware schedule lowers the ASR by about 24 relative to no filtering (Section G.6). In the regimes we study, the risk that sample-level filters estimate is thus a function $R(z; \theta)$ of both sample and state, and a filtering decision made once need not remain valid as training proceeds.

3.3 RQ2: Which States Are Suitable for Measurement?

We first locate when rankings change, then examine $\theta^{(0)}$, the only state available before fine-tuning.

Rankings are rewired in the first epoch. At half-epoch spacing (Figure 3b), the two checkpoint pairs within the first epoch agree at only $\rho = 0.444$ and 0.426. Agreement then jumps to 0.906 and rises steadily to 0.998. On a finer 0.2-epoch grid (inset), the minimum of 0.471 lies between 0.4 and 0.6 epochs, and it falls within the first epoch in all six settings (Section F.2). Rewiring is thus a one-time event, and a ranking taken after the first epoch remains largely valid.

Little safety signal at initialization. At $\theta^{(0)}$, the gradients of benign samples share a dominant common component. The leading principal component carries 26.5% of the gradient energy, is

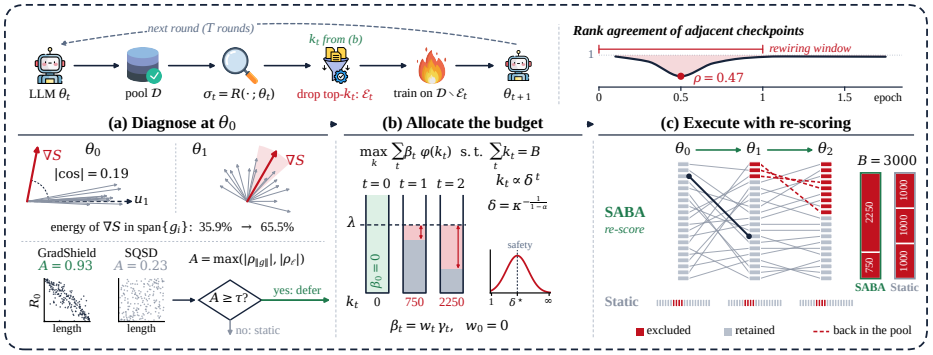

Figure 4: Overview of SABA. **Top:** one round of fine-tuning with re-scoring, and the rewiring of rankings in the first epoch. **(a)** The check A detects an uninformative ranking at θ_0. **(b)** Budget allocation yields $k_0 = 0$ and a geometric schedule. **(c)** SABA re-scores the pool every round; static filtering removes the same samples under the same budget.

nearly parallel to the mean gradient ($\cos = 0.934$), and is almost orthogonal to the safety direction ($|\cos(u_1, \nabla S)| = 0.192$). The span of all 5,000 gradients captures only 35.9% of the energy of the safety direction, against 65.5% after one round of training (Figure A1). The safety signal at $\theta^{(0)}$ is therefore largely absent, not merely masked by the common component.

Two failure modes. This geometry breaks the two families of gradient scorers in different ways. Inner-product scorers inherit the ordering by gradient magnitude: GradShield correlates with gradient norm at $+0.93$, and because gradient norm decreases with response length, its ranking favors short responses ($\rho = -0.86$). This length dependence vanishes by $\theta^{(2)}$ ($+0.09$) and carries no safety signal (Section F.5). Projection scorers such as SQSD cancel most of the common component, but their scores are compressed instead: the standard deviation of R is 0.031 at $\theta^{(0)}$, against 0.336 at $\theta^{(1)}$ and 0.454 at $\theta^{(2)}$, so the ranking at $\theta^{(0)}$ is decided by small residuals in a gradient space that carries little safety signal.

No repair at $\theta^{(0)}$ recovers the later ranking. If the safety signal is absent at $\theta^{(0)}$, no post-processing there should recover it, and none does. Redrawing the initialization, using larger LoRA ranks or full-parameter gradients, projecting out the leading components, and training a 500-sample probe epoch all leave the ranking unrelated to those at later states (Table A6).

RQ1 and RQ2 thus change what a filter must decide: instead of ranking once and deleting permanently, it must decide in every round where to measure risk and how much to exclude (Section 4).

4 Filtering as Budget Allocation

Building on Section 3, we propose SABA (Figure 4), which reformulates safety filtering as the allocation of a fixed exclusion budget across training rounds, re-scores the pool at the current state in every round, and spends no budget at initialization when the initial ranking is uninformative.

4.1 Budget-Constrained Allocation

Let $\mathcal{D}$ be a candidate pool of N samples, and let fine-tuning run for T rounds starting from $\theta_0 = \theta^{(0)}$. In round t, the filter ranks the pool by a risk score $R(\cdot\,; s_t)$ computed at a state $s_t \in \{\theta_0, \ldots, \theta_t\}$, removes the k_t highest-ranked samples, and trains for one epoch on the remaining samples to obtain θ_{t+1}. A filter is thus specified by the schedule $\pi = \{(s_t, k_t)\}_{t=0}^{T-1}$, and its total amount of training is

$$\sum_{t=0}^{T-1} \big(N - k_t\big) \;=\; TN - B, \qquad B = \sum_{t=0}^{T-1} k_t, \tag{2}$$

where B is the exclusion budget. SABA fixes the amount of training, and therefore B, so that excluding more samples in one round requires excluding fewer in another. To compare schedules

with the same budget, we assign each round a value. Let $\varphi(k)$ denote the harm avoided by removing the k most harmful samples of a round under an ideal ranking. Since a ranked filter removes the most harmful samples first, φ is increasing and concave with $\varphi(0) = 0$. The value realized in round t is discounted by two state-dependent factors, which gives the allocation problem

$$\max_{k \geq 0} U(k) \;=\; \sum_{t=0}^{T-1} \beta_t \, \varphi(k_t) \qquad \text{s.t.} \qquad \sum_{t=0}^{T-1} k_t = B, \qquad \beta_t = w_t \gamma_t, \tag{3}$$

where $w_t \in [0, 1]$ is the fidelity, the degree to which the ranking at s_t orders samples by the harm they cause from θ_t onward, and $\gamma_t \in (0, 1]$ is the persistence, the fraction of the benefit that survives later rounds. Fidelity varies with the state (Section 3); persistence is below one because excluded samples return to the pool.

4.2 Geometric Deferral Schedule

Equation (3) is a concave program over a simplex, and its maximizer takes the standard water-filling form (Proposition 2 in Section C): a round with $\beta_t = 0$ receives no budget however large B is, and $k_t^\star$ is nondecreasing in β_t. Under a fixed training budget, the model state thus enters the allocation only through the coefficients β_t. To obtain a deployable schedule, SABA adopts the simplest form of each factor that is consistent with Section 3:

$$\varphi(k) = k^\alpha, \qquad \gamma_t = \kappa^{\,T-1-t}, \qquad w_0 \in \{0, w\}, \qquad w_t = w \;\; (t \geq 1), \tag{4}$$

where $\alpha \in (0, 1)$ controls how quickly returns diminish, $\kappa \in (0, 1]$ is the fraction of an exclusion's benefit that survives one further round, and $w > 0$ is the common fidelity of trained states. A power law makes the schedule independent of the pool size, and geometric decay means each further round undoes a fixed fraction of the remaining benefit, with $\kappa = 1$ meaning permanent exclusion. A single w suffices because rankings at trained states agree closely (Section 3.3), and it cancels because it scales every funded round equally; only whether $w_0 = 0$ matters, which holds for a scorer whose initial ranking is dominated by the common gradient component. Substituting Equation (4) into the water-filling solution gives $k_t^\star \propto \beta_t^{1/(1-\alpha)}$ on funded rounds, and hence the following schedule.

Proposition 1 (Geometric deferral)**.** *Under Equation* (4) *with* $w_0 = 0$*, the optimal schedule is* $k_0^\star = 0$ *and*

$$k_t^\star \;=\; B \, \frac{\delta^{\,t}}{\sum_{u=1}^{T-1} \delta^{\,u}}, \qquad t \geq 1, \qquad \textit{where} \qquad \delta \;=\; \kappa^{-1/(1-\alpha)} \;\geq\; 1. \tag{5}$$

The deferral ratio δ is the only free parameter. It ranges from an even split over the trained rounds ($\delta = 1$) to spending the whole budget in the final round ($\delta \to \infty$), and the uniform schedule $k_t = B/T$ of existing filters is recovered with $w_0 = w$ and $\kappa = 1$. Proposition 1 also makes a prediction that does not depend on the exact values of α and κ: safety should be single-peaked in δ, with its optimum strictly inside $(1, \infty)$. The prediction fails if the curve is monotone or peaks at $\delta \leq 1$; we test it in Section 5.3.

4.3 State-Aware Scoring and Initialization Check

Proposition 1 leaves two decisions open: the state at which each round scores the pool, and whether $w_0 = 0$. We write $R_t = R(\cdot\,; \theta_t)$ for the scores of the pool at θ_t.

Scoring at the current state. Since no single state orders the pool well for all later states, SABA scores the whole pool at the current state in every round, $s_t = \theta_t$. Exclusion is therefore a per-round decision: a sample withheld in one round is re-scored in the next and may be trained on.

Initialization check. To decide whether $w_0 = 0$, SABA measures at θ_0 the symptom identified in Section 3.3:

$$A \;=\; \max\big(\, |\rho(R_0, \|g\|)|,\; |\rho(R_0, \text{len})| \,\big), \tag{6}$$

where $\rho(\cdot, \cdot)$ is the Spearman rank correlation over the pool, $\|g\|$ is the norm of the sample gradient, and len is the response length. If a gradient-based scorer has $A \geq \tau = 0.3$, its initial ranking is dominated by the common component; SABA then sets $k_0 = 0$ and allocates the budget by

Proposition 1, which we call the SABA schedule. Representation-based scorers are exempt, because the adapter is zero at θ_0 and their scores are properties of the aligned model. All other scorers keep their original static usage, $k_t = B/T$ on the initial ranking, with one refinement: if the ranking changes at the first trained state, $B_r = \rho(R_0, R_1) < \tau$, SABA uses the ranking at θ_1 for the remaining rounds. For the eight scorers of Section 5.1, the assignments are the same for any $\tau \in (0.226, 0.356]$ (Section G.5).

In every configuration, SABA excludes B samples in total and therefore trains on $TN - B$ sample-passes; it reduces to static filtering when both checks fail and costs one additional scoring pass in each round after the first (Algorithm 1 in Section C).

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTING

Protocol. Every arm fine-tunes Qwen3-8B with LoRA on the same $N = 5{,}000$ Dolly samples for $T = 3$ one-epoch rounds with budget $B = 3{,}000$, and thus trains on exactly $12{,}000$ sample-passes (Equation (2)), so no arm can appear safer merely by training less. Baselines exclude $k = (1000, 1000, 1000)$; for a scorer that passes the initialization check, SABA excludes $k = (0, 750, 2250)$, i.e. $\delta = 3$, fixed in advance rather than tuned on the benchmarks. Our main scorer GradShield passes the check ($A = 0.93$). Refresh-only and memory-ratchet score the full pool at the same three states as SABA, so the comparison with them is also compute-matched.

Scorers and baselines. We pair SABA with eight scorers: the gradient-based GradShield (Hu et al., 2026), SQSD (Wang et al., 2026), He-grad (He et al., 2024) and Self-Inf-N (Guan et al., 2025); the representation-based He-rep (He et al., 2024), LARF (Li et al., 2025a) and DS-CSS (Zhang et al., 2026); and SCOPE, a control that keeps the form of GradShield with a different safety direction (definitions in Section B). The nineteen baselines in Table 1 are no filtering and three groups of filters. Static baselines score the pool once at $\theta^{(0)}$ with each of the eight scorers, and with SQSD at the reference checkpoint of its official pipeline. Uninformative baselines exclude random samples, the shortest or the longest responses. Dynamic baselines re-score during training under the uniform budget, each with GradShield and SQSD: refresh-only re-scores in every round, delayed-static re-scores once at θ_1, and memory-ratchet additionally penalizes samples ranked dangerous earlier.

Metrics. Safety is the ASR of the final adapter on CatQA (550 prompts; Bhardwaj et al., 2024) and HEx-PHI (290 prompts; Qi et al., 2024) under greedy decoding, judged separately by Llama-Guard-3-8B (Team, 2024) (Guard) and the Beaver cost model (Dai et al., 2024) (Beaver). Utility is held-out negative log-likelihood (NLL, nats per token) and ROUGE-L on Dolly, and zero-shot ARC-Easy accuracy. All numbers are mean $\pm$ standard deviation over three seeds.

5.2 COMPARISON WITH BASELINES

Table 1 reports the main comparison. Without filtering, fine-tuning on benign data raises the ASR of Qwen3-8B to 63.3% on CatQA and 53.1% on HEx-PHI under Llama-Guard-3. SABA lowers these to 18.3% and 14.4%; the reduction is between 39 and 45 on every judge-benchmark pair, while its utility stays within the range of the static baselines (NLL 1.676, against 1.671 to 1.683).

Static filtering leaves most of the risk in place. Scored once at $\theta^{(0)}$, the eight scorers reach between 39.3% and 67.2% on CatQA, where He-grad and Self-Inf-N even end above the unfiltered model. The best static scorer on CatQA is DS-CSS, which remains 21 above SABA there and 17 on HEx-PHI, with a standard deviation across seeds six times ours (9.3 against 1.5); on HEx-PHI every static arm is at least 15 above SABA. Scoring SQSD at its official reference state instead of $\theta^{(0)}$ does not close this gap. Static GradShield is indistinguishable from no filtering (62.4% against 63.3%), as Section 3.3 predicts, yet the same scorer under SABA is the best arm in the table.

Dynamic baselines close part of the gap. Every dynamic baseline built on GradShield lowers the CatQA ASR of static GradShield by 15 to 28, confirming the diagnosis of Section 3 independently of our schedule. The strongest of them is refresh-only, which remains 16.1 above SABA on CatQA and 9.3 on HEx-PHI under Guard. The remaining gap comes from the allocation alone: refresh-only re-scores at the current state, but it spends a third of the budget at $\theta^{(0)}$ and spreads the rest evenly.

Table 1: Safety (ASR in %) and utility under an identical training budget (12,000 sample-passes, three seeds). Baselines exclude $k = (1000, 1000, 1000)$ per round and SABA $k = (0, 750, 2250)$. **Bold**: best; underline: second best; red: less safe than no filtering.

	CatQA ↓		HEx-PHI ↓		NLL ↓	ROUGE-L ↑	ARC-E ↑
Method	Guard	Beaver	Guard	Beaver			
No filtering							
no filter (all 5,000 samples)	63.3±0.7	73.7±1.7	53.1±1.9	58.0±1.6	1.667	0.334	0.813
Static filtering, scored once							
GradShield	62.4±2.5	72.7±1.9	43.9±2.4	48.7±2.0	1.671	0.322	0.808
He-grad	63.8±6.8	74.1±6.7	46.4±4.5	51.3±3.5	1.671	0.329	0.811
Self-Inf-N	67.2±3.4	76.8±3.9	50.6±1.7	53.8±2.5	1.672	0.334	0.813
SQSD	44.8±3.9	58.4±3.6	36.0±3.5	42.0±2.6	1.675	0.325	0.814
DS-CSS	39.3±9.3	51.7±8.5	31.8±8.3	37.0±7.4	1.680	0.340	0.818
LARF	46.5±6.5	59.0±5.0	34.3±0.5	40.8±0.9	1.678	0.334	0.820
He-rep	48.4±5.6	60.6±6.0	30.2±3.8	38.0±4.9	1.673	0.329	0.815
SCOPE	45.6±1.3	58.2±1.3	34.9±3.9	39.8±3.3	1.683	0.333	0.818
SQSD at official θ_{ref}	46.4±7.0	58.3±6.3	29.7±4.6	34.7±5.5	1.673	0.327	0.814
Uninformative exclusion							
random	61.9±3.2	72.7±2.0	46.2±3.6	51.6±4.1	1.673	0.333	0.814
shortest responses	68.3±4.4	79.2±4.9	50.1±4.8	54.3±4.7	1.672	0.315	0.807
longest responses	54.1±5.8	68.2±4.9	39.9±2.9	45.4±2.9	1.685	0.341	0.818
Dynamic filtering, re-scored during training							
delayed static · GradShield	47.4±3.1	58.4±3.3	34.0±1.8	39.4±1.9	1.672	0.341	0.812
delayed static · SQSD	42.9±1.0	53.5±1.0	30.1±1.1	35.1±1.1	1.674	0.326	0.814
memory ratchet · GradShield	41.1±2.7	51.8±2.2	27.5±3.8	33.1±5.2	1.671	0.330	0.812
memory ratchet · SQSD	43.8±5.0	55.9±4.5	29.7±4.0	35.1±4.7	1.674	0.329	0.813
refresh only · GradShield	34.4±5.9	43.8±5.1	23.7±1.1	28.9±1.2	1.672	0.331	0.813
refresh only · SQSD	47.0±4.1	59.9±4.5	33.2±6.2	39.3±6.8	1.676	0.324	0.814
SABA (ours) · GradShield	**18.3**±1.5	**30.2**±2.0	**14.4**±1.1	**19.2**±1.2	1.676	0.330	0.812
gain over the best baseline	▼ 16.1	▼ 13.5	▼ 9.3	▼ 9.7			

The effect is not a length effect. Excluding the shortest responses raises the ASR above no filtering (68.3% on CatQA), excluding the longest lowers it only modestly (54.1%), and random exclusion has no effect (61.9%). SABA therefore does not work by removing short answers, which is effectively what the $\theta^{(0)}$ GradShield ranking does; instead, it removes samples ranked at trained states, where the length dependence has disappeared.

5.3 ABLATION STUDIES

Table 2: Component ablation across budgets B (percentage: pool fraction excluded per round), GradShield, Llama-Guard-3, final ASR in %, three seeds; paired difference to the full method in parentheses, red if it hurts and green if it helps by more than 1.

	$B = 1500$ (10%)		$B = 3000$ (20%)		$B = 4500$ (30%)		$B = 6000$ (40%)	
Variant	CatQA	HEx-PHI	CatQA	HEx-PHI	CatQA	HEx-PHI	CatQA	HEx-PHI
SABA (full)	**40.3**	**27.2**	**18.3**	**14.4**	**12.4**	**10.0**	**11.5**	**10.5**
− skip: $k_0 > 0$, uniform k	44.7 (+4.4)	34.6 (+7.4)	34.4 (+16.1)	23.7 (+9.3)	29.4 (+17.0)	19.7 (+9.7)	25.6 (+14.2)	14.9 (+4.5)
− defer: $\delta = 1$	39.0 (−1.3)	29.7 (+2.4)	28.5 (+10.2)	20.3 (+6.0)	22.2 (+9.8)	16.3 (+6.3)	16.2 (+4.8)	12.6 (+2.2)
− return: permanent exclusion	37.9 (−2.4)	27.2 (+0.0)	20.9 (+2.6)	16.4 (+2.1)	12.1 (−0.2)	10.3 (+0.3)	11.2 (−0.2)	12.0 (+1.5)
− re-score: $s_t \equiv \theta^{(0)}$	67.0 (+26.7)	51.6 (+24.4)	69.1 (+50.8)	54.5 (+40.1)	67.7 (+55.3)	54.5 (+44.5)	68.3 (+56.8)	54.1 (+43.7)

Components. Table 2 removes each of the four decisions of SABA in turn, at four budgets. Re-scoring matters most: if the $\theta^{(0)}$ ranking is frozen under the deferred schedule (− re-score), the ASR returns to about the unfiltered level or above it at every budget. Spending budget at initialization (− skip) costs 16.1 at $B = 3000$, the empirical counterpart of $w_0 = 0$. Spreading the remaining budget evenly (− defer) costs 10.2 and hurts at every budget from 3000 upward. Making exclusions permanent (− return) costs 2.6 at $B = 3000$ and at most 1.5 at larger budgets. The ordering is the same under the Beaver judge (Table A9).

Deferral ratio. Figure 5a tests the prediction of Section 4.2 with seven schedules, from $\delta = 0$ (all exclusions in the second round) to $\delta = \infty$ (all in the final round). On all four judge-benchmark

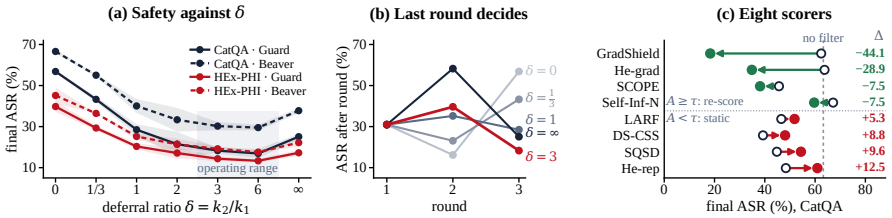

Figure 5: The schedule, three seeds. (a) Final ASR against the deferral ratio $\delta = k_2/k_1$ at $k_0 = 0$, $B = 3000$. (b) ASR after each round. (c) Static filtering (open) and the SABA schedule (filled) for eight scorers; Δ is the paired change. (b, c): Llama-Guard-3 on CatQA.

pairs the curve is single-peaked, with its optimum inside the range, and the value $\delta = 3$ that we fixed before the sweep lies within 1.6 of the best schedule in Figure 5a. On CatQA under Guard, the ASR is 28.5% at $\delta = 1$, 18.3% at $\delta = 3$, 16.9% at $\delta = 6$ and 25.1% at $\delta = \infty$. Both extremes fail because withheld samples return to the pool (Figure 5b): budget spent in the second round protects that round's model but not the final one, and budget spent only in the final round cannot undo an unfiltered second round.

Budget, rounds and cost. Safety saturates beyond $B = 4500$ while held-out NLL keeps rising. At a fixed training volume, fewer rounds are better, and deferral still helps at $T = 5$ (Figure A3). Re-scoring costs one extra scoring pass in each round after the first, and every cheaper approximation we tried loses safety, although a single-pass approximation at $0.37\times$ the re-scoring time still improves on the best baseline on three of the four judge-benchmark pairs (Table A11).

5.4 When Does Re-Scoring Pay Off?

Applied to all eight scorers without the check, the SABA schedule improves exactly the four that the first check flags (Figure 5c; both benchmarks in Figure A2): on CatQA, GradShield by 44.1, He-grad by 28.9, and Self-Inf-N and SCOPE by 7.5 each. LARF, DS-CSS, SQSD and He-rep have informative initial rankings and are better kept static, since the schedule would raise their ASR by 5.3 to 12.5. These regressions come from deferring, not from re-scoring: relative to a control that defers each scorer's own frozen $\theta^{(0)}$ ranking, re-scoring changes the ASR by only $+1.0$ for DS-CSS and -0.1 for SQSD, against -5.3 to -50.8 for the four scorers that benefit (Table A12).

The checks of Section 4.3 separate the two groups (dotted line in Figure 5c). The four scorers that benefit are gradient functionals with A between 0.42 and 0.96, whose single-direction inner product the common component reduces to essentially the gradient norm; SQSD cancels this component ($A = 0.23$), and representation scores are properties of the aligned model that training does not overturn. Among the scorers that fail the first check, DS-CSS, LARF and He-rep keep their ranking ($B_r \geq 0.67$) and stay static, whereas SQSD ($B_r = 0.14$) switches to its θ_1 ranking and improves on static filtering on all four judge-benchmark pairs. For every scorer run in more than one configuration on Dolly, the checks select the best configuration or one within noise (Table A13), and again for GradShield on Alpaca; the two misses, SQSD on Alpaca and GradShield on Llama-3, have A in the middle range (0.50 and 0.27), yet their selected branch still lowers the CatQA ASR by 6.6 and 6.3 relative to no filtering (Section G.6). In a pre-registered Llama-3 test (A and predicted branch recorded before training), they select the best branch, or one within 2, for five of seven further scorers, including all three sent to the schedule (Section D).

6 Conclusion

The safety risk of benign fine-tuning samples is not a property of the samples alone but depends on the state at which they are learned, and gradient-based scores carry little safety signal at initialization. SABA turns this diagnosis into a budget allocation that re-scores at the current state and defers exclusions, lowering the ASR by 39 to 45 under an identical training budget with utility within the range of the baselines, and its gains carry over to a second corpus and model family.

REPRODUCIBILITY STATEMENT

Section 3.1 and Section B specify the diagnostic trajectory, the base models, the LoRA configuration and every risk scorer, and Section E.4 describes the matching procedure behind the intervention. Section 4 and Algorithm 1 in Section C define SABA, including both checks and the fixed threshold. Section 5.1 gives the filtering protocol, the nineteen baselines and the evaluation metrics, and all results are reported as mean and standard deviation over three seeds. The code, the per-sample risk scores at every state, the frozen matched sets and the per-seed evaluation results will be released upon publication.

ETHICS STATEMENT

This work aims to preserve the safety alignment of language models when they are fine-tuned on benign data. The evaluation uses the public CatQA and HEx-PHI benchmarks, which contain harmful prompts, and the responses of the fine-tuned models are used only for automatic judgment by Llama-Guard-3 and the Beaver cost model. The per-sample risk scores we will release identify benign samples that degrade safety, and could in principle be used to select data that weakens alignment; they are, however, the same quantities that existing filters already compute, and we release them to support the development of defenses. No human subjects were involved in this study.

AI USE STATEMENT

Large language model assistants were used for grammar checking and sentence-level editing of the manuscript. An LLM-based coding assistant was also used to write, edit, and execute experiment and analysis code based on specifications provided by the authors. Generative AI was not used for research ideation, methodology or experiment design, interpretation of results, generation of synthetic datasets, or mathematical proofs. All AI-assisted text and code were reviewed, tested, and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 14138–14149. Association for Computational Linguistics, 2024.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm`, 2023. Accessed: 2023-06-30.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. In Aarti Singh, Maryam Fazel,

Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety. In *First Conference on Language Modeling*, 2024.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe lora: The silver lining of reducing safety risks when finetuning large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Zhanhao Hu, Xiao Huang, Patrick Mendoza, Emad A. Alghamdi, Basel Alomair, Raluca Ada Popa, and David A. Wagner. Gradshield: Alignment preserving finetuning. *CoRR*, abs/2605.14194, 2026.

Zixuan Hu, Li Shen, Zhenyi Wang, Yongxian Wei, and Dacheng Tao. Adaptive defense against harmful fine-tuning for large language models via bayesian data scheduler. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim F. Tekin, and Ling Liu. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 31983–32016. Association for Computational Linguistics, 2025.

Jin Hwa Lee, Matthew Smith, Maxwell Adam, and Jesse Hoogland. Influence dynamics and stagewise data attribution. *CoRR*, abs/2510.12071, 2025.

Hao Li, Lijun Li, Zhenghao Lu, Xianyi Wei, Rui Li, Jing Shao, and Lei Sha. Layer-aware representation filtering: Purifying finetuning data to preserve LLM safety alignment. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 8030–8050. Association for Computational Linguistics, 2025a.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. Salora: Safety-alignment preserved low-rank adaptation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

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! In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. SEAL: safety-enhanced aligned LLM fine-tuning via bilevel data selection. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

Xiao Wang, Yifei Zhang, YongKang Liu, Xiaocui Yang, Zihan Wang, Shi Feng, and Daling Wang. From parameter dynamics to risk scoring: Quantifying sample-level safety degradation in LLM fine-tuning. In *International Conference on Machine Learning (ICML 2026)*, 2026.

Zefeng Wu, Weiwei Qi, Jielong Chen, Tianhang Zheng, Di Hong, Chaochao Lu, Liang He, Zhan Qin, and Kui Ren. Datashield: Uncovering risky fine-tuning data across llms through consensus subspace alignment. *CoRR*, abs/2607.15081, 2026.

Jie Xu and Zihan Wu. Dynamic influence tracker: Measuring time-varying sample influence during training. *CoRR*, abs/2502.10793, 2025.

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

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing RLHF protections in GPT-4 via fine-tuning. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 681–687. Association for Computational Linguistics, 2024.

Junbo Zhang, Qianli Zhou, Xinyang Deng, Wen Jiang, Jie Pan, and Jinbiao Zhu. Datashield: Safety-degrading data filtering for LLM benign instruction fine-tuning. *CoRR*, abs/2606.00160, 2026.

A Extended Related Work

Safety degradation during fine-tuning. Safety alignment established for large language models does not reliably survive fine-tuning. Even when the fine-tuning data contain no explicit malicious instructions, standard downstream adaptation can weaken the model's safety behavior (Qi et al.,

2024). One response is to make the model, or the optimization, robust to later fine-tuning. Vaccine (Huang et al., 2024b) uses perturbation-aware alignment so that the aligned model resists the representation shifts caused by harmful fine-tuning, and SafeLoRA (Hsu et al., 2024) restricts low-rank updates to a subspace that better preserves safety. These and related methods (Huang et al., 2024a; Li et al., 2025b) reduce safety degradation by constraining the optimization trajectory.

Later work showed that, even when all training data are benign, individual samples differ widely in how much they degrade safety (He et al., 2024; Guan et al., 2025). The harmfulness of a sample's content is therefore not the same as its effect on safety during training: a sample whose text is entirely benign can still weaken alignment through the update it induces. This observation motivates a second line of work, which identifies and removes such samples before or during fine-tuning.

Safety-aware data selection and risk estimation. Most existing methods estimate the safety risk of a benign sample from signals inside the model. He et al. (2024) proposed two such measures, one based on representation matching and one on gradient matching. Later representation-based methods restrict the comparison to specific layers or subspaces: LARF (Li et al., 2025a) locates safety-sensitive layers, and the DataShield methods (Zhang et al., 2026; Wu et al., 2026) project hidden states onto a compliance direction or a safety-consensus subspace. Gradient-based methods instead measure how a sample affects the update. Self-Inf-N (Guan et al., 2025) flags samples with unusually high normalized self-influence, and GradShield (Hu et al., 2026) compares a sample's fine-tuning gradient with the gradient of a safety objective. Other methods learn the selection itself: SEAL (Shen et al., 2025) trains a bilevel data ranker, and BDS (Hu et al., 2025) adjusts sample weights during training. SQSD (Wang et al., 2026) starts from the observation that, as fine-tuning proceeds, the parameters drift along directions associated with safety degradation, and scores a sample by projecting the update it induces onto a safe and an unsafe direction. Apart from BDS, which reweights samples during training, these methods compute one ranking and apply it throughout fine-tuning, which is the assumption examined in Section 3.

B Implementation Details

Fine-tuning. All arms use LoRA (Hu et al., 2022) with rank $r = 8$, scaling factor 16 and no dropout, applied to the query, key, value, output, gate, up and down projections of every layer. Training uses LLaMA-Factory with AdamW at a constant learning rate of 5×10^{-6} after a 3% warm-up, an effective batch size of 8 (per-device batch 1, gradient accumulation 8), a cut-off length of $2{,}048$ tokens and bf16 precision, with the Qwen3 chat template and thinking disabled. One epoch over the $5{,}000$-sample pool is 625 optimizer steps. In the filtering protocol, the adapter is carried over between rounds, and each round trains for one epoch on the retained samples. In the two- and five-round arms of Figure A3c, the epochs per round are scaled to 1.71 and 0.55 so that the total number of sample-passes stays at $12{,}000$. The seed controls only the LoRA initialization and the data order.

Diagnostic trajectory. A separate trajectory is needed because the filtering protocol changes the model state and the training subset at the same time in every round, so the two effects are confounded. The diagnostic trajectory keeps the data fixed and lets only the state vary. It uses the same configuration for 10 epochs without filtering and saves a checkpoint every 313 steps; $\theta^{(1)}$ and $\theta^{(2)}$ are checkpoints 626 and $1{,}252$ (epochs 1.002 and 2.003). $\theta^{(0)}$ is the LoRA initialization itself (Kaiming-initialized down-projection and zero up-projection, seed 0), not the first saved checkpoint. The half-epoch grid of Figure 3b is this checkpoint grid. The 0.2-epoch grid of its inset comes from a rerun of the first epoch with the same seed and configuration that saves a checkpoint every 125 steps (Section F.2).

Scoring. All scorers score the frozen $5{,}000$-sample pool at the indicated state, with dropout disabled and a maximum sequence length of $1{,}024$ tokens. For 45 samples the prompt alone exceeds this length, so their induced update is identically zero at every state; Section E.1 describes how they are handled. The SQSD directions D and S are estimated with the released code; on Llama-3 the two projections are nearly collinear (Section D).

Risk scorers. GradShield is the inner product between a sample's fine-tuning gradient and the gradient of a refusal margin. SQSD is the difference $D - S$ between the projections of the sample-induced update onto the dangerous and the safe direction. He-grad is the cosine between the sample gradient and the mean gradient of harmful anchors, and Self-Inf-N is length-normalized self-influence.

Among the representation-based scorers, He-rep is the similarity of the final-token representation to harmful and safe anchors, LARF is the cosine similarity to unsafe and safe anchors at a safety-sensitive layer, and DS-CSS is the projection of hidden states onto a compliance direction. SCOPE keeps the first-order form of GradShield but replaces its refusal-margin probe with a preference-based safety functional. From PKU-SafeRLHF (Ji et al., 2025) we take 64 prompts, each with one safe and one unsafe response, define $J(\theta) = \mathrm{mean}_k \,\mathrm{softplus}\big(\log p_\theta(\text{unsafe}_k) - \log p_\theta(\text{safe}_k)\big)$, and score a sample by $-\langle g_i, \nabla J\rangle$. We validated the seven published implementations against their released code.

Evaluation. Responses to the 550 CatQA and 290 HEx-PHI prompts are generated greedily from the final adapter and judged by Llama-Guard-3-8B and by the Beaver cost model, each applied to the full prompt-response pair. All differences reported in the text are computed from unrounded means. Utility is measured as negative log-likelihood and ROUGE-L on a held-out split of Dolly, and general ability as zero-shot multiple-choice accuracy on the 2,376 questions of the ARC-Easy test set, scored by log-likelihood with length normalization (`acc_norm`).

C Allocation Details

Proposition 2 states the water-filling solution of Equation (3) used in Section 4.2, and Algorithm 1 summarizes the full procedure of SABA.

Proposition 2 (Allocation)**.** *Let φ be strictly increasing, strictly concave and differentiable on $(0, N]$, and let $\beta_t > 0$ for at least one t. Then $k_t^\star = 0$ whenever $\beta_t = 0$, and for some $\lambda > 0$*

$$k_t^\star \;=\; \begin{cases} (\varphi')^{-1}\big(\lambda/\beta_t\big), & \beta_t\,\varphi'(0^+) > \lambda, \\ 0, & \textit{otherwise}, \end{cases} \qquad \sum_t k_t^\star = B, \tag{7}$$

so that $k_t^\star$ is nondecreasing in β_t.

Proposition 2 implies that a round whose ranking carries no information receives no budget, however large B is.

Algorithm 1 SABA: state-aware budget allocation

Require: pool $\mathcal{D}$ of N samples; rounds T; budget B; risk score R; deferral ratio $\delta \geq 1$; threshold $\tau = 0.3$
Ensure: fine-tuned model θ_T

$\sigma \leftarrow R(\cdot\,;\theta_0)$ over $\mathcal{D}$; compute A by Equation (6); $\theta \leftarrow \theta_0$
if R is gradient-based **and** $A \geq \tau$ **then** ▷ initial ranking uninformative: $w_0 = 0$
  $m \leftarrow$ DEFER; $k_0 \leftarrow 0$, $k_t \leftarrow B\,\delta^{\,t} / \sum_{u=1}^{T-1} \delta^{\,u}$ for $t \geq 1$ ▷ Equation (5), rounded to sum to B
else
  $m \leftarrow$ STATIC; $k_t \leftarrow B/T$ for all t
end if
for $t = 0$ **to** $T-1$ **do**
  if $m =$ DEFER **and** $t \geq 1$ **then**
    $\sigma \leftarrow R(\cdot\,;\theta)$ ▷ re-score the whole pool at the current state
  else if $m =$ STATIC **and** $t = 1$ **then**
    $\sigma' \leftarrow R(\cdot\,;\theta)$; **if** $\rho(\sigma, \sigma') < \tau$ **then** $\sigma \leftarrow \sigma'$ ▷ second check
  end if
  $\mathcal{E}_t \leftarrow$ the k_t samples of $\mathcal{D}$ with largest σ
  $\theta \leftarrow \textsc{TrainOneEpoch}\big(\theta,\ \mathcal{D} \setminus \mathcal{E}_t\big)$ ▷ $\mathcal{E}_t$ returns to the pool
end for
return θ

D Results on Llama-3-8B-Instruct

This appendix collects every Llama-3-8B-Instruct run. The protocol is that of Section 5.1 with the base model replaced: the same 5,000 Dolly samples, $T = 3$, $B = 3{,}000$, three seeds unless marked,

and Llama-3's own θ_0 adapter and θ_0 scores. Two scorer implementations differ from Qwen3-8B. The SQSD directions D and S are estimated with the released code, and on Llama-3 the per-sample projections onto them are nearly collinear (Pearson correlation 0.96 over the pool), so $D - S$ is degenerate and SQSD uses D alone; the official-reference-state baseline depends on both and is not run. LARF's safety-sensitive layer and DS-CSS's compliance direction are re-selected on Llama-3 with each method's own procedure.

Diagnostics. The state–sample interaction of Section 3 is reproduced on Llama-3 in Table A4 (single seed), and the rewiring window in Table A7. The θ_0 gradient spectrum is in Table A8: the common component is present ($|\cos(u_1, \bar{g})| = 0.84$) and even more nearly orthogonal to the safety direction than on Qwen3 ($|\cos(u_1, \nabla S)| = 0.02$). However, GradShield's score follows it much less closely ($\rho(R_0, \|g\|) = 0.29$ on count-sketched gradients). Its θ_0 ranking is reproducible across fresh LoRA initializations (mean $\rho = 0.98$), and its probe direction is nearly orthogonal to the mean gradient ($\cos = 0.08$, against 0.26 on Qwen3), which is why A is low on this model.

Pre-registered test of the two checks.

Table A1: Pre-registered test of the two checks on Llama-3-8B-Instruct: A and the predicted branch of each scorer were recorded before any of these arms was trained, with $\tau = 0.3$ fixed on Qwen3-8B. The last columns give the final ASR under Llama-Guard-3 (CatQA / HEx-PHI) of every branch run, three seeds; bold marks the selected branch, and ✓ a selected branch within 2 of the best on both benchmarks. GradShield is reported in Table A15, since its arms predate the test.

						CatQA / HEx-PHI, Guard, by branch			
Scorer	form	A	B_r	check 1	selected	static	delayed static	schedule	
SQSD	gradient	0.61	n/a	$k_0 = 0$	schedule	74.0 / 54.0	n/a	**63.6 / 43.1**	✓
He-grad	gradient	0.20	0.24	no	delayed static	72.5 / 50.7	**64.2 / 45.2**	59.8 / 39.2	×
Self-Inf-N	gradient	0.36	n/a	$k_0 = 0$	schedule	75.0 / 58.9	n/a	**74.4 / 60.0**	✓
SCOPE	gradient	0.60	n/a	$k_0 = 0$	schedule	70.8 / 55.4	n/a	**64.4 / 46.9**	✓
He-rep	representation	0.30	0.89	no	static	**69.7 / 51.3**	69.2 / 52.4	71.1 / 55.3	✓
LARF	representation	0.84	0.97	no	static	**73.5 / 50.8**	73.8 / 49.8	73.4 / 47.7	×
DS-CSS	representation	0.74	0.90	no	static	**72.1 / 47.1**	74.8 / 47.6	78.5 / 56.2	✓

Table A1 tests the two checks on a base model that played no part in setting them. For each of the seven scorers other than GradShield, we computed A and recorded the predicted branch before any of these arms was trained, with $\tau = 0.3$ as fixed on Qwen3-8B; B_r was then measured at each static arm's θ_1, as the second check prescribes, and every branch was run with three seeds. The selected branch is the best one, or within 2 of it on both benchmarks, for five of the seven scorers. All three scorers that the first check sends to the schedule are assigned correctly. The schedule lowers the ASR of SQSD from 74.0% to 63.6% on CatQA and from 54.0% to 43.1% on HEx-PHI, and that of SCOPE by 6.5 and 8.5, whereas for DS-CSS, which the checks keep static, it would have raised the ASR by 6.4 and 9.1. The two misses are He-grad and LARF. He-grad's A of 0.20 falls below the threshold, and its three random-adapter draws (0.15 to 0.37) straddle it; on Qwen3 the draws vary as much (0.36 to 0.47), although all of them pass (Section G.5). It therefore takes the static branch, where the second check fires ($B_r = 0.24$), and the resulting delayed-static arm improves on static filtering by 8.2 and 5.5 but remains 4.5 and 6.0 above the schedule. LARF is representation-based and keeps its static ranking as the rule prescribes ($B_r = 0.97$); its three branches agree on CatQA to within 0.5, and the schedule is 3.1 lower on HEx-PHI.

Full comparison.

Table A2 repeats the main comparison. Llama-3 is less safe than Qwen3 after benign fine-tuning (73.7% on CatQA and 61.3% on HEx-PHI without filtering), and static filtering barely changes its CatQA ASR: the eight scorers reach between 69.7% and 75.0%, and GradShield, Self-Inf-N and SQSD end above the unfiltered model. Excluding random, short or long responses leaves the CatQA ASR within 1.3 of no filtering. Re-scoring during training is what helps on this model. The schedule with GradShield is the safest arm on all four judge-benchmark pairs, 4.7 and 0.9 below the best baseline under Guard (SQSD re-scored once at θ_1), and the checks select the schedule with SQSD, which reaches 63.6% and 43.1%. For GradShield, whose initial ranking the first check does not flag ($A = 0.27$), the checks select delayed-static filtering (67.4% and 47.8%), which improves on static filtering by 7.5 and 12.1 but gives up part of the gain of the schedule (Section G.6).

Table A2: The comparison of Table 1 on Llama-3-8B-Instruct, same protocol and budget; three seeds, ASR in %. SQSD uses D alone on this model, and its official reference state is not run. The highlighted rows are SABA: the schedule with GradShield, GradShield as selected by the two checks, and SQSD as selected. †: mean over fewer than three seeds; n/a: not evaluated. Formatting as in Table 1.

Method	CatQA ↓		HEx-PHI ↓		NLL ↓	ROUGE-L ↑	ARC-E ↑
	Guard	Beaver	Guard	Beaver			
No filtering							
no filter (all 5,000 samples)	73.7±1.4	84.7±0.9	61.3±2.4	69.5±2.6	1.662	0.327	0.815
Static filtering, scored once							
GradShield	74.8±1.2	87.2±1.5	59.9±3.3	70.2±3.5	1.665	0.327	0.810
He-grad	72.5±1.4	83.8±2.4	50.7±2.8	61.5±3.5	1.668	0.327	0.826
Self-Inf-N	75.0±0.3	86.4±0.9	58.9±2.2	70.0±1.8	1.667	0.340	0.825
SQSD (D only)	74.0±2.7	86.8±1.5	54.0±2.4	64.4±1.8	1.674	0.334	0.828
DS-CSS	72.1±1.3	83.6±1.9	47.1±0.9	55.6±1.7	1.678	0.335	0.830
LARF	73.5±2.2	85.8±2.0	50.8±4.4	61.7±4.5	1.678	0.333	0.828
He-rep	69.7±1.5	81.2±0.4	51.3±0.4	62.2±1.1	1.665	0.339	0.827
SCOPE	70.8±0.5	83.6±1.4	55.4±0.7	67.2±0.3	1.669	0.332	0.824
Uninformative exclusion							
random	73.0±3.7	85.2±2.8	56.9±4.6	65.7±3.0	1.667	0.327	0.826
shortest responses	72.4±1.9	84.0±2.1	60.0±2.5	69.9±2.2	1.665	0.328	0.824
longest responses	74.1±0.3	87.1±0.9	53.9±2.9	66.6±0.3	1.681	0.351	0.827
Dynamic filtering, re-scored during training							
delayed static · SQSD	63.9±2.8	75.8†	42.1±2.4	51.8±2.3	1.671	0.329	0.829
memory ratchet · GradShield	70.7±1.0	n/a	49.1±2.0	n/a	1.666	0.334	0.826
memory ratchet · SQSD	65.9±1.9	n/a	43.6±1.7	n/a	1.672	0.327	0.828
refresh only · GradShield	67.2±1.3	79.5†	47.1±0.2	57.8±1.0	1.666	0.340	0.825
refresh only · SQSD	65.3±2.0	n/a	43.9±0.9	n/a	1.671	0.324	0.827
SABA · GradShield, schedule	**59.2±0.9**	**70.9±0.7**	**41.1±0.9**	**50.8±0.9**	1.669	0.342	0.811
gain over the best baseline	▼4.7	▼4.9	▼0.9	▼1.0			
SABA · GradShield, as selected (delayed static)	67.4±1.2	79.4±1.0	47.8±0.5	58.5±1.4	1.666	0.343	0.812
SABA · SQSD, as selected (schedule)	63.6±0.3	76.1±2.3	43.1±0.7	53.2±1.6	1.672	0.324	0.826

E SUPPORTING MATERIAL FOR RQ1

This appendix gives the details of the ranking-transfer and intervention results of Section 3.2: the handling of the samples whose scores are identically zero, the matching pipeline behind the 2×2 intervention with its balance audit, and the full table of interaction values.

E.1 SAMPLES WITH IDENTICALLY ZERO SCORES

For 45 of the 5,000 samples, the prompt alone exceeds the scoring pipeline's maximum length of 1,024 tokens. Every response token is then truncated and all labels are masked, so the induced update and every gradient-based score derived from it are identically zero at every state. These samples will be listed by index in the released artifact. We keep them in Table A3, where they tie at $R = 0$ and count as tied pairs, so that the pool of that table matches the pool of every filtering arm. Their effect on the table is negligible. Over the remaining 4,955 samples, the $\theta^{(0)} \to \theta^{(1)}$ agreement is $\rho = 0.100$ instead of 0.104, the top-20% overlap is 23.7% instead of 24.0%, and the flip rate is 0.466 instead of 0.465, while the number of tied pairs falls from 993 to 3. We remove them from the dispersion statistics of Figure A1b, where 45 identical values would distort the standard deviation and the share of $D > S$, and from the seed-stability numbers of Table A6. Excluding them changes A by at most 0.03 for any scorer and changes no assignment of the first check.

E.2 SCOPE OF THE INTERVENTION

This section describes the scope of the intervention in four respects. First, the intervention also detects state-dependent harm on Alpaca (Table A5). In a test whose direction was recorded before it was run, we keep E, the matching and the replicates of Section 3 and place L at $\theta^{(1)}$, the first checkpoint after the rewiring window. The interaction is then positive on CatQA in all three seeds and under both judges (+2.05 [0.83, 3.08] and +2.53 [1.31, 3.36]), and positive but smaller on HEx-PHI (+0.5 and +0.6, intervals including zero). At the later state $L = \theta^{(2)}$ used on Dolly, the pooled CatQA interaction over six seeds keeps the same sign under both judges (+2.1 and +2.0, positive

in five and four of the six seeds), while between-seed variance widens its nested intervals to include zero; on HEx-PHI it is close to zero. The simple effect at E is close to zero on both corpora, so the corpora differ only in what happens at L. The unfiltered trajectories lose safety at similar rates on the two corpora (CatQA ASR of 41% to 46% on Alpaca and 44% to 48% on Dolly at $\theta^{(2)}$, on a 110-prompt subset), so the smaller effect at $\theta^{(2)}$ on Alpaca is not a matter of how far safety has degraded; on this corpus the harm of the two groups converges later in training, and the state dependence is clearest just after the rewiring window.

Second, the intervention compares two states on opposite sides of the rewiring window. It therefore establishes that the safety impact differs between these states, but does not imply that harm changes continuously along the whole trajectory. In particular, $E = \theta^{(0.5)}$ lies near the minimum of the rewiring window, whereas $L = \theta^{(2)}$ lies well after it. This reading is consistent with the high agreement between $\theta^{(1)}$ and $\theta^{(2)}$ in Table A3.

Third, the effect is identified at the group level. The safety effect of a single sample is below the resolution of our evaluation, which is why the intervention is carried out on matched groups. Finally, robustness across random seeds is established only for Qwen3-8B, since the Llama-3 intervention uses a single seed.

E.3 Rank-transfer statistics

Table A3 gives the agreement statistics behind Section 3.2 and Figure 2.

Table A3: Agreement between SQSD risk rankings of the 5,000-sample pool at different states. Overlap is the fraction of the top-20% exclusion set shared by the two states; flip rate is the fraction of sample pairs whose order differs. The last row scores the same state twice.

State pair	Spearman ρ	Kendall τ	Top-20% overlap	Flip rate
$\theta^{(0)} \to \theta^{(1)}$	0.104	0.070	24.0%	0.465
$\theta^{(0)} \to \theta^{(2)}$	0.076	0.051	24.0%	0.475
$\theta^{(1)} \to \theta^{(2)}$	0.852	0.669	69.2%	0.166
$\theta^{(1)}$, scored twice	1.000	0.999	99.9%	0.0004
chance	0.000	0.000	20.0%	0.500

E.4 Matching procedure and covariate balance

Why matching is needed. Migration groups defined directly on the raw scores are severely imbalanced in length. On the Qwen3-8B trajectory, samples ranked high-risk at E and low-risk at L (EH-L) average 103.2 response words, against 38.9 for the converse group EL-H (Mann-Whitney $p = 5.6 \times 10^{-8}$), and instruction length and character count are also imbalanced at $p < 10^{-4}$. Left uncorrected, "risky early" would largely mean "long answer". The imbalance remains even though SQSD normalizes each module to counter the known bias of unnormalized updates toward short responses, and it is specific to the migration groups.

Pipeline. Scores at E and L are first residualized by ordinary least squares on $\log$(response length) and $\log$(instruction length), and the migration groups are redefined on the ranks of the residual scores: EH-L is the upper residual tail at E intersected with the lower tail at L, and EL-H is the converse. The two tails are then matched by caliper nearest-neighbor matching on the log covariates, with a caliper of 0.7 standard deviations. A small integer program enforces an exact quota on the open- versus closed-domain instruction category, and a mean-balance caliper of 0.1 standard deviations is applied to the matched sets. Repeating the matching with different random restarts gives the replicates of Table A4. Each replicate trains every cell of the 2×2 design on its own matched sets, and all matched sets were frozen and written to disk before any outcome was evaluated. On the Qwen3-8B trajectory, the residualized tails contain 33 (EH-L) and 39 (EL-H) candidates, and the matched members of the twelve replicates cover 31 and 23 distinct samples. The balance table below is computed on this union, and Table A4 gives the mean matched-set size per replicate.

Balance.

Covariate	EH-L ($n = 31$)	EL-H ($n = 23$)	Mann-Whitney p
Response length (words)	61.4	65.8	1.000
Instruction length (words)	22.9	39.2	0.430
Tokens	84.3	105.0	0.073
Risk at E (separated by design)	1.187	0.843	2.2×10^{-9}

After matching, the groups are indistinguishable in response length, while the separation in early risk that the design requires is retained. Learning difficulty is ruled out as a confounder in the same way. Risk correlates with the sample's training loss at both states ($+0.49$ at E and $+0.46$ at L), but the change in risk between the states is only weakly related to the change in loss ($r = -0.376$, about 14% of the variance), and the matched groups are balanced on loss at both states ($p = 0.75$ and 0.71). Update magnitude cannot drive the effect either: the cosine-based score is scale invariant, and the correlation between risk and update norm changes sign between the two states (-0.43 at E, $+0.16$ at L).

Per-replicate audit. We compute standardized mean differences for every replicate. On the Qwen3-8B trajectory, the groups separate on the scoring axis as intended (standardized difference $+2.08$ at E and -4.87 at L), and the mean length difference is -0.14 (largest 0.40); on Llama-3-8B-Instruct the corresponding values are $+3.77$, -4.17 and -0.19 (largest 0.33). The instruction category has too many levels for a fixed threshold to be meaningful at $n \approx 15$, so we compare the total-variation distance between the groups with an exchangeable permutation null. The observed distances (0.31 on Qwen3 and 0.33 on Llama-3) sit at or below the mean of the null (0.31 and 0.37; 95th percentiles 0.46 and 0.57), and no replicate exceeds the 95th percentile. A residual length imbalance above 0.25 standardized units remains in 1 of 12 Qwen3 replicates and 5 of 12 Llama-3 replicates. Three facts limit its influence. First, where it remains, the EH-L responses are the longer ones, yet EH-L is the group that does less harm at L, so the imbalance works against the effect it would have to explain. Second, any length effect that is constant across states cancels in the difference of differences, so only a length effect that itself interacts with the state could bias I. Third, the interaction replicates on the Qwen3 trajectory, where eleven of the twelve replicates are balanced. On Alpaca, where the groups are drawn from a 500-sample sub-pool, the largest per-replicate standardized difference on the matching covariates is 0.15 over all six seeds at $L = \theta^{(2)}$ and 0.19 over the three seeds at $L = \theta^{(1)}$, well below the 0.25 used above.

E.5 Full interaction values

Missing combinations are not padded: only one Llama-3 seed was trained, and the Alpaca setting was run on Qwen3-8B only. Pooling is always done within a cell of corpus, model and judge, and never across judges, whose absolute rates are not comparable. The pooled intervals are nested bootstraps that resample seeds and then replicates within each resampled seed, weighting seeds equally. Flat intervals over the pooled replicates are narrower; we report the more conservative nested intervals throughout. A one-way analysis of variance over seeds rejects a common magnitude in three of the four Dolly cells and in all four Alpaca cells at $L = \theta^{(2)}$; on Dolly the sign is nevertheless the same in every seed, and we pool seeds with a nested bootstrap whose intervals account for this variation.

Table A5 gives every Alpaca run. At $L = \theta^{(1)}$ all three seeds agree in sign on CatQA and the pooled interval excludes zero under both judges; at $L = \theta^{(2)}$ the per-seed CatQA estimates range from -3.8 to $+6.1$ across both judges, so the between-seed variance dominates the interval. The matched sets at $\theta^{(1)}$ are smaller (10 to 15 per group against 11 to 22), because fewer samples change rank between $\theta^{(0.5)}$ and $\theta^{(1)}$ than between $\theta^{(0.5)}$ and $\theta^{(2)}$.

F Supporting Material for RQ2

This appendix supports Section 3.3: the gradient geometry at initialization and the five attempted repairs, the rewiring window in the other five settings, the spectral analysis behind Figure A1c and its extension to the settings of Section G.6, the probe experiment, and the length correlations that separate the harm from the gradient norm.

Table A4: State–sample interaction I on Dolly for every model, judge and seed; brackets give the 2.5th and 97.5th percentiles of 20,000 bootstrap resamples over replicates. n is the mean number of matched samples per group per replicate, and Rep. the number of replicates. Pooled rows combine the three seeds of one cell by a nested bootstrap. Seed 0 of Qwen3-8B is the diagnostic trajectory of Section 3; the Alpaca rows are in Table A5.

Model	Judge	Seed	Benchmark	I	n / Rep.
Dolly-5k, $\eta = 5 \times 10^{-6}$					
Qwen3-8B	Guard	0	CatQA	$+6.70$ [$+5.32, +8.14$]	22 / 12
Qwen3-8B	Guard	0	HEx-PHI	$+4.80$ [$+3.82, +5.78$]	22 / 12
Qwen3-8B	Guard	1	CatQA	$+7.79$ [$+7.12, +8.61$]	15.7 / 6
Qwen3-8B	Guard	1	HEx-PHI	$+3.97$ [$+3.28, +4.71$]	15.7 / 6
Qwen3-8B	Guard	2	CatQA	$+1.85$ [$+0.06, +3.61$]	16.5 / 6
Qwen3-8B	Guard	2	HEx-PHI	$+2.01$ [$+0.69, +3.33$]	16.5 / 6
Qwen3-8B	Beaver	0	CatQA	$+7.35$ [$+6.14, +8.67$]	22 / 12
Qwen3-8B	Beaver	0	HEx-PHI	$+2.96$ [$+1.55, +4.31$]	22 / 12
Qwen3-8B	Beaver	1	CatQA	$+8.79$ [$+8.36, +9.21$]	15.7 / 6
Qwen3-8B	Beaver	1	HEx-PHI	$+2.70$ [$+1.84, +3.79$]	15.7 / 6
Qwen3-8B	Beaver	2	CatQA	$+4.76$ [$+3.42, +6.48$]	16.5 / 6
Qwen3-8B	Beaver	2	HEx-PHI	$+0.63$ [$-1.38, +2.64$]	16.5 / 6
Llama-3	Guard	0	CatQA	$+3.38$ [$+2.64, +4.03$]	14 / 12
Llama-3	Guard	0	HEx-PHI	$+5.78$ [$+4.94, +6.61$]	14 / 12
Llama-3	Beaver	0	CatQA	$+3.68$ [$+3.32, +4.05$]	14 / 12
Llama-3	Beaver	0	HEx-PHI	$+6.90$ [$+6.06, +7.67$]	14 / 12
Pooled over three seeds, nested bootstrap					
Dolly	Guard	all	CatQA	$+5.45$ [$+2.15, +7.85$]	n/a
Dolly	Guard	all	HEx-PHI	$+3.59$ [$+2.05, +4.87$]	n/a
Dolly	Beaver	all	CatQA	$+6.97$ [$+4.89, +8.75$]	n/a
Dolly	Beaver	all	HEx-PHI	$+2.10$ [$+0.49, +3.34$]	n/a

F.1 Gradient geometry at initialization

Figure A1 shows the common gradient component at $\theta^{(0)}$ and its two symptoms discussed in Section 3.3.

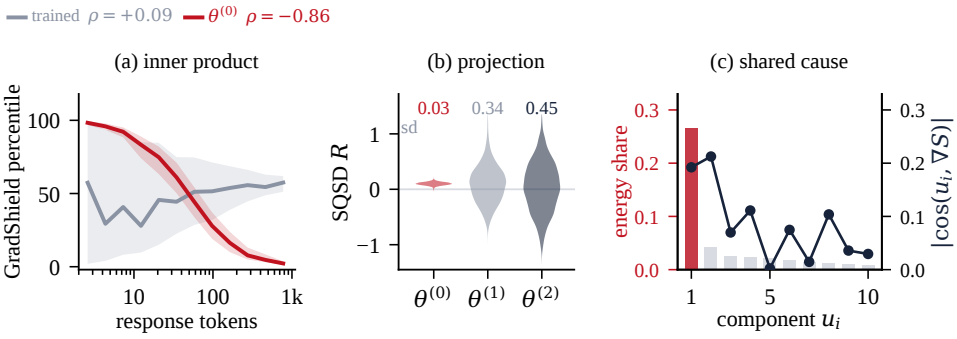

Figure A1: The common gradient component at $\theta^{(0)}$ and its consequences. (a) GradShield percentile against response length, shown by binned median and interquartile range, at $\theta^{(0)}$ and $\theta^{(2)}$. (b) Distribution of the SQSD score R at three states, with standard deviations. (c) Energy share of the ten leading principal components of the sample gradients at $\theta^{(0)}$ (bars) and their absolute cosine similarity with the safety direction (points).

The five repairs. If the safety direction is largely absent from the sample gradients at $\theta^{(0)}$, no post-processing of those gradients should recover it, and the failure should survive every repair that stays at $\theta^{(0)}$. Table A6 tests five such repairs, and none succeeds. Scoring noise is ruled out, because rerunning the scorer reproduces the ranking ($\rho = 1.000$), and redrawing the Kaiming initialization

Table A5: The Alpaca-5k intervention on Qwen3-8B at two late states, $E = \theta^{(0.5)}$ throughout; brackets as in Table A4, six replicates per seed. $L = \theta^{(2)}$ is the design of Section 3, run on six seeds; $L = \theta^{(1)}$ was run on seeds 0 to 2 after the first three $\theta^{(2)}$ seeds, with the direction of the effect recorded before training. n is the mean matched-set size per group. Pooled rows are nested bootstraps over three and six seeds respectively.

Seed	Judge	Benchmark	$L = \theta^{(1)}$	$L = \theta^{(2)}$	n at $\theta^{(1)}$ / $\theta^{(2)}$
0	Guard	CatQA	$+2.85$ [$+1.64, +3.73$]	-3.18 [$-4.91, -1.33$]	10 / 16
0	Guard	HEx-PHI	$+0.23$ [$-0.29, +0.86$]	-3.85 [$-6.15, -2.07$]	10 / 16
0	Beaver	CatQA	$+3.24$ [$+2.91, +3.58$]	-3.82 [$-5.18, -2.00$]	10 / 16
0	Beaver	HEx-PHI	$+0.63$ [$-0.63, +1.67$]	-4.89 [$-6.55, -2.82$]	10 / 16
1	Guard	CatQA	$+0.76$ [$-0.12, +1.64$]	$+3.39$ [$+1.00, +5.79$]	15 / 11
1	Guard	HEx-PHI	-0.23 [$-1.26, +0.75$]	-2.13 [$-5.52, +1.38$]	15 / 11
1	Beaver	CatQA	$+1.21$ [$+0.30, +2.30$]	$+4.64$ [$+1.52, +7.85$]	15 / 11
1	Beaver	HEx-PHI	$+0.98$ [$-0.46, +2.24$]	-2.47 [$-5.98, +1.09$]	15 / 11
2	Guard	CatQA	$+2.55$ [$+1.94, +3.30$]	$+6.12$ [$+4.85, +7.39$]	14 / 20
2	Guard	HEx-PHI	$+1.44$ [$+0.63, +2.07$]	$+4.77$ [$+3.10, +6.44$]	14 / 20
2	Beaver	CatQA	$+3.15$ [$+2.76, +3.61$]	$+5.91$ [$+4.73, +7.12$]	14 / 20
2	Beaver	HEx-PHI	$+0.11$ [$-1.15, +1.49$]	$+4.77$ [$+3.22, +6.09$]	14 / 20
3	Guard	CatQA	n/a	$+4.70$ [$+4.00, +5.36$]	n/a / 20
3	Guard	HEx-PHI	n/a	$+3.22$ [$+2.41, +3.91$]	n/a / 20
3	Beaver	CatQA	n/a	$+4.76$ [$+4.21, +5.27$]	n/a / 20
3	Beaver	HEx-PHI	n/a	$+2.87$ [$+2.18, +3.45$]	n/a / 20
4	Guard	CatQA	n/a	$+0.27$ [$-1.55, +1.82$]	n/a / 16
4	Guard	HEx-PHI	n/a	-0.92 [$-2.64, +0.80$]	n/a / 16
4	Beaver	CatQA	n/a	-0.03 [$-1.27, +1.03$]	n/a / 16
4	Beaver	HEx-PHI	n/a	-2.41 [$-4.14, -0.57$]	n/a / 16
5	Guard	CatQA	n/a	$+1.27$ [$-1.00, +3.48$]	n/a / 22
5	Guard	HEx-PHI	n/a	-3.33 [$-4.77, -1.78$]	n/a / 22
5	Beaver	CatQA	n/a	$+0.73$ [$-2.09, +3.18$]	n/a / 22
5	Beaver	HEx-PHI	n/a	-3.16 [$-4.60, -1.49$]	n/a / 22
pooled	Guard	CatQA	$+2.05$ [$+0.83, +3.08$]	$+2.10$ [$-0.56, +4.51$]	3 / 6 seeds
pooled	Guard	HEx-PHI	$+0.48$ [$-0.40, +1.42$]	-0.37 [$-2.96, +2.45$]	3 / 6 seeds
pooled	Beaver	CatQA	$+2.53$ [$+1.31, +3.36$]	$+2.03$ [$-0.92, +4.72$]	3 / 6 seeds
pooled	Beaver	HEx-PHI	$+0.57$ [$-0.31, +1.44$]	-0.88 [$-3.61, +2.13$]	3 / 6 seeds

gives $\rho = 0.9995$ for GradShield and 0.641 for SQSD across seeds, against 0.104 across states. A larger adapter does not help: at ranks 32 and 128 the ranking is unchanged ($\rho \geq 0.9997$) and the length correlation stays at -0.86. Leaving the LoRA subspace does not help either, since full-parameter gradients give $\rho = 0.998$ against the LoRA ranking, and static filtering on them reaches an ASR of 62.1 ± 2.9, level with 63.3 for no filtering. Projecting out the leading principal components removes the symptoms (the correlation with gradient norm falls to 0.01 and with length to 0.02), but the residual ranking remains unrelated to later states ($\rho \in [-0.03, +0.08]$ for all $k \leq 10$). Finally, a small amount of training is not enough: after a 500-sample probe the ranking still agrees with $\theta^{(0)}$ at $\rho = 0.97$, and static filtering on it reaches an ASR of 64.4. The failure at $\theta^{(0)}$ therefore reflects missing information rather than a correctable bias.

F.2 The rewiring window across settings

Table A7 repeats the adjacent-checkpoint analysis of Figure 3b in the six settings of Section 3.1. In every setting, the least correlated adjacent pair is the one within the first epoch, and the correlation reaches above 0.99 by the end of training. At the largest learning rate (5×10^{-5}) the minimum still falls in the first epoch, although the drop is shallower (0.838, against 0.862 for the next-lowest pair); in the other five settings it is pronounced. Because the checkpoint grid is 313 steps, these trajectories locate the minimum only to within half an epoch, which is why we ran the dense grid in the inset of Figure 3b.

Table A6: Five candidate repairs of the failure at $\theta^{(0)}$, none of which succeeds. Correlations are Spearman over the pool, and ASR is on CatQA (no filter: 63.3). Row 1 covers GradShield and SQSD; the other rows use GradShield, whose variants can be recomputed from cached gradients.

If the failure were…	we would see it go away when…	but instead
scoring noise	the scorer is rerun, or the initialization is redrawn	$\rho = 1.000$ on a rerun; $\rho = 0.9995$ (GradShield) and 0.641 (SQSD) across three Kaiming seeds, against $\rho = 0.104$ across states
an artifact of the rank-8 update	the LoRA rank is increased	$\rho \geq 0.9997$ for $r = 32$ and $r = 128$; the length correlation stays at -0.86
confinement to the LoRA subspace	scores are recomputed from full-parameter gradients	$\rho = 0.998$ against the LoRA ranking; static filtering reaches ASR 62.1 ± 2.9, matching no-filter
a dominant common component	the leading principal components are projected out	the symptoms vanish (ρ with $\|g\|$ falls to 0.01 and with length to $+0.02$) yet the residual ranking stays unrelated to later states, $\rho \in [-0.03, +0.08]$ for all $k \leq 10$
the absence of any training signal	a short probe epoch is run before scoring	scores after a 500-sample probe still match $\theta^{(0)}$ at $\rho = 0.97$; static filtering reaches ASR 64.4
The safety direction is not reachable from the sample gradients at $\theta^{(0)}$: they span only 35.9% of its energy, so no number of principal components captures half. After one round of the filtering protocol this rises to 65.5%, with half reached at $k \approx 1165$.		

Table A7: Spearman correlation between SQSD rankings at consecutive checkpoints, half an epoch apart, in six settings. These trajectories cache only the dangerous-direction projection D on a fixed 500-sample subpool, so the values are not comparable in level with Figure 3b, which uses $R = D - S$ on the full pool; the location of the minimum is what the table establishes. The first saved checkpoint is at 0.5 epochs, so the pair $\theta^{(0)} \to \theta^{(0.5)}$ is not available here.

Setting	$0.5 \to 1$	$1 \to 1.5$	$1.5 \to 2$	$2 \to 2.5$	$9.5 \to 10$	min. pair
Dolly, Qwen3, $\eta = 5 \times 10^{-6}$ (primary)	**0.275**	0.939	0.972	0.981	0.998	$0.5 \to 1$
Dolly, Qwen3, $\eta = 3 \times 10^{-6}$	**0.057**	0.902	0.970	0.984	0.998	$0.5 \to 1$
Dolly, Qwen3, $\eta = 2 \times 10^{-6}$	**0.064**	0.775	0.958	0.984	0.998	$0.5 \to 1$
Dolly, Qwen3, $\eta = 5 \times 10^{-5}$	**0.838**	0.888	0.897	0.862	0.993	$0.5 \to 1$
Alpaca, Qwen3, $\eta = 5 \times 10^{-6}$	**0.313**	0.942	0.977	0.982	0.999	$0.5 \to 1$
Dolly, Llama-3, $\eta = 5 \times 10^{-6}$	**0.517**	0.895	0.969	0.985	0.999	$0.5 \to 1$

The dense grid. The 0.2-epoch grid in the inset of Figure 3b is a rerun of the first 1.2 epochs of the primary trajectory with the same seed, data order and hyperparameters, saving a checkpoint every 125 steps. The rerun is not bit-identical to the original trajectory, but its state at step 625 agrees with the original checkpoint 626 at a Spearman correlation of 0.998 between the two SQSD rankings of the pool and a top-20% overlap of 97.7%, so the two grids describe the same trajectory to the resolution of the analysis. The inset reports full-pool values; removing the 45 zero-score samples changes them by at most 0.011 (to 0.966, 0.837, 0.460, 0.802, 0.932 and 0.968).

F.3 Spectral analysis

Table A8 extends Figure A1c to the two further settings. The common component is present in all three: the leading component carries 15% to 27% of the gradient energy, is nearly parallel to the mean gradient ($|\cos| \geq 0.84$), tracks the gradient norm ($\rho \geq 0.85$), and is nearly orthogonal to the safety direction ($|\cos| \leq 0.19$, and 0.02 on Llama-3). The settings differ in how closely the scorer follows this component. GradShield's association with gradient norm is 0.93 on Dolly, 0.74 on Alpaca and 0.29 on Llama-3 (on count-sketched gradients; the check itself gives $A = 0.27$), so the

Table A8: The leading principal component u_1 of the sample-gradient matrix at θ_0 in the three settings of Section G.6 (count-sketched GradShield gradients, 5,000 samples). $\bar{g}$ is the mean gradient and ∇S the safety direction of SQSD; the last two columns are GradShield's association with gradient norm and length on the same sketched gradients, whose maximum is A up to sketching error (0.29 against $A = 0.27$ on Llama-3, Section D).

Setting	$\lambda_1/\sum\lambda$	top-10 energy	$\lvert\cos(u_1,\bar{g})\rvert$	$\rho(\lvert\langle g,u_1\rangle\rvert,\lVert g\rVert)$	$\lvert\cos(u_1,\nabla S)\rvert$	∇S energy in top-10	$\rho(R_0,\lVert g\rVert)$	$\rho(R_0,\text{len})$
Dolly-5k · Qwen3-8B	0.265	0.440	0.934	+0.95	0.192	0.118	+0.93	-0.81
Alpaca-5k · Qwen3-8B	0.212	0.384	0.895	+0.89	0.159	0.124	+0.74	-0.55
Dolly-5k · Llama-3-8B-Instruct	0.149	0.337	0.837	+0.85	0.016	0.053	+0.29	-0.16

first check fires in the first two settings but not in the third, where the scorer inherits the component only weakly and the second check takes over (Section G.6).

F.4 The probe experiment

The last row of Table A6 asks whether the failure at $\theta^{(0)}$ is merely the absence of any training signal, in which case a small amount of training should already repair it. Starting from $\theta^{(0)}$, we train for one epoch on 500 samples of the pool, score the whole pool with GradShield at the resulting state, and run static filtering with that ranking, $k = (1000, 1000, 1000)$, under the protocol of Section 5.1 with three seeds. The probe does not move the ranking: its agreement with the $\theta^{(0)}$ ranking is $\rho = 0.971$ to 0.974 across seeds, the correlation with response length remains at -0.86, and the 1,000-sample exclusion set shares 852 to 866 members with the $\theta^{(0)}$ set. Accordingly the final ASR is $64.4 \pm 0.4\%$ on CatQA and $47.5 \pm 4.4\%$ on HEx-PHI under Llama-Guard-3, level with static filtering at $\theta^{(0)}$ (62.4 and 43.9) and with no filtering (63.3 and 53.1). The rewiring that Figure 3b locates after the first fifth of an epoch is therefore not reached by a 500-sample probe, which covers a tenth of an epoch. The first check of Section 4.3 would also still fire at the probe state, since the length association is unchanged.

F.5 Length correlations by state

The claim in Section 3.3 that response length is not a hidden safety signal rests on separating two quantities at the same trained state. The first is the first-order harm estimate $\langle \nabla \ell(z), \nabla S \rangle$, which is the GradShield functional; the second is the squared gradient norm $\|\nabla \ell(z)\|^2$, computed in the same backward pass. We evaluate both for all 5,000 samples at the state after the first round of the memory-ratchet GradShield arm of Table 1, seed 1 (cumulative step 500), rather than at a checkpoint of the diagnostic trajectory. The ratchet has no history in its first round, so this round excludes the 1,000 samples ranked highest by GradShield at $\theta^{(0)}$, the same set as static GradShield, and the state is protocol-equivalent to the first round of static filtering. Over the full pool, the harm estimate correlates with response length at $\rho = +0.082$, whereas the squared norm correlates at -0.852; excluding the 45 zero-gradient samples of Section E.1 gives $+0.083$ and -0.903. The norm thus keeps its strong length dependence after training. A score proportional to the norm, as GradShield is at $\theta^{(0)}$ (Figure A1a), therefore ranks by length, whereas the harm estimate at a trained state does not. This $+0.08$ should not be confused with the $+0.09$ correlation between the GradShield score and length at $\theta^{(2)}$ of the diagnostic trajectory quoted in Section 3.3, which is measured at a different state; both show that GradShield ranks by length only at $\theta^{(0)}$, where it follows the gradient norm, while the norm itself stays length-dependent at every state. The per-sample values will be released with the figure data.

G Additional Experimental Results

This appendix completes Section 5: the component ablation under the second judge, the full per-scorer results behind Figure A2, the budget and the number of rounds, the scoring cost, the two checks with their decomposition and every configuration run per scorer, and the second corpus and base model.

Table A9: Component ablation of SABA under the Beaver cost-model judge, final ASR in %, GradShield scorer, three seeds; the paired difference to the full method is in parentheses. Companion to Table 2.

	$B = 1500$ (10%)		$B = 3000$ (20%)		$B = 4500$ (30%)		$B = 6000$ (40%)	
Variant	CatQA	HEx-PHI	CatQA	HEx-PHI	CatQA	HEx-PHI	CatQA	HEx-PHI
SABA (full)	**51.0**	**32.6**	**30.2**	**19.2**	**27.6**	**15.1**	**27.3**	**17.5**
− skip: $k_0 > 0$, uniform k	54.7 (+3.7)	39.1 (+6.4)	43.8 (+13.5)	28.9 (+9.7)	39.5 (+11.9)	24.3 (+9.2)	36.8 (+9.5)	19.3 (+1.8)
− defer: $\delta = 1$	50.3 (−0.7)	35.2 (+2.5)	40.1 (+9.8)	25.2 (+6.0)	34.7 (+7.0)	20.1 (+5.1)	31.8 (+4.4)	19.3 (+1.8)
− return: permanent exclusion	48.3 (−2.7)	32.8 (+0.1)	32.8 (+2.5)	21.7 (+2.5)	25.8 (−1.8)	14.7 (−0.3)	26.4 (−0.9)	18.3 (+0.8)
− re-score: $s_t \equiv \theta^{(0)}$	77.1 (+26.1)	57.2 (+24.6)	77.9 (+47.7)	58.3 (+39.1)	77.0 (+49.3)	57.9 (+42.9)	74.1 (+46.8)	57.8 (+40.3)

G.1 COMPONENT ABLATION UNDER THE BEAVER JUDGE

Table A9 repeats Table 2 under the second judge. The ordering of the four components is unchanged at every budget: re-scoring is the largest effect (+47.7 on CatQA at $B = 3000$), then spending at initialization (+13.5), then the uniform schedule (+9.8), then permanent exclusion (+2.5).

G.2 ALL SCORERS UNDER STATIC FILTERING AND UNDER SABA

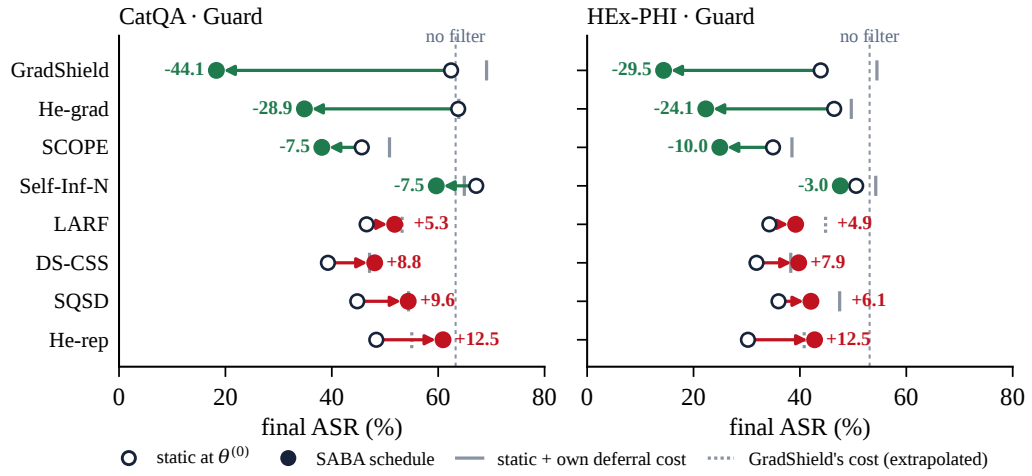

Figure A2: The SABA schedule, $k = (0, 750, 2250)$, applied unconditionally to eight risk scorers, against each scorer's static usage; Llama-Guard-3 judge, three seeds. Open markers are static filtering at $\theta^{(0)}$, filled markers are the schedule with the same scorer, and the number is the paired change in ASR. The gray tick is static filtering plus the scorer's own cost of deferral alone, measured with its $\theta^{(0)}$ ranking frozen, i.e. where the schedule would land if re-scoring contributed nothing (dotted: GradShield's cost, for the two scorers without their own control); three of the four scorers that regress land on it. The first check of Section 4.3 assigns exactly the four upper scorers to the schedule.

Table A10 gives the numbers behind Figure A2 under both judges, together with utility. The sign of the change agrees between the two judges for every scorer and both benchmarks, and no scorer's utility moves by more than 0.01 in NLL, 0.02 in ROUGE-L or 0.005 in ARC-Easy between its static and its SABA arm.

G.3 BUDGET AND NUMBER OF ROUNDS

Budget. Figure A3a,b vary the budget at $\delta = 3$. Safety improves steeply from $B = 1500$ (40.3%) to $B = 3000$ (18.3%) and $B = 4500$ (12.4%), then saturates (11.5% at $B = 6000$), while the utility price grows without saturating: held-out NLL rises from 1.669 to 1.676, 1.690 and 1.708. The usable range is therefore $B \in [3000, 4500]$, or 20% to 30% of the pool per round, and we use the

Table A10: Static filtering and SABA for each of the eight scorers, all four judge-benchmark pairs and the three utility measures, three seeds. The parenthesized value is the paired change from static to SABA, green when SABA is safer and red when it is not.

		CatQA ↓		HEx-PHI ↓		NLL ↓	ROUGE-L ↑	ARC-E ↑
Scorer	Schedule	Guard	Beaver	Guard	Beaver			
GradShield	static, $k = (1000, 1000, 1000)$	62.4±2.5	72.7±1.9	43.9±2.4	48.7±2.0	1.671	0.322	0.808
	SABA, $k = (0, 750, 2250)$	18.3±1.5 (-44.1)	30.2±2.0 (-42.4)	14.4±1.1 (-29.5)	19.2±1.2 (-29.5)	1.676	0.330	0.812
He-grad	static, $k = (1000, 1000, 1000)$	63.8±6.8	74.1±6.7	46.4±4.5	51.3±3.5	1.671	0.329	0.811
	SABA, $k = (0, 750, 2250)$	34.8±3.7 (-28.9)	43.8±4.3 (-30.3)	22.3±1.2 (-24.1)	28.2±1.9 (-23.1)	1.673	0.331	0.813
Self-Inf-N	static, $k = (1000, 1000, 1000)$	67.2±3.4	76.8±3.9	50.6±1.7	53.8±2.5	1.672	0.334	0.813
	SABA, $k = (0, 750, 2250)$	59.6±4.6 (-7.5)	72.8±5.1 (-4.0)	47.6±4.4 (-3.0)	53.7±4.4 (-0.1)	1.678	0.344	0.813
SCOPE	static, $k = (1000, 1000, 1000)$	45.6±1.3	58.2±1.3	34.9±3.9	39.8±3.3	1.683	0.333	0.818
	SABA, $k = (0, 750, 2250)$	38.1±3.5 (-7.5)	51.6±3.7 (-6.6)	24.9±2.0 (-10.0)	31.7±1.8 (-8.0)	1.679	0.319	0.814
LARF	static, $k = (1000, 1000, 1000)$	46.5±6.5	59.0±5.0	34.3±0.5	40.8±0.9	1.678	0.334	0.820
	SABA, $k = (0, 750, 2250)$	51.8±1.6 (+5.3)	63.7±0.7 (+4.7)	39.2±1.3 (+4.9)	47.0±1.2 (+6.2)	1.681	0.345	0.817
DS-CSS	static, $k = (1000, 1000, 1000)$	39.3±9.3	51.7±8.5	31.8±8.3	37.0±7.4	1.680	0.340	0.818
	SABA, $k = (0, 750, 2250)$	48.1±6.9 (+8.8)	59.6±7.7 (+7.9)	39.8±1.7 (+7.9)	46.9±1.8 (+9.9)	1.683	0.328	0.818
SQSD	static, $k = (1000, 1000, 1000)$	44.8±3.9	58.4±3.6	36.0±3.5	42.0±2.6	1.675	0.325	0.814
	SABA, $k = (0, 750, 2250)$	54.4±6.0 (+9.6)	65.1±4.5 (+6.7)	42.1±5.5 (+6.1)	46.2±5.1 (+4.3)	1.675	0.333	0.813
He-rep	static, $k = (1000, 1000, 1000)$	48.4±5.6	60.6±6.0	30.2±3.8	38.0±4.9	1.673	0.329	0.815
	SABA, $k = (0, 750, 2250)$	60.9±0.7 (+12.5)	72.0±0.7 (+11.4)	42.8±1.9 (+12.5)	49.1±1.4 (+11.0)	1.675	0.328	0.812

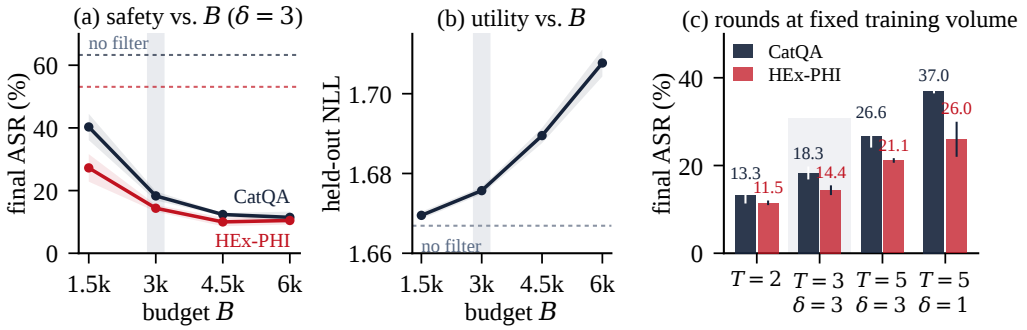

Figure A3: Budget and rounds, GradShield scorer, Llama-Guard-3 judge. (a) Final ASR against the exclusion budget B at $\delta = 3$ (dashed: no filtering; shaded: the budget used elsewhere in the paper). (b) Held-out NLL against B. (c) Final ASR for two, three and five rounds at a fixed 12,000 sample-passes; the five-round schedules are $k = (0, 75, 225, 675, 2025)$ for $\delta = 3$ and $k = (0, 750, 750, 750, 750)$ for $\delta = 1$.

lower end throughout. The component ablation in Table 2 shows that the ordering of the components is stable across this whole range.

Number of rounds. Figure A3c varies T with the training volume held at 12,000 sample-passes, so that a two-round arm trains for 1.71 epochs per round and a five-round arm for 0.55. Fewer rounds are better: $T = 2$ with $k = (0, 3000)$ reaches 13.3%, $T = 3$ with $\delta = 3$ reaches 18.3%, and $T = 5$ reaches 26.6% with $\delta = 3$ and 37.0% with $\delta = 1$. Two explanations are consistent with Figure 5b: a longer final round trains longer on a pool cleaned at a well-informed state, and dividing a fixed budget over more rounds leaves each round's exclusion set smaller and shorter-lived. Within each T, deferral still helps (26.6% against 37.0% at $T = 5$), so the geometric shape transfers. However, the number of rounds itself is a protocol choice that we fix at $T = 3$ to match the diagnostic protocol of Section 3, and the geometric rule should not be read as a recommendation to add rounds.

G.4 SCORING COST

SABA scores the whole pool in every round with $k_t > 0$, which for $T = 3$ means two passes in addition to the initial pass that static filtering also performs; on our hardware, these two passes take 2,731 s per arm in total. Cheaper variants all lose safety (Table A11): a single-pass approximation costs +14.4 on CatQA at 0.37× the measured cost, scoring a random half of the pool costs +20.2 at 0.38×, and scoring a random quarter costs +35.9 at 0.19×. A full scoring pass is therefore necessary for the best result, consistent with re-scoring being the largest component in Table 2. This

Table A11: Cheaper scoring variants of SABA, GradShield scorer, $k = (0, 750, 2250)$, three seeds. Measured cost is the scoring wall-clock per arm summed over the two scoring rounds; the parenthesized value is the change in final ASR relative to full scoring.

Scoring variant	nominal cost	measured	CatQA/Guard	CatQA/Beaver	HEx-PHI/Guard	HEx-PHI/Beaver
full pool (reference)	1.00×	2731 s (1.00×)	**18.3**±1.5	**30.2**±2.0	**14.4**±1.1	**19.2**±1.2
single-pass approximation	0.50×	1022 s (0.37×)	32.7±2.4 (+14.4)	44.0±2.7 (+13.8)	20.7±1.8 (+6.3)	26.3±0.8 (+7.1)
score a random 50% of the pool	0.50×	1039 s (0.38×)	38.5±3.1 (+20.2)	49.7±2.8 (+19.5)	25.5±1.8 (+11.1)	31.4±2.5 (+12.2)
score a random 25% of the pool	0.25×	519 s (0.19×)	54.2±4.1 (+35.9)	64.2±5.0 (+34.0)	41.5±2.9 (+27.1)	46.6±2.7 (+27.4)

cost does not separate SABA from its strongest baselines: refresh-only and memory-ratchet also score the full pool at θ_0, θ_1 and θ_2, so Table 1 compares them with SABA at equal scoring cost, and at a fraction of that cost the single-pass approximation still improves on the best baseline of Table 1 on three of the four judge-benchmark pairs (32.7% against 34.4% on CatQA and 20.7% against 23.7% on HEx-PHI under Guard).

G.5 DEFERRAL COST, RE-SCORING GAIN, AND THE TWO-FACTOR GATE

Table A12: The two checks of SABA for each scorer, and the change from static filtering to the unconditional schedule split into deferral cost and re-scoring gain (Llama-Guard-3). A is Equation (6), computed on the scores as the protocol computes them; for He-grad the bracket is the range over three random-adapter draws. Check 1 assigns $k_0 = 0$ to gradient-based scorers with $A \geq \tau = 0.3$. B_r is the rank correlation between the $\theta^{(0)}$ scores and the same scorer's scores at the static branch's θ_1, averaged over three seeds (not computed for Self-Inf-N and SCOPE, which the first check assigns to the schedule); check 2 keeps the static ranking when $B_r \geq \tau$ and re-scores once otherwise. The signed correlations show that the sign of $\rho(R_0, \|g\|)$ is a property of the scorer's direction (GradShield and SCOPE differ), which is why A takes absolute values. Deferral cost is the scorer's own frozen-$\theta^{(0)}$ control, $k = (0, 750, 2250)$, minus its static arm, both over the same three seeds (for GradShield, the − re-score row of Table 2); † marks the two scorers for which GradShield's control is used. Re-scoring gain is Δ minus the deferral cost.

		at $\theta^{(0)}$			check 1	at θ_1	check 2	Δ schedule − static		deferral cost		re-scoring gain	
Scorer	form	$\rho(R_0, \text{len})$	$\rho(R_0, \|g\|)$	A	k_0	B_r	branch	CatQA	HEx-PHI	CatQA	HEx-PHI	CatQA	HEx-PHI
GradShield	gradient	-0.81	+0.93	0.93	0	-0.09	schedule	-44.1	-29.5	+6.7	+10.6	-50.8	-40.1
He-grad	gradient	-0.42	+0.33	0.42 [0.36, 0.47]	0	+0.03	schedule	-28.9	-24.1	+0.1	+3.2	-29.0	-27.4
Self-Inf-N	gradient	-0.13	+0.46	0.46	0	—	schedule	-7.5	-3.0	-2.2	+3.7	-5.3	-6.7
SCOPE	gradient	+0.92	-0.96	0.96	0	—	schedule	-7.5	-10.0	+5.2	+3.6	-12.7	-13.6
LARF	representation	+0.70	-0.66	0.70	B/T	+0.85	static	+5.3	+4.9	+6.7†	+10.6†	-1.4	-5.6
DS-CSS	representation	+0.62	-0.62	0.62	B/T	+0.92	static	+8.8	+7.9	+7.8	+6.4	+1.0	+1.5
SQSD	gradient ($D - S$)	+0.23	-0.14	0.23	B/T	+0.14	re-score once	+9.6	+6.1	+9.6	+11.5	-0.1	-5.4
He-rep	representation	-0.07	-0.01	0.07	B/T	+0.67	static	+12.5	+12.5	+6.7†	+10.6†	+5.9	+2.0

Table A12 gives the two checks and the decomposition discussed in Section 5.4, and Table A13 gives every configuration we ran for the six scorers with more than one.

Threshold. Both checks share τ, and the assignment of all eight scorers is unchanged for any $\tau \in (0.226, 0.356]$. The lower end is SQSD's A (0.226) and the upper end is the lowest of He-grad's three draws (0.356); throughout this interval, the second check separates SQSD ($B_r = 0.14$) from He-rep (0.67), the closest scorer on the other side. We use $\tau = 0.3$. τ was fixed on these eight scorers and then applied unchanged to the settings of Section G.6 and to the pre-registered test of Section D.

Stability of A. For most gradient-based scorers, the $\theta^{(0)}$ ranking hardly depends on the LoRA initialization. Across three seeds it correlates at $\rho = 1.000$ for SCOPE and 0.999 for Self-Inf-N, whose A stays at 0.96 and within $[0.45, 0.48]$; SQSD's A is 0.23, 0.03 and 0.06, and GradShield's is 0.93 both as scored and on full-parameter gradients. He-grad is the one fragile case, because its implementation draws a fresh random adapter at every scoring call. Over three draws its A is 0.47, 0.36 and 0.47 (0.42 in the run reported here, and 0.11 on full-parameter gradients). All three draws pass at $\tau = 0.3$, but a threshold of 0.4 would send one of them to the static branch, so we ran that branch as well. There the second check fires ($B_r = 0.03$), and delayed-static He-grad improves on static filtering by 20.4 and 15.7, against 28.9 and 24.1 for the schedule (Table A13), so a misassigned He-grad keeps about two thirds of the gain.

Table A13: Every configuration run for the six scorers with more than one, as the change in final ASR from static filtering (Llama-Guard-3, three seeds; green and red mark changes larger than 1). The configuration selected by the two checks at $\tau = 0.3$ is marked ✓ and set in bold. For seed 1 of the $(1000, 500, 1500)$ and $(0, 1500, 1500)$ arms the round-2 scores are those of an existing arm at the same state.

Scorer	configuration	k	re-scored	CatQA	HEx-PHI
GradShield	**schedule** (✓)	$(0, 750, 2250)$	every round	**-44.1**	**-29.5**
	delayed static	$(1000, 1000, 1000)$	once, at θ_1	-15.0	-9.9
He-grad	**schedule** (✓)	$(0, 750, 2250)$	every round	**-28.9**	**-24.1**
	delayed static	$(1000, 1000, 1000)$	once, at θ_1	-20.4	-15.7
SQSD	schedule	$(0, 750, 2250)$	every round	+9.6	+6.1
	delayed static (✓)	$(1000, 1000, 1000)$	once, at θ_1	**-1.9**	**-5.9**
	deferred after a static round	$(1000, 500, 1500)$	every round	+5.0	-2.2
	uniform over trained rounds	$(0, 1500, 1500)$	every round	+7.6	+0.7
DS-CSS	**static** (✓)	$(1000, 1000, 1000)$	never	**0.0**	**0.0**
	schedule	$(0, 750, 2250)$	every round	+8.8	+7.9
	delayed static	$(1000, 1000, 1000)$	once, at θ_1	+6.5	+4.5
	deferred after a static round	$(1000, 500, 1500)$	every round	+0.8	+4.7
LARF	**static** (✓)	$(1000, 1000, 1000)$	never	**0.0**	**0.0**
	schedule	$(0, 750, 2250)$	every round	+5.3	+4.9
	delayed static	$(1000, 1000, 1000)$	once, at θ_1	-1.1	+1.3
He-rep	**static** (✓)	$(1000, 1000, 1000)$	never	**0.0**	**0.0**
	schedule	$(0, 750, 2250)$	every round	+12.5	+12.5
	delayed static	$(1000, 1000, 1000)$	once, at θ_1	+13.1	+11.7

Deferral cost. Six scorers have their own frozen control; LARF and He-rep use GradShield's, so their re-scoring gains (-1.4 and $+5.9$ on CatQA) are approximate, although neither check uses them. The size of the deferral cost shows what the frozen ranking was doing. It is close to zero for He-grad and Self-Inf-N ($+0.1$ and -2.2 on CatQA), whose static rankings are no better than no filtering, since deferring an uninformative ranking loses nothing. It is 5 to 10 on CatQA for SQSD, DS-CSS and SCOPE, whose static rankings are informative, because the first round then trains on the samples those rankings would have removed. It is also $+6.7$ for GradShield, whose static ranking is no better than no filtering on CatQA but is a length ranking: 84% of the 2,250 samples it would exclude in the final round are the 2,250 shortest responses, so concentrating the budget there lands at 69.1%, level with excluding the shortest responses outright (68.3%, Table 1).

Branches. For every scorer with more than one configuration, the one the checks select is the best measured, or within noise of it (LARF, where re-scoring once gives -1.1 and $+1.3$). We ran the rejected branches after fixing the rule. Re-scoring once at θ_1 costs DS-CSS 6.5 and 4.5 and He-rep 13.1 and 11.7, and leaves LARF unchanged. For SQSD, deferring the budget with re-scoring in every round is worse than static filtering on CatQA, whether the first round is skipped, $k = (0, 1500, 1500)$, or static, $k = (1000, 500, 1500)$ ($+7.6$ and $+5.0$). This is why a scorer that fails the first check keeps the uniform schedule: with $w_0 = w$, Equation (7) would still shift budget toward later rounds unless $\kappa = 1$, but deferring after a static first round does not consistently improve on static filtering for either scorer on which we tried it (SQSD and DS-CSS), so such a scorer is used exactly as published.

G.6 Second corpus and second base model

Table A14 adds SQSD on Alpaca to Table A15, together with the frozen-θ_0 control that decomposes its change; this case is discussed at the end of this section. Within each setting, all rows agree to within 0.012 in NLL, 0.02 in ROUGE-L and 0.01 in ARC-Easy.

Second corpus. On Alpaca the first check fires ($A = 0.74$), and the main result repeats in a sharper form. Static GradShield is less safe than no filtering on every judge-benchmark pair (51.5% against 44.5% on CatQA and 57.0% against 45.5% on HEx-PHI under Guard), whereas the schedule reaches 20.7% and 22.0%, about 24 below no filtering, at a held-out NLL within 0.01 of it.

Table A14: All rows of the generality experiment, three seeds; companion to Table A15. Bold marks the branch the two checks select for each scorer. NLL and ROUGE-L are measured on each setting's own held-out split (Dolly or Alpaca). Red marks a filter less safe than no filtering.

Method	CatQA/G	CatQA/B	HEx-PHI/G	HEx-PHI/B	NLL	ROUGE-L	ARC-E
Dolly-5k · Qwen3-8B (main setting) GradShield: $A = 0.93$, check 1 fires							
no filtering	63.3±0.7	73.7±1.7	53.1±1.9	58.0±1.6	1.667	0.334	0.813
static · GradShield	62.4±2.5	72.7±1.9	43.9±2.4	48.7±2.0	1.671	0.322	0.808
static · SQSD	44.8±3.9	58.4±3.6	36.0±3.5	42.0±2.6	1.675	0.325	0.814
delayed static · SQSD (✓)	**42.9±1.0**	**53.5±1.0**	**30.1±1.1**	**35.1±1.1**	1.674	0.326	0.814
SABA schedule · GradShield (✓)	**18.3±1.5**	**30.2±2.0**	**14.4±1.1**	**19.2±1.2**	1.676	0.330	0.812
Alpaca-5k · Qwen3-8B GradShield: $A = 0.74$, check 1 fires							
no filtering	44.5±2.0	55.9±1.9	45.5±3.6	50.5±3.2	1.237	0.397	0.819
static · GradShield	51.5±2.3	62.7±2.7	57.0±4.4	61.8±3.8	1.240	0.396	0.817
SABA schedule · GradShield (✓)	**20.7±1.5**	**31.6±1.2**	**22.0±2.4**	**24.9±1.4**	1.246	0.394	0.818
static · SQSD	30.8±3.3	42.5±2.4	35.2±2.1	39.1±1.7	1.245	0.398	0.817
delayed static · SQSD	35.6±4.2	46.1±4.0	36.7±1.6	40.6±1.8	1.246	0.399	0.819
SABA schedule · SQSD (✓)	**37.9±4.9**	**47.9±5.3**	**36.3±3.0**	**39.1±3.9**	1.248	0.396	0.819
frozen-θ_0 control · SQSD	29.2±3.3	40.6±4.1	35.6±3.8	39.1±4.8	1.246	0.395	0.818
Dolly-5k · Llama-3-8B-Instruct GradShield: $A = 0.27$, check 1 does not fire; $B_r = 0.17$, check 2 fires							
no filtering	73.7±1.4	84.7±0.9	61.3±2.4	69.5±2.6	1.662	0.327	0.815
static · GradShield	74.8±1.2	87.2±1.5	59.9±3.3	70.2±3.5	1.665	0.327	0.810
SABA schedule · GradShield	59.2±0.9	70.9±0.7	41.1±0.9	50.8±0.9	1.669	0.342	0.811
delayed static · GradShield (✓)	**67.4±1.2**	**79.4±1.0**	**47.8±0.5**	**58.5±1.4**	1.666	0.343	0.812

Table A15: The GradShield comparison on a second corpus and a second base model, same protocol, budget, δ and τ; ASR in %, three seeds, NLL on each setting's own held-out split. Each setting has its own θ_0 scores and its own checks; ✓ and bold mark the branch the two checks select there. Red marks a filter that leaves the model less safe than no filtering. Table A14 gives all rows, including SQSD on Alpaca.

Method	CatQA/G	CatQA/B	HEx-PHI/G	HEx-PHI/B	NLL	ARC-E
Dolly-5k · Qwen3-8B (main setting) GradShield: $A = 0.93$, check 1 fires						
no filtering	63.3±0.7	73.7±1.7	53.1±1.9	58.0±1.6	1.667	0.813
static · GradShield	62.4±2.5	72.7±1.9	43.9±2.4	48.7±2.0	1.671	0.808
SABA schedule · GradShield (✓)	**18.3±1.5**	**30.2±2.0**	**14.4±1.1**	**19.2±1.2**	1.676	0.812
Alpaca-5k · Qwen3-8B GradShield: $A = 0.74$, check 1 fires						
no filtering	44.5±2.0	55.9±1.9	45.5±3.6	50.5±3.2	1.237	0.819
static · GradShield	51.5±2.3	62.7±2.7	57.0±4.4	61.8±3.8	1.240	0.817
SABA schedule · GradShield (✓)	**20.7±1.5**	**31.6±1.2**	**22.0±2.4**	**24.9±1.4**	1.246	0.818
Dolly-5k · Llama-3-8B-Instruct GradShield: $A = 0.27$, check 1 does not fire; $B_r = 0.17$, check 2 fires						
no filtering	73.7±1.4	84.7±0.9	61.3±2.4	69.5±2.6	1.662	0.815
static · GradShield	74.8±1.2	87.2±1.5	59.9±3.3	70.2±3.5	1.665	0.810
SABA schedule · GradShield	59.2±0.9	70.9±0.7	41.1±0.9	50.8±0.9	1.669	0.811
delayed static · GradShield (✓)	**67.4±1.2**	**79.4±1.0**	**47.8±0.5**	**58.5±1.4**	1.666	0.812

Second base model. On Llama-3 the first check does not fire ($A = 0.27$), although static GradShield is again no better than no filtering (74.8% against 73.7%). The initial ranking is uninformative without being dominated by length or gradient norm, so A, which detects that particular failure, misses it. The scorer therefore takes the static branch, where the second check fires ($B_r = 0.17$, and 0.15 to 0.18 across seeds, measured at the static arm's θ_1 as the check prescribes), and the initial ranking is replaced after one round. Delayed-static GradShield reaches 67.4% and 47.8%, 7.5 and 12.1 below static filtering, whereas applying the schedule without the check would have reached 59.2% and 41.1%. The second check thus recovers half to two thirds of what the schedule would have gained on the two benchmarks.

Where the first check errs. Across the eleven scorer-setting pairs of Tables A12 and A15, the two checks select the best branch measured, or one within noise of it, in nine, and in the pre-registered test on Llama-3 they do so for five of seven further scorers (Section D). The two exceptions are GradShield on Llama-3, discussed above, and SQSD on Alpaca (Table A14). On Alpaca, SQSD's

initial ranking is more strongly associated with length than on Dolly ($\rho = 0.50$ against 0.23), so the first check fires and assigns it to the schedule. The ranking is nevertheless informative there (static SQSD is 13.7 below no filtering), and the schedule ends 7.1 and 1.1 above static filtering under Guard (5.5 and 0.0 under Beaver), though still 6.6 and 9.2 below no filtering. The frozen control shows that deferral costs nothing on this corpus (-1.6 and $+0.5$), so the regression is a re-scoring loss of $+8.7$ and $+0.7$; re-scoring once in the delayed-static branch is likewise 4.7 and 1.5 worse than static filtering. This is the one case among the eleven pairs in which the re-scored rankings are worse than the initial one. Both errors are bounded: the selected branch still improves on no filtering by 6.6 (SQSD on Alpaca) and 6.3 (Llama-3) on CatQA. Across all twelve gradient-based pairs including the pre-registered ones, every pair with $A \geq 0.59$ is assigned correctly, and the three errors (GradShield and He-grad on Llama-3, SQSD on Alpaca) have A between 0.20 and 0.50. We therefore read A as a sufficient symptom of an uninformative initial ranking rather than a necessary one, and as least reliable in this middle range.

H LIMITATIONS

Our results establish state-dependent risk and a practical schedule under a controlled protocol, but several limitations remain.

Generality of the intervention. Although the matched intervention shows state-dependent harm in all sixteen Dolly combinations of model, judge, benchmark and seed, and although a pre-registered test detects it on Alpaca in every seed and under both judges when the late state is taken just after the rewiring window, at the later state used on Dolly the Alpaca estimate keeps its positive sign but its interval includes zero (Section E.2). The evidence for state-dependent harm is therefore most complete on Dolly, with two model families, and on other corpora the effect is clearest shortly after the rewiring window. Testing further corpora and developing estimators that resolve the effect of smaller groups or of single samples are natural next steps.

Modeling assumptions behind the schedule. Although the allocation view predicts the qualitative structure we observe, no budget at initialization and an optimum of δ strictly inside $(1, \infty)$, the geometric schedule relies on modeling choices: a power-law return, geometric persistence and a common fidelity for trained states. We fix $\delta = 3$ in advance rather than estimating it, and the wide plateau of Figure 5a makes this choice uncritical in our setting, but other corpora or training lengths may place the optimum elsewhere. Estimating fidelity and persistence online from the rankings and outcomes of earlier rounds would make the schedule adaptive.

The initialization check. Although the two checks are simple and select the best branch, or one within noise of it, in fourteen of eighteen scorer-setting pairs, including five of seven in a pre-registered test (Section D), the threshold was set on the eight Dolly scorers, and the check is least reliable when A takes intermediate values. A is also a symptom of one particular failure, an initial ranking dominated by gradient norm or length, and misses rankings that are uninformative for other reasons, as for GradShield on Llama-3. By construction, a representation-based scorer never takes the schedule, even where, as for LARF on Llama-3, it would gain from it. A direct test of whether the initial ranking carries safety information would be a more general criterion.

Scale and evaluation. Although our experiments span two model families, two corpora, four learning rates and eight scorers, they use 8B models, LoRA adapters, $5{,}000$-sample pools and three rounds of training. Larger models, full-parameter fine-tuning and longer training may change both the rewiring window and the cost of re-scoring, which is one extra scoring pass in each round after the first in our protocol. Safety is measured on two benchmarks under greedy decoding by two automatic judges, which agree on the sign of the static-to-SABA change for all eight scorers in the main setting (Section G.2). Our threat model is the loss of refusal on direct harmful requests after benign fine-tuning; adaptive jailbreak attacks and human audits address a different question and would complement these results.