WHAT MAKES AN UPDATE HARMFUL? FUNCTIONAL RISK IN MODEL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

What makes an adaptation update harmful to what a model already knows? Existing preservation methods constrain updates using historical gradients, activation-derived directions, or protected subspaces. We take a more direct functional view: for a protected loss $\mathcal{L}_c$, we ask how a concrete update Δ_t proposed by the optimizer changes it. To first order this change is $G_t^\top \Delta_t$ with $G_t = \nabla_\theta \mathcal{L}_c(\theta_t)$, which we call the functional risk of the update. Estimating G_t from finite reference data raises two questions for any method built on such estimates: how accurately must the signal be estimated to judge an update, and how long does it remain informative as the model changes? Judging an update requires only the gradient component along it, but AdamW proposals are nearly orthogonal to the protected gradient, and we prove that any procedure then needs a number of reference examples that grows with the inverse squared cosine between the two. Correcting an update is easier: a correction along the estimated gradient lowers the protected loss whenever the estimate is aligned with the true gradient. Across model states, a cached estimate changes later decisions, most strongly after a task switch. Motivated by these findings, we introduce Dynamic Functional Risk Control (DFRC), a lightweight controller that periodically refreshes its estimates and minimally modifies candidate updates whose predicted risk exceeds a tolerance; its forgetting over a run is bounded independently of detection accuracy. In sequential adaptation of two model families over three task orders, DFRC improves average accuracy and reduces forgetting relative to sequential LoRA, and further reduces forgetting when combined with an existing LoRA continual-learning method.

1 INTRODUCTION

Foundation models are routinely adapted after pretraining. For large language models (LLMs), supervised fine-tuning and parameter-efficient adaptation are standard ways to add domain knowledge and new behaviors (Hu et al., 2022; Lin et al., 2026), but adaptation can degrade capabilities the model already has (McCloskey & Cohen, 1989; French, 1999), including previously learned tasks, general knowledge, and safety behavior (Kotha et al., 2024; Biderman et al., 2024; Qi et al., 2024; Lin et al., 2026; Tang et al., 2026). Methods for mitigating this degradation constrain adaptation through parameter importance, stored examples or gradients, or protected subspaces (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Farajtabar et al., 2020; Wang et al., 2023; De Lange et al., 2022; Wang et al., 2024), and recent work observes that protected directions themselves drift during training (Zhang et al., 2025). We study a more local question: what makes a particular optimizer update harmful to what the model already knows?

For a protected loss $\mathcal{L}_c$ and the parameter change Δ_t proposed by the optimizer at state θ_t, the change of $\mathcal{L}_c$ is, to first order, $\langle \nabla \mathcal{L}_c(\theta_t), \Delta_t \rangle$, and the update is harmful when this functional-risk score is positive. Estimating this score from a finite reference set with a gradient computed at an earlier state introduces an error that splits exactly into a sampling term, a staleness term, and a higher-order remainder (Section 3). Measuring these terms on AdamW proposals during LoRA adaptation, we find that sampling dominates for a geometric reason: the proposals are nearly orthogonal to the protected gradient, with cosines below 0.04 for every proposal we measured, so the harm of one update is a small signal that a few hundred reference examples often cannot resolve. The higher-order remainder is small except right after a task switch.

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 difficulty is asymmetric. Deciding whether an update is harmful requires resolving the sign of its small projection, and we prove that any procedure needs a number of reference examples that grows with the inverse squared cosine between the update and the protected gradient. Removing harm is different: a correction along the estimated protected gradient lowers the protected loss whenever the estimate points within ninety degrees of the true gradient, a condition on the estimate alone that can hold even when the full gradient is estimated poorly (Section 3.5). Harm is therefore hard to detect but easy to correct.

Dynamic Functional Risk Control (DFRC) is built on this asymmetry: it executes a proposal unchanged if its estimated score is within tolerance and otherwise applies the smallest correction that satisfies the estimated constraint, which moves along the estimated protected gradient; it also refreshes protected gradients periodically and constrains several objectives separately. Our contributions are as follows.

- An exact decomposition of the error of update-level risk estimates into sampling, staleness, and higher-order terms, with measurements identifying sampling as dominant (Section 3).
- A detection–correction asymmetry with a matching lower bound: any detector needs reference data growing with the inverse squared cosine between update and protected gradient, whereas a correction's benefit depends only on the alignment of the estimate (Section 3.5).
- DFRC, whose forgetting over a run is bounded independently of detection accuracy, with causal evidence that its benefit comes from the protected-gradient direction and improved retention (Sections 4 and 5).

2 Related Work

EWC penalizes changes to parameters deemed important to previous tasks (Kirkpatrick et al., 2017); GEM and A-GEM use gradients on stored examples to constrain interfering updates (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019); OGD projects new gradients to preserve previously learned outputs (Farajtabar et al., 2020); and Progressive Prompts learns a new soft prompt for each task (Razdaibiedina et al., 2023). In parameter-efficient adaptation, O-LoRA separates tasks through approximately orthogonal low-rank subspaces (Wang et al., 2023), LoRA-Null initializes adapters in the null space of activations of a small knowledge sample so that layer outputs on that sample are approximately preserved (Tang et al., 2026), MBF combines orthogonal initialization with continual merging (Qiao & Mahdavi, 2025), and DOC tracks drifting functional directions (Zhang et al., 2025). Loss-side methods such as TALR (Lin et al., 2026) are complementary. Theoretical analyses of forgetting study linear regression and the NTK regime (Evron et al., 2022; Doan et al., 2021; Lin et al., 2023; Zheng et al., 2026), and first-order estimates of how training changes another loss also underlie influence estimation (Koh & Liang, 2017; Pruthi et al., 2020; Xia et al., 2024). We instead analyze the update the optimizer proposes, the reliability of each decision, and the error of estimating it from a finite sample and a stale state; Appendix D discusses further related work.

3 Understanding Functional Risk

In this section, we define the functional-risk score of an optimizer step, bound the error of its practical estimate in Theorem 1, analyze the factors that govern this error, and prove in Section 3.5 that detecting harm is intrinsically harder than correcting it. Measurements come from the mechanism study of Section 5 (Study A); proofs are in Appendix A.

Problem setup. Consider an LLM at adaptation step t with trainable parameters $\theta_t \in \mathbb{R}^p$ and adaptation gradient g_t. The base optimizer, for example AdamW, combines g_t with its internal state and returns a *proposal* Δ_t, the increment it would apply. A controller may modify the proposal, so the *executed update* $U_t = \theta_{t+1} - \theta_t$ can differ from Δ_t. A capability or behavior c that should be preserved is represented by a *protected objective*, the expected loss over a reference distribution $\mathcal{D}_c$, with *protected gradient* G_c:

$$\mathcal{L}_c(\theta) = \mathbb{E}_{z\sim\mathcal{D}_c}[\ell_c(z;\theta)], \qquad G_c(\theta) = \nabla_\theta \mathcal{L}_c(\theta). \tag{1}$$

A proposal is harmful to c if it increases $\mathcal{L}_c$. If $\mathcal{L}_c$ is continuously differentiable on the segment from θ_t to $\theta_t + \Delta_t$, integrating the gradient along this segment gives the change caused by the proposal

exactly:

$$\Delta\mathcal{L}_{c,t} := \mathcal{L}_c(\theta_t + \Delta_t) - \mathcal{L}_c(\theta_t) = \int_0^1 \langle G_c(\theta_t + u\Delta_t), \Delta_t\rangle \, du = R_{c,t} + \varepsilon^{\mathrm{ho}}_{c,t}, \tag{2}$$

where

$$R_{c,t} = \langle G_c(\theta_t), \Delta_t\rangle, \qquad \varepsilon^{\mathrm{ho}}_{c,t} = \int_0^1 \langle G_c(\theta_t + u\Delta_t) - G_c(\theta_t), \Delta_t\rangle \, du.$$

We call $R_{c,t}$, the first-order prediction of $\Delta\mathcal{L}_{c,t}$, the *functional-risk score* of the proposal; a positive score predicts harm. The remainder $\varepsilon^{\mathrm{ho}}_{c,t}$ is due to the change of the gradient along the step.

During training, the protected gradient is estimated from a finite reference set and, because each estimate costs a pass over this set, reused for several steps. An estimate $\widehat{G}_{c,N}(\theta_s) = N^{-1}\sum_{i=1}^N \nabla_\theta \ell_c(z_i;\theta_s)$ from N reference examples at an earlier state θ_s assigns the current proposal the score $\widehat{R}_{c,t;s,N} = \langle \widehat{G}_{c,N}(\theta_s), \Delta_t\rangle$. Its error splits exactly into three terms, one for each approximation:

$$\begin{aligned}\Delta\mathcal{L}_{c,t} - \widehat{R}_{c,t;s,N} = &\underbrace{\langle G_c(\theta_s) - \widehat{G}_{c,N}(\theta_s), \Delta_t\rangle}_{\text{sampling error}} \\ &+ \underbrace{\langle G_c(\theta_t) - G_c(\theta_s), \Delta_t\rangle}_{\text{staleness error}} + \underbrace{\varepsilon^{\mathrm{ho}}_{c,t}}_{\text{higher-order remainder}}.\end{aligned} \tag{3}$$

With $\Sigma_c(\theta) = \mathrm{Cov}_{z\sim\mathcal{D}_c}[\nabla_\theta \ell_c(z;\theta)]$, the covariance of per-example protected gradients, bounding each term gives the following guarantee.

Theorem 1 (Functional-risk prediction error)**.** *Fix θ_s, θ_t, and Δ_t independently of N i.i.d. reference examples from $\mathcal{D}_c$, and assume that $\Sigma_c(\theta_s)$ is finite and $\mathbb{E}[\nabla_\theta \ell_c(z;\theta_s)] = G_c(\theta_s)$. Suppose $\nabla\mathcal{L}_c$ is β_{temp}-Lipschitz on the segment $[\theta_s, \theta_t]$ and β_{step}-Lipschitz on $[\theta_t, \theta_t + \Delta_t]$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\Delta\mathcal{L}_{c,t} - \widehat{R}_{c,t;s,N}\right| \le \sqrt{\frac{\Delta_t^\top \Sigma_c(\theta_s)\Delta_t}{N\delta}} + \beta_{\mathrm{temp}}\|\theta_t - \theta_s\|\|\Delta_t\| + \frac{\beta_{\mathrm{step}}}{2}\|\Delta_t\|^2. \tag{4}$$

Norms are Euclidean in the trainable-parameter coordinates; we omit c for a single capability and N when it is fixed, and index multiple capabilities by $j = 1, \ldots, m$. Based on Theorem 1, we analyze four factors that determine whether the estimated score can be trusted: the update that is scored (Section 3.1), the reference sample (Section 3.2), gradient staleness (Section 3.3), and the higher-order remainder, which persists with an exact and fresh gradient (Section 3.4).

3.1 The Scored Update

A common test of whether a step harms c checks the adaptation gradient and accepts the step if $G^\top g \ge 0$; GEM and A-GEM use this test (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019). For gradient descent with a scalar learning rate, the test is equivalent to testing the update, because $\Delta = -\eta g$ turns $G^\top g \ge 0$ into $G^\top \Delta \le 0$. Adaptive optimizers, however, transform the gradient before applying it, which raises the following question.

Q1: Why assess the optimizer proposal?

To address Q1, consider the preconditioned proposal

$$\Delta = -\eta P g, \tag{5}$$

with $\eta > 0$ and P nonsingular. Its first-order effect on the protected loss is $G^\top\Delta = -\eta\,(P^\top G)^\top g$, so the gradient test and the update test agree for every g exactly when the half-spaces $\{g : G^\top g \ge 0\}$ and $\{g : (P^\top G)^\top g \ge 0\}$ coincide. Proposition 2 characterizes when this happens; the proof is in Appendix A.1.

Proposition 2 (Gradient- vs. update-space compatibility)**.** *Let $G \in \mathbb{R}^p \setminus \{0\}$, $\eta > 0$, and let $P \in \mathbb{R}^{p\times p}$ be nonsingular. For $\Delta = -\eta P g$, the conditions $G^\top g \ge 0$ and $G^\top\Delta \le 0$ are equivalent for every g if and only if $P^\top G = \alpha G$ for some $\alpha > 0$.*

Remark 1. Adam and AdamW (Kingma & Ba, 2015) depart further from Eq. (5), because its scaling and momentum depend on the gradient history, and decoupled weight decay adds a term proportional to $-\theta_t$ that no test on g_t can detect (Loshchilov & Hutter, 2019). We therefore score and correct the proposal Δ_t, the increment the optimizer would apply. In Study A (Section 5; three LLMs, 16 AdamW proposals at each of three states), harm labels from 4,096 independent reference examples agreed with the proposal test for 134 of 144 proposals and with the gradient test for 106. The runs used no weight decay; adding it analytically with $\lambda \leq 0.1$ changes no decision.

3.2 The Impact of Finite Reference Sets

In practice the score must be estimated, because the protected distribution $\mathcal{D}_c$ is available only through a finite reference set. Most methods that preserve model behavior likewise derive what to protect from a finite sample, whether a Fisher diagonal, stored gradients, or subspaces of sampled representations (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Farajtabar et al., 2020; Saha et al., 2021; Wang et al., 2023). In each case the sample size limits how accurately the protected quantity is known, which raises the following question.

Q2: When is the functional-risk score reliable?

To address Q2, we analyze the three error terms of Theorem 1 in turn. For a fixed proposal Δ, the estimated score $\widehat{G}_N^\top \Delta$ averages the N scalars $\nabla_\theta \ell(z_i;\theta)^\top \Delta$, so its accuracy depends only on how much the per-example gradients vary along Δ, and the error of the full gradient vector accumulates the variation in all p directions. Lemma 3 makes this precise; the proof is in Appendix A.2.

Lemma 3 (Update-specific estimation and decision reliability)**.** *Fix θ and Δ independently of N i.i.d. reference examples. Assume a finite gradient covariance $\Sigma = \Sigma(\theta)$ and $\mathbb{E}[\nabla_\theta \ell(z;\theta)] = G = G(\theta)$. Then*

$$\mathrm{Var}(\widehat{G}_N^\top \Delta) = \frac{\Delta^\top \Sigma \Delta}{N}, \qquad \mathbb{E}\|\widehat{G}_N - G\|^2 = \frac{\mathrm{tr}\,\Sigma}{N}. \tag{6}$$

Classify a score as violating when it exceeds a threshold τ. If the population margin $m_\tau(\Delta) = |G^\top \Delta - \tau|$ is positive, the estimated and population decisions differ with probability at most $\min\{1,\ \Delta^\top \Sigma \Delta / [N m_\tau(\Delta)^2]\}$.

Remark 2. Lemma 3 indicates that the score is governed by the variance $\Delta^\top \Sigma \Delta$ along the proposal, and the vector by the total variance $\mathrm{tr}\,\Sigma$. For $\Delta \neq 0$, let $v = \Delta/\|\Delta\|$ and, when $v^\top \Sigma v > 0$, define

$$d_{\mathrm{eff}}(\Delta) = \frac{\mathrm{tr}\,\Sigma}{v^\top \Sigma v}. \tag{7}$$

This ratio is at least one, equals p for isotropic noise, and is also the ratio of two sufficient sample sizes: a correct decision for Δ is guaranteed with $d_{\mathrm{eff}}(\Delta)$ times fewer reference examples than a gradient estimate that fixes the decisions of all proposals with the same norm and margin (Appendix A.2). Both sizes grow as $m_\tau(\Delta)^{-2}$, so scores close to the threshold need precise estimates and scores far from it tolerate large errors.[1] Sampling is therefore unlikely to change a decision unless the standard error of the score, $\sqrt{\Delta^\top \Sigma \Delta / N}$, is comparable to the margin. In Study A, the median d_{eff} of the proposals ranged from 3.9×10^3 to 2.7×10^4, 60 to 530 times below its value for random directions, so proposals point along noisier directions than average. The signal-to-noise ratio of the score grew roughly as $\sqrt{N}$ (Figure 2a), yet decisions from random 256-example reference sets still disagreed with stable evaluation labels in 20% of cases (Figure 1a).

3.3 The Impact of Gradient Staleness

To analyze the staleness error, compare the exact scores that the current gradient and a gradient cached at step s assign to the same proposal, $R_t = \langle G(\theta_t), \Delta_t \rangle$ and $R_{t;s} = \langle G(\theta_s), \Delta_t \rangle$. If $\mathcal{L}$ is twice continuously differentiable near the segment $[\theta_s, \theta_t]$, then

$$R_t - R_{t;s} = \Delta_t^\top \overline{H}_{s:t} (\theta_t - \theta_s), \qquad \overline{H}_{s:t} = \int_0^1 \nabla^2 \mathcal{L}\big(\theta_s + u(\theta_t - \theta_s)\big)\, du. \tag{8}$$

[1] These probability statements assume that the state and the proposal do not depend on the reference sample. Appendix A.2 discusses the dependence introduced by reusing a fixed reference set.

Only the component of the gradient change along Δ_t affects the score. We call $|R_t - R_{t;s}|$ the projected drift; it is at most $b_t = \beta_{\text{temp}}\|\theta_t - \theta_s\|\|\Delta_t\|$ if the gradient is β_{temp}-Lipschitz on the segment. Since a decision changes only if the score moves by more than its distance to the threshold, we obtain Corollary 4; the proof is in Appendix A.3.

Corollary 4 (Stability of a cached decision)**.** *If $|R_{t;s} - \tau| > b_t$, then $R_{t;s}$ and R_t lie on the same side of τ. More generally, if $|\widehat{R}_{t;s} - R_{t;s}| \le a_t$ and $|\widehat{R}_{t;s} - \tau| > a_t + b_t$, then the estimated stale score $\widehat{R}_{t;s}$ yields the same threshold decision as the fresh population score R_t.*

Remark 3. Corollary 4 indicates that a cached gradient remains usable as long as the projected drift stays below the margin of the decision it supports. A gradient computed many steps ago can still decide a proposal with a large margin, and a recent one can fail near the threshold. Cache age matters through the displacement it permits, which is how the refresh interval of Section 4.2 controls this term. In Study A, a gradient cached at step 0 changed 18 of 48 decisions by step 200 (12 of 16 for Gemma), always where the projected drift exceeded the margin; one cached at step 100 changed 2 of 48 (Figure 2b).

3.4 The Impact of Higher-Order Effects

If the gradient is β_{step}-Lipschitz along the step, the third term of Theorem 1 satisfies $|\varepsilon_t^{\text{ho}}| \le \frac{1}{2}\beta_{\text{step}}\|\Delta_t\|^2$, which bounds the error of a single prediction. To separate harmful from harmless proposals, the scores should also order different proposals consistently with their actual effects. To analyze this, fix θ_t and treat the proposal Δ as random, representing the proposals the optimizer could produce at this state (for example, from different mini-batches), with covariance $C_t = \text{Cov}_\Delta(\Delta)$ and $G = G(\theta_t)$. The score $R_t = G^\top\Delta$ has variance $G^\top C_t G$, and we define

$$\kappa_t = \frac{\text{Std}_\Delta(\varepsilon_t^{\text{ho}})}{\sqrt{G^\top C_t G}}, \qquad \gamma_t = \text{Corr}_\Delta(R_t, \varepsilon_t^{\text{ho}}), \tag{9}$$

so that κ_t compares the variation of the remainder with that of the score and γ_t is their correlation. Together they determine how well the score tracks the actual change (Appendix A.4):

$$\text{Corr}_\Delta(R_t, \Delta\mathcal{L}_t) = \frac{1 + \gamma_t\kappa_t}{\sqrt{1 + \kappa_t^2 + 2\gamma_t\kappa_t}}. \tag{10}$$

Remark 4. For $\gamma_t = 0$ the correlation in Eq. (10) is $(1 + \kappa_t^2)^{-1/2}$, close to one for $\kappa_t \ll 1$ and decaying like $1/\kappa_t$ for $\kappa_t \gg 1$. κ_t grows when the remainder varies more or the scores vary less. Because the correlation concerns ordering, threshold decisions can remain reliable when all scores lie far on one side of the threshold. Concentrating the proposals does not remove this limit (Appendix A.4). In Study A, scoring with the gradient of the evaluation sample itself, which leaves only the remainder in the residual, gave κ_t of 0.01 to 0.15 and correlations of at least 0.99 in eight of nine model–state cells; at the first update after the task switch in Gemma, $\kappa_t = 0.67$ and the correlation was 0.92.

3.5 Hard to Detect, Easy to Correct

Remarks 2–4 answer Q2: a decision can be trusted when the estimated score is farther from the threshold than the bound in Eq. (4), and more data and refreshes shrink only its first two terms. The resulting difficulty has a geometric origin: the score $G^\top\Delta = \|G\|\|\Delta\|\cos\theta$ is small whenever Δ is nearly orthogonal to G, as AdamW proposals driven by a new task are for the protected gradient of an old one. A controller that cannot reliably tell harmful proposals from harmless ones may still act on them, which raises the following question.

Q3: Can harm be removed when it cannot be reliably detected?

To address Q3, consider a correction along the estimated protected gradient, $d = \lambda\widehat{G}$ with $\lambda > 0$, which is the form DFRC uses (Section 4.1). To first order it lowers the protected loss by $\lambda\langle G, \widehat{G}\rangle$, so its sign depends on the alignment of the estimate with the true gradient and not on the proposal. Theorem 5 shows that detection is intrinsically harder than correction; the proof is in Appendix A.6.

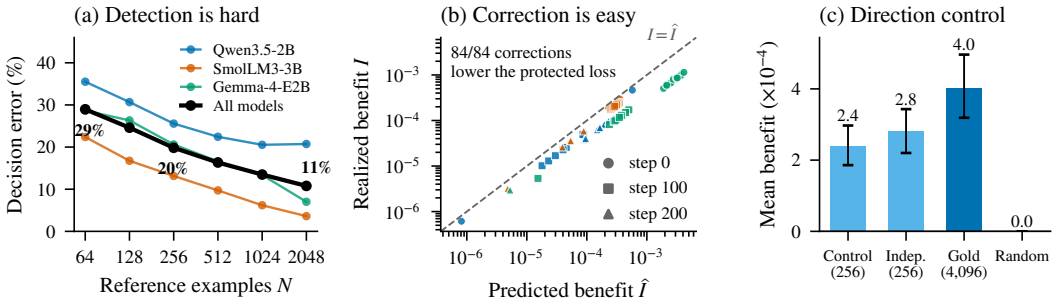


Figure 1: Hard to detect, easy to correct (Study A). (a) Disagreement of decisions from random reference sets with stable labels, over 125 stably labeled proposals. (b) Realized versus predicted benefit of the 84 triggered corrections; all are positive and lie below the diagonal by the alignment ratio of Remark 5. (c) Protected-loss reduction at equal correction norm for four directions, averaged over 83 proposals with 95% bootstrap intervals.

Theorem 5 (Detection–correction asymmetry)**.** *Let $\widehat{G}_N$ average N i.i.d. per-example gradients with mean $G \neq 0$ and covariance Σ, and let Δ be fixed with $G^\top \Delta \neq 0$ and $\Delta^\top \Sigma \Delta > 0$. Write $\cos\theta = G^\top\Delta/(\|G\|\|\Delta\|)$, $v = \Delta/\|\Delta\|$, and $u = G/\|G\|$.*

(i) Detection. *Fix $m > 0$. Any procedure that decides the sign of $G^\top\Delta$ from N Gaussian per-example gradients with covariance Σ, and errs with probability at most $\delta < 1/4$ for every such distribution with $|G^\top\Delta| = m$, requires, with $m = \|G\|\|\Delta\||\cos\theta|$,*

$$N \;\geq\; \frac{v^\top \Sigma v}{2\|G\|^2 \cos^2\theta}\,\log\frac{1}{4\delta}. \tag{11}$$

(ii) Correction. *In general, and for Gaussian gradients,*

$$\Pr\left[\langle G, \widehat{G}_N\rangle \leq 0\right] \leq \frac{u^\top\Sigma u}{N\|G\|^2} \quad \textit{and} \quad \Pr\left[\langle G, \widehat{G}_N\rangle \leq 0\right] \leq \exp\Big(-\frac{N\|G\|^2}{2u^\top\Sigma u}\Big), \tag{12}$$

respectively. Outside this event, every correction $d = \lambda\widehat{G}_N$ with $\lambda > 0$ has positive first-order benefit, whether or not the corrected proposal was harmful.

Remark 5. Part (i) holds for any procedure that sees only the N reference gradients, and the plug-in sign test attains it up to a constant (Appendix A.6), so the factor $1/\cos^2\theta$ is intrinsic to detection and absent from part (ii). To first order, a correction realizes the fraction $\langle G, \widehat{G}\rangle/\|\widehat{G}\|^2$ of its predicted benefit $\widehat{G}^\top d$, the same for all corrections made with one estimate and below one when $\widehat{G}$ is noisy. In Study A, every proposal had a cosine below 0.04 with the estimated protected gradient, with medians of 0.004 to 0.027, so $1/\cos^2\theta$ is of order 10^3 to 10^5. Figure 1 shows the resulting contrast between unreliable decisions and reliable corrections.

This answers Q3: harm that cannot be reliably detected can still be removed, because the correction direction carries the full signal $\|G\|^2$, whereas a decision sees only the small projection $G^\top\Delta$.

4 Dynamic Functional Risk Control

In this section, we introduce Dynamic Functional Risk Control (DFRC), which exploits the asymmetry of Section 3.5. DFRC scores the optimizer proposal (Remark 1), leaves it unchanged when the estimated score is within tolerance, and otherwise corrects it along the estimated protected gradient by the smallest amount that satisfies the estimated constraint (Section 4.1). It refreshes the protected gradients every K executed updates and at task boundaries (Section 4.2) and constrains several protected objectives separately (Section 4.3); Section 4.4 bounds its forgetting over a run. Algorithm 1 in Appendix A.9 summarizes the procedure.

4.1 Minimal Functional Intervention

When DFRC replaces a proposal Δ by a corrected update U at state θ, let $d = \Delta - U$ be the correction and define its benefit as the loss difference between the two endpoints, $I = \mathcal{L}(\theta + \Delta) - \mathcal{L}(\theta + U)$. A positive I means that executing U leaves the protected loss lower than executing Δ, although $\mathcal{L}(\theta+U)$ may still exceed $\mathcal{L}(\theta)$. With a gradient estimate $\widehat{G}$, the first-order prediction of the benefit is $\widehat{I} = \widehat{G}^\top d$. The error of this prediction is proportional to the length of the correction, with a curvature term of order $(\|\Delta\| + \|U\|)\|d\|$ compared with $\|\Delta\|^2$ for the step itself (Proposition 8 in Appendix A.5), so for a small correction the higher-order effects at the two endpoints largely cancel.

DFRC therefore executes a proposal unchanged if it satisfies the estimated protection constraint, and otherwise executes the closest update that does. With $\widehat{G}_t$ the protected-gradient estimate available at step t, fresh or cached, and a tolerance $\tau \geq 0$, DFRC executes

$$U_t^\star = \arg\min_U \frac{1}{2}\|U - \Delta_t\|^2 \quad \text{s.t.} \quad \langle \widehat{G}_t, U \rangle \leq \tau. \tag{13}$$

For $\widehat{G}_t \neq 0$ the solution is

$$U_t^\star = \Delta_t - \frac{(\widehat{R}_t - \tau)_+}{\|\widehat{G}_t\|^2}\widehat{G}_t, \qquad \widehat{R}_t = \langle \widehat{G}_t, \Delta_t \rangle, \tag{14}$$

and for $\widehat{G}_t = 0$ the constraint is vacuous and $U_t^\star = \Delta_t$. From a violating proposal, DFRC subtracts the multiple of $\widehat{G}_t$ that brings its estimated score exactly to τ, so the predicted benefit of the correction equals the estimated excess, $\widehat{G}_t^\top(\Delta_t - U_t^\star) = \widehat{R}_t - \tau$. Because the feasible half-space contains the origin, the correction never lengthens the step, $\|U_t^\star\| \leq \|\Delta_t\|$, and among feasible updates it minimizes the worst-case first-order disturbance to adaptation (Appendix A.8).

Benefit of the correction. If $\widehat{R}_t > \tau$, the correction $d_t = \Delta_t - U_t^\star$ is a positive multiple of $\widehat{G}_t$, so it moves along the estimated protected gradient and Theorem 5 applies.

Corollary 6 (Benefit of a single-constraint projection)**.** *Let $\tau \geq 0$ and suppose $\widehat{R}_t > \tau$, so that Eq.* (13) *produces a nonzero correction d_t. If $\nabla\mathcal{L}$ is β_{int}-Lipschitz on the convex hull of θ_t, $\theta_t + \Delta_t$, and $\theta_t + U_t^\star$, then*

$$I \;\geq\; \|d_t\|\Big(\frac{\langle G(\theta_t), \widehat{G}_t\rangle}{\|\widehat{G}_t\|} - \beta_{\text{int}}\|\Delta_t\|\Big), \tag{15}$$

so $I > 0$ whenever $\langle G(\theta_t), \widehat{G}_t\rangle > \beta_{\text{int}}\|\Delta_t\|\,\|\widehat{G}_t\|$. If in addition $\|G(\theta_t) - \widehat{G}_t\| \leq e$, then $|I - \widehat{I}|/\widehat{I} \leq (e + \beta_{\text{int}}\|\Delta_t\|)/\|\widehat{G}_t\|$.

Remark 6. Only the component of the true gradient along the estimate enters the condition, and by Theorem 5 its first-order part fails with probability at most $u^\top\Sigma u/(N\|G\|^2)$. In Study A, the full error of the 256-example estimate, measured against 4,096 independent examples, was 0.70 to 1.30 times the norm of the estimate, so the stronger full-error condition $\|\widehat{G}_t\| > e + \beta_{\text{int}}\|\Delta_t\|$ had little or no margin in every cell. All 84 triggered corrections nevertheless lowered the measured protected loss, by a nearly constant fraction (0.27 to 0.83) of the predicted benefit within each cell, as Remark 5 implies, and random directions of equal norm had no effect (Figure 1b,c).

4.2 Periodic Gradient Refresh

DFRC recomputes the protected gradients on the fixed reference sets every K executed updates and whenever the protected objectives change, in particular at task boundaries; between refreshes, each proposal is scored against the cached gradients. Corollary 4 shows how the interval controls the staleness error. Because corrections never lengthen a step, the displacement since the last refresh at step s satisfies $\|\theta_t - \theta_s\| \leq \sum_{i=s}^{t-1}\|U_i\| \leq K\bar{u}$ with $\bar{u} = \max_{s\leq i<t}\|\Delta_i\|$, so

$$|R_t - R_{t;s}| \leq \beta_{\text{temp}}\, K\, \bar{u}\, \|\Delta_t\|. \tag{16}$$

The staleness error thus grows at most linearly in K, while each refresh costs one pass over the N reference examples, or N/K per-example gradients per step on average. A drift-triggered variant (Appendix A.11) was not selective: in Study A its Lipschitz-type drift estimate exceeded the actual drift by a median factor of 85 to 300, so it refreshed at every step.

4.3 MULTI-CAPABILITY CONTROL

With several protected objectives, a constraint on their averaged gradient can hide an increase in one loss behind a decrease in another, a conflict familiar from multi-task learning (Sener & Koltun, 2018; Yu et al., 2020). DFRC-MC therefore constrains each objective separately, with tolerances $\tau_j \geq 0$:

$$U_t^\star = \arg\min_U \frac{1}{2}\|U - \Delta_t\|^2 \quad \text{s.t.} \quad \langle \widehat{G}_{j,t}, U\rangle \leq \tau_j, \quad j = 1, \ldots, m. \tag{17}$$

The feasible set is an intersection of half-spaces containing the origin, so the solution is unique, satisfies $\|U_t^\star\| \leq \|\Delta_t\|$, and is computed from an $m \times m$ dual problem on the Gram matrix of the protected gradients (Appendix A.8). The variant DFRC-Avg imposes the single averaged constraint $\langle \sum_j w_j \widehat{G}_{j,t}, U\rangle \leq \sum_j w_j \tau_j$ with positive weights summing to one, in which estimated violations of different objectives can cancel; Section 5.3 compares both variants.

4.4 TRAJECTORY-LEVEL GUARANTEE

The results so far concern one step. Because every update DFRC executes satisfies its estimated constraints, they combine into a bound on the change of each protected loss over a whole run; the proof is in Appendix A.7.

Theorem 7 (Trajectory-level guarantee)**.** *Run DFRC-MC for T updates with refreshes at most K updates apart. For objective j, let $\xi_{j,s} = \widehat{G}_{j,s} - G_j(\theta_s)$ be the error of the gradient computed at refresh s, let D_s be the net displacement over the interval that starts at s, and let $\bar{u} = \max_t \|\Delta_t\|$. If $\nabla \mathcal{L}_j$ is β_j-Lipschitz on a convex set containing the trajectory, then*

$$\mathcal{L}_j(\theta_T) - \mathcal{L}_j(\theta_0) \;\leq\; T\tau_j + \sum_s |\langle \xi_{j,s}, D_s\rangle| + \beta_j K \bar{u} \sum_{t<T} \|U_t\| + \frac{\beta_j}{2} \sum_{t<T} \|U_t\|^2 \tag{18}$$

for every objective j, whereas DFRC-Avg satisfies it only for the weighted average $\sum_j w_j \mathcal{L}_j$. Moreover, for $\tau \geq 0$ a single-constraint correction satisfies $\|d_t\| \leq \|\Delta_t\|\,|\cos\angle(\Delta_t, \widehat{G}_t)|$, which bounds its first-order cost to adaptation.

Remark 7. Eq. (18) accumulates the sampling, staleness, and curvature terms of Theorem 1 along the run: sampling errors enter only through their projections on the net displacement of each refresh interval, the staleness term grows with K, and no term depends on whether individual detections were correct. It also predicts that forgetting decreases with larger N and smaller K. The plasticity statement rests on the geometry that makes detection hard. Every proposal in Study A had a cosine below 0.04, and in Study B the median correction was 0.45% to 1.82% of the proposal norm. In the sequential-adaptation runs, the averaged constraint left at least one protected task with positive predicted risk on 55.5% of steps, which per-objective constraints exclude.

5 EXPERIMENTS

5.1 SETUP

Mechanism studies. Study A adapts Qwen3.5-2B, SmolLM3-3B, and Gemma-4-E2B with rank-8 LoRA from a DBpedia-trained checkpoint toward Amazon and saves 16 exact AdamW proposals at steps 0, 100, and 200. Protected gradients use a 256-example control set, loss changes are measured in FP32 on 2,048 disjoint examples, and a preregistered follow-up adds 4,096 more (Appendix B.1). Study B repeats the decision-reliability, state-dependence, and one-step-intervention measurements on SmolLM3-3B, Qwen3-8B, and Ministral3-8B (Bakouch et al., 2025; Yang et al., 2025; Mistral AI, 2025) with 32 to 128 candidates per state (Appendix B.2); both studies use $\tau = 0$.

Sequential adaptation. We adapt Qwen3-8B-Base and SmolLM3-3B-Base on DBpedia, Amazon Reviews, Yahoo Answers, and AG News (Zhang et al., 2015) in three task orders, with rank-8 LoRA on every `q_proj` and `v_proj` module, AdamW with learning rate 10^{-4}, and 1,000 executed updates per task. DFRC uses $N = 256$ task-balanced protected examples in total, $K = 50$, and $\tau = 0$. We compare sequential LoRA (SeqLoRA), DFRC with an averaged constraint (DFRC-Avg), DFRC-MC, DOC (Zhang et al., 2025), MBF (Qiao & Mahdavi, 2025), and MBF+DFRC, and report average accuracy (AA) and forgetting (F) as defined in Appendix C.

Table 1: Sequential adaptation (seed 42), averaged over three task orders. F: forgetting.

	Qwen3-8B		SmolLM3-3B	
Method	AA ↑	F ↓	AA ↑	F ↓
SeqLoRA	0.7637	0.0349	0.7562	0.0367
DFRC-Avg	0.7769	0.0175	0.7612	0.0328
DFRC-MC	0.7688	0.0214	0.7723	0.0212
DOC	0.7640	0.0302	0.7393	0.0577
MBF	0.7788	0.0155	0.7724	0.0144
MBF+DFRC	**0.7800**	**0.0116**	**0.7812**	**0.0109**

5.2 Detection and Correction on Saved Proposals

Study A appears alongside each result above, and Appendix B.1 collects it. Study B shows the same pattern with larger models (Appendix B.2). At $N = 256$ and the initial state, its decisions reach balanced accuracy 0.80 to 0.85 although the gradient cosine is only 0.31 to 0.71. In paired one-step interventions, DFRC significantly reduces protected damage in five of nine model–state pairs and is never significantly worse, with median corrections of 0.45% to 1.82% of the proposal norm. Two results qualify the analysis. For Gemma at step 0 in Study A, the negative score–loss correlation persisted with 4,096 reference examples, contrary to the preregistered sampling explanation; two halves of the evaluation set agree on only 6 of 16 labels there, so the effects lie below what the available data can resolve (Theorem 5). In Study B, a correction direction from a larger reference set raised the protected loss at the initial SmolLM state, so the conditions of Corollary 6 can fail.

5.3 Sequential Adaptation

Table 1 reports the seed-42 matrix. Relative to SeqLoRA, DFRC-MC improves average accuracy in all six model–order pairs (one-sided sign test, $p = 1/64$), by 1.06 points on average, and reduces forgetting by 1.45 points; DFRC-Avg improves five of six pairs. Separate constraints matter most for SmolLM order O2, where the averaged controller loses 1.31 points and DFRC-MC gains 1.11; on the same 18,000 updates, separate constraints intervene on 13.4% of steps that the averaged constraint leaves untouched, which diagnoses cancellation among task risks. On Qwen, DFRC-Avg attains a higher mean accuracy than DFRC-MC. MBF is a strong baseline that improves all six pairs. MBF+DFRC, which uses the averaged controller, also improves all six pairs, with the largest mean gains of $+2.07$ accuracy points and -2.45 forgetting points relative to SeqLoRA; relative to MBF alone, it gains 0.12 accuracy points on Qwen and 0.88 on SmolLM and lowers forgetting on both (0.0155 to 0.0116 and 0.0144 to 0.0109). Current-task accuracy is essentially unchanged (Appendix C).

6 Conclusion

We studied what makes an update harmful by analyzing the update the optimizer proposes. Because AdamW proposals are nearly orthogonal to the protected gradient, detecting harm is intrinsically hard, whereas correcting it along the estimated protected gradient is reliable whenever the estimate is aligned with the true gradient; DFRC turns this asymmetry into a trajectory-level guarantee and improved retention. Our guarantees rest on unmeasured smoothness constants and, for the lower bound, on a Gaussian model; the condition of Corollary 6 cannot be checked during training; the experiments use one seed; the drift-triggered refresh we tried was not selective; and DFRC's 256-example memory has not been compared with replay of the same examples.

AI use statement

In this work, we used generative AI tools for language editing, manuscript organization, literature search, code review, and drafting of experiment scripts and analysis utilities. We have reviewed all AI-assisted work: AI-assisted text was revised by the authors, experimental claims were checked against executable code and logged outputs, and AI-assisted code was inspected and tested before

use in reported experiments. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This work studies optimization and continual adaptation of language models using existing public benchmarks and model checkpoints. It does not involve human subjects or collection of private user data. The proposed controller can reduce unintended degradation during model adaptation, but it does not itself guarantee preservation of safety, factuality, or other properties that are not represented by the chosen protected objective.

Reproducibility statement

The appendix specifies the mathematical derivations, model and dataset configurations, controller implementation, evaluation metrics, and data-isolation rules. The experimental code records protocol hashes, model and module manifests, optimizer-state checks, reference-set provenance, and per-run outputs. The preregistered protocols, per-order results, and aggregation scripts will be released as anonymized supplementary material.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, 2018.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, et al. SmolLM3: Smol, multilingual, long-context reasoner. Hugging Face Blog, 2025.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, et al. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Jean Bretagnolle and Catherine Huber. Estimation des densités: Risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Proceedings of the 35th Conference on Learning Theory*, 2022.

Junfeng Fang, Houcheng Jiang, Kun Wang, et al. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, 2025.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

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 *International Conference on Learning Representations*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, 2024.

Jiacheng Lin, Zhongruo Wang, Kun Qian, Tian Wang, Arvind Srinivasan, Hansi Zeng, Ruochen Jiao, Xie Zhou, Jiri Gesi, Dakuo Wang, Yufan Guo, Kai Zhong, Weiqi Zhang, Sujay Sanghavi, Changyou Chen, Hyokun Yun, and Lihong Li. SFT doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in LLMs. In *International Conference on Learning Representations*, 2026.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. TRGP: Trust region gradient projection for continual learning. In *International Conference on Learning Representations*, 2022.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6467–6476, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. KeepLoRA: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, 2026.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

Mistral AI. Ministral 3 8B. Technical documentation, 2025.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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 *International Conference on Learning Representations*, 2024.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single LoRA continual learning via continual merging. *arXiv preprint arXiv:2512.23017*, 2025.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. In *International Conference on Learning Representations*, 2023.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Pengwei Tang, Xiaolin Hu, Yong Liu, Lizhong Ding, Dongjie Zhang, Xing Wu, and Debing Zhang. Put the space of LoRA initialization to the extreme to preserve pre-trained knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33232–33240, 2026.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 10658–10671, 2023.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.

Zhixin Zhang, Zeming Wei, and Meng Sun. Dynamic orthogonal continual fine-tuning for mitigating catastrophic forgettings. *arXiv preprint arXiv:2509.23893*, 2025.

Guodong Zheng, Peng Wang, Shengchao Hu, Quan Zheng, and Li Shen. Understanding the dynamics of forgetting and generalization in continual learning via the neural tangent kernel. In *International Conference on Learning Representations*, 2026.

A Proofs and Additional Derivations

A.1 Gradient-space and update-space tests

Proof of Proposition 2. Set $h = P^\top G$, which is nonzero because P is nonsingular. Since $G^\top \Delta = -\eta\, h^\top g$, the condition $G^\top \Delta \le 0$ is equivalent to $h^\top g \ge 0$. If $h = \alpha G$ with $\alpha > 0$, then $h^\top g$ and $G^\top g$ have the same sign for every g. Conversely, suppose the two conditions agree for every g. Any $x \perp G$ satisfies $G^\top(\pm x) \ge 0$, hence $h^\top x \ge 0$ and $-h^\top x \ge 0$, so $h^\top x = 0$. Thus $h \in \operatorname{span}(G)$, say $h = \alpha G$. Taking $g = G$ gives $\alpha \ge 0$, and $h \neq 0$ excludes $\alpha = 0$. □

A diagonal example. Proposition 2 indicates that the two tests agree for all gradients only if G is an eigenvector of $P^\top$ with a positive eigenvalue; for a symmetric positive-definite P that is not a multiple of the identity, this holds only on a set of measure zero. A diagonal preconditioner already separates the tests. For $G = (1, 1)^\top$, $g = (1, -0.5)^\top$, and $P = \operatorname{diag}(0.1, 10)$, the gradient passes the test with $G^\top g = 0.5$, but $G^\top \Delta = 4.9\eta > 0$.

For symmetric positive-definite P, the condition $PG = \alpha G$ states that G is an eigenvector of P. If P is not a multiple of the identity, its eigenspaces are proper subspaces, and their finite union has Lebesgue measure zero. For a general nonsingular P, the condition involves $P^\top G$, and it can fail even when G is an eigenvector of P.

Minimal correction under preconditioning. The map $g' \mapsto U = -\eta P g'$ is a bijection, so every candidate update corresponds to exactly one modified gradient. Substituting $U = -\eta P g'$ and $\Delta = -\eta P g$ into $\min_U \frac{1}{2}\|U - \Delta\|^2$ subject to $G^\top U \le 0$, and dividing the objective by η^2, gives

$$\min_{g'} \ \frac{1}{2}(g' - g)^\top P^\top P (g' - g) \quad \text{s.t.} \quad (P^\top G)^\top g' \ge 0. \tag{19}$$

Euclidean distance in update space thus corresponds to the $P^\top P$ metric in gradient space, and the constraint normal becomes $P^\top G$. A-GEM projects the gradient in the Euclidean metric with constraint normal G, which coincides with Eq. (19) when P is a positive multiple of the identity.

Decoupled weight decay. In AdamW the proposal has the form $\Delta_t = -\eta\,(a_t + \lambda \theta_t)$, where a_t is the adaptive moment step and $\lambda \ge 0$ is the decoupled weight-decay coefficient. Its functional-risk score therefore contains the term $-\eta\lambda\, G^\top \theta_t$. At a fixed state this term does not depend on the current gradient, so a compatibility test based only on $G^\top g_t$ cannot account for it. The term vanishes when $\lambda = 0$.

A.2 Sampling: proofs and assumptions

Proof of Lemma 3. Write $g_i = \nabla_\theta \ell(z_i; \theta)$. Independence gives $\mathbb{E}\widehat{G}_N = G$ and $\operatorname{Cov}(\widehat{G}_N) = \Sigma/N$. Projecting the covariance onto Δ yields the scalar variance, and its trace yields the mean-squared error. If the two decisions differ and $m_\tau(\Delta) > 0$, then $|(\widehat{G}_N - G)^\top \Delta| \ge m_\tau(\Delta)$; for instance, $G^\top \Delta > \tau \ge \widehat{G}_N^\top \Delta$ implies $(G - \widehat{G}_N)^\top \Delta \ge G^\top \Delta - \tau$, and the reverse case is symmetric. Chebyshev's inequality, together with the trivial bound of one, gives the claim. □

Uniform and update-specific certificates. If $\|\widehat{G}_N - G\| < m/\|\Delta\|$, then $|(\widehat{G}_N - G)^\top \Delta'| < m$ for every Δ' with $\|\Delta'\| \le \|\Delta\|$, so no decision with margin at least m can change. Markov's inequality applied to $\|\widehat{G}_N - G\|^2$ bounds the probability that this event fails by $\|\Delta\|^2 \operatorname{tr}\Sigma/(Nm^2)$, which is at most δ once $N \ge B_{\mathrm{unif}} = \|\Delta\|^2 \operatorname{tr}\Sigma/(\delta m^2)$. The update-specific bound of Lemma 3 is at most δ once $N \ge B_{\mathrm{dir}} = \Delta^\top \Sigma \Delta/(\delta m^2)$. The ratio of the two sufficient budgets is $B_{\mathrm{unif}}/B_{\mathrm{dir}} = \|\Delta\|^2 \operatorname{tr}\Sigma/(\Delta^\top \Sigma \Delta) = d_{\mathrm{eff}}(\Delta)$.

Sub-Gaussian tails. If $X_i = \Delta^\top(g_i - G)$ satisfies $\mathbb{E}\exp(\lambda X_i) \le \exp(\lambda^2 v_\Delta/2)$ for all $\lambda \in \mathbb{R}$, independence and the Chernoff bound give

$$\Pr\Big(\Big|N^{-1}\textstyle\sum_i X_i\Big| \ge m\Big) \le 2\exp\Big(-\tfrac{Nm^2}{2v_\Delta}\Big),$$

so $N \ge 2v_\Delta m_\tau(\Delta)^{-2}\log(2/\delta)$ ensures decision error at most δ. For T proposals fixed independently of the reference sample, a union bound makes all T decisions correct with probability at least $1-\delta$ once $\log(2/\delta)$ is replaced by $\log(2T/\delta)$ and m_τ by the smallest margin, so a whole run costs only a factor logarithmic in T; the dependence created by reusing one sample along a trajectory is discussed below. The variance proxy v_Δ is an upper bound on the directional variance $\Delta^\top\Sigma\Delta$ and can be larger than it. In Study A the per-example contributions $\Delta^\top g_i$ are heavy-tailed, with a median excess kurtosis of 142 (76 to 1,832 across the 144 proposals), so a few reference examples can dominate a score (Appendix B.1). The proxy can then greatly exceed the directional variance, and the Chebyshev form of Lemma 3 is the relevant one.

Properties of d_{eff}. For unit v, $v^\top\Sigma v \le \lambda_{\max}(\Sigma) \le \operatorname{tr}\Sigma$, so $d_{\mathrm{eff}} \ge 1$ whenever it is defined; it is infinite if $v^\top\Sigma v = 0 < \operatorname{tr}\Sigma$ and undefined if $\Sigma = 0$. For v uniform on the unit sphere, $\mathbb{E}_v[v^\top\Sigma v] = \operatorname{tr}\Sigma/p$. The normalized noise alignment $a_\Sigma(v) = p\, v^\top\Sigma v/\operatorname{tr}\Sigma$ therefore has mean one over random directions and satisfies $a_\Sigma(v) = p/d_{\mathrm{eff}}(v)$, so low directional variance corresponds to large d_{eff}. The average of a_Σ over proposals is in general not the reciprocal of the average of d_{eff}, because averaging does not commute with taking reciprocals; we therefore state which quantity each reported average uses.

Reference reuse. If each reference sample is drawn independently of the current state and proposal given the preceding history, Lemma 3 holds conditionally. A fixed reference set reused throughout adaptation does not provide this independence. Neither does a correction $d = \Delta - U(\widehat{G})$ computed from the estimate itself, because the projected error $(\widehat{G} - G)^\top d$ need not have mean zero. Two remedies are available. One is an independent evaluation sample drawn after the correction is selected. The other is a uniform event $\|\widehat{G} - G\| \le e$, which gives $|(\widehat{G} - G)^\top d| \le e\|d\|$ for every d and underlies Proposition 8.

Weighting. The $1/N$ expressions assume equally weighted i.i.d. example gradients. A loss averaged over all answer tokens of variable-length examples is a ratio estimator, and its variance must be computed under that weighting. Sampling n examples uniformly without replacement from a pool of size M gives covariance $(1 - n/M)\Sigma_M/n$, where Σ_M is the pool covariance with denominator $M - 1$.

A.3 Staleness, the higher-order remainder, and Theorem 1

The fundamental theorem of calculus applied to $\nabla\mathcal{L}$ along $[\theta_s, \theta_t]$ gives $G(\theta_t) - G(\theta_s) = \overline{H}_{s:t}(\theta_t - \theta_s)$; taking the inner product with Δ_t proves Eq. (8). Lipschitz continuity of $\nabla\mathcal{L}$ and the Cauchy–Schwarz inequality give $|R_t - R_{t;s}| \le b_t$ without requiring second derivatives.

Proof of Corollary 4. A real number cannot cross a threshold under a perturbation smaller than its distance to that threshold. Apply this to $|R_t - R_{t;s}| \le b_t$. For the estimated score, the triangle inequality gives $|R_t - \widehat{R}_{t;s}| \le a_t + b_t < |\widehat{R}_{t;s} - \tau|$. □

For a continuously differentiable loss, $\Delta\mathcal{L} = \int_0^1 G(\theta + u\Delta)^\top\Delta\, du$. Subtracting $G(\theta)^\top\Delta$ and using the Lipschitz gradient bounds the remainder by $\beta_{\mathrm{step}}\|\Delta\|^2/2$. For a twice continuously differen-

tiable loss, the remainder has the integral form

$$\varepsilon^{\mathrm{ho}} = \int_0^1 (1-u)\,\Delta^\top \nabla^2 \mathcal{L}(\theta + u\Delta)\Delta\, du.$$

Proof of Theorem 1. Bound the three terms of Eq. (3) separately. By Lemma 3 at $\theta = \theta_s$ and Chebyshev's inequality, the sampling term is at most $\sqrt{\Delta_t^\top \Sigma(\theta_s)\Delta_t/(N\delta)}$ with probability at least $1-\delta$. The staleness term is at most b_t, and the higher-order remainder is at most $\beta_{\mathrm{step}}\|\Delta_t\|^2/2$, both deterministically. The triangle inequality completes the proof. □

A.4 First-order signal and higher-order variation

Set $R = G^\top\Delta$ and $E = \varepsilon^{\mathrm{ho}}$, so that $\Delta\mathcal{L} = R + E$ and $\mathrm{Var}(R) = G^\top CG$ for candidate covariance C. With $\kappa = \mathrm{Std}(E)/\mathrm{Std}(R)$ and $\gamma = \mathrm{Corr}(R, E)$, we have $\mathrm{Cov}(R, R+E) = \mathrm{Var}(R)(1+\gamma\kappa)$ and $\mathrm{Var}(R+E) = \mathrm{Var}(R)(1+\kappa^2+2\gamma\kappa)$. Dividing the covariance by the product of the standard deviations gives Eq. (10), whenever the variances involved are finite and nonzero. If E is constant and R is not, the correlation equals one although γ is undefined.

Quadratic model. Let $q(\Delta) = \frac{1}{2}\Delta^\top H\Delta$ with $H = \nabla^2\mathcal{L}(\theta_t)$ be the quadratic part of the remainder, and model the proposals as Gaussian, $\Delta \sim \mathcal{N}(\mu, C)$. Then

$$\mathrm{Var}_\Delta(q) = \underbrace{\frac{1}{2}\,\mathrm{tr}[(HC)^2]}_{\text{spread}} + \underbrace{\mu^\top HCH\mu}_{\text{mean}}, \qquad \mathrm{Cov}_\Delta(G^\top\Delta, q) = G^\top CH\mu. \tag{20}$$

The spread term comes from the fluctuation of the proposals around their mean, and the mean term from the cross term between this fluctuation and the mean proposal μ. Only the mean term is correlated with the score, so an analysis that centers the proposals ($\mu = 0$) removes exactly the correlated part.

To prove these identities, write $\Delta = \mu + x$ with $x \sim \mathcal{N}(0, C)$. Then

$$q(\Delta) - \mathbb{E}q(\Delta) = \mu^\top Hx + \frac{1}{2}\{x^\top Hx - \mathrm{tr}(HC)\}.$$

Centered Gaussian third moments vanish, so the quadratic part is uncorrelated with every linear function of x, in particular with $\mu^\top Hx$ and with $G^\top x$. The two parts have variances $\mu^\top HCH\mu$ and $\mathrm{tr}[(HC)^2]/2$, using $\mathrm{Var}(x^\top Hx) = 2\,\mathrm{tr}[(HC)^2]$ for symmetric H, and $\mathrm{Cov}(G^\top x, \mu^\top Hx) = G^\top CH\mu$. This proves Eq. (20), including for singular C. Under $C \mapsto sC$, the mean component and $G^\top CG$ scale as s, and the spread component scales as s^2. In particular, for $G^\top CG > 0$ and $\Delta_s \sim \mathcal{N}(\mu, sC)$ with $s > 0$, fixing G, H, μ gives

$$\frac{\mathrm{Var}(q(\Delta_s))}{\mathrm{Var}(G^\top\Delta_s)} = \frac{\mu^\top HCH\mu}{G^\top CG} + s\,\frac{\mathrm{tr}[(HC)^2]}{2G^\top CG}.$$

The ratio is nondecreasing in s and tends to $\mu^\top HCH\mu/(G^\top CG)$ as $s \to 0$. For an exactly quadratic loss, $\varepsilon^{\mathrm{ho}} = q$, so this limit is the lower limit of κ_t^2 stated in Section 3.4, for fixed G, H, μ, and shape of C.

If the Hessian is L_H-Lipschitz in operator norm along the step, Taylor's theorem with integral remainder gives $|\varepsilon^{\mathrm{ho}} - q(\Delta)| \le L_H\|\Delta\|^3/6$, which bounds the gap between the quadratic model and the full remainder.

A.5 Intervention reliability

The following bound supports the discussion of corrections in Section 4.1.

Proposition 8 (Intervention prediction error)**.** *Suppose $\|G(\theta) - \widehat{G}\| \le e$ and $\nabla\mathcal{L}$ is β_{int}-Lipschitz on the convex hull of θ, $\theta+\Delta$, and $\theta+U$. Then*

$$|I - \widehat{I}| \le \Big[e + \frac{\beta_{\mathrm{int}}}{2}\big(\|\Delta\| + \|U\|\big)\Big]\|d\|. \tag{21}$$

In particular, $I > 0$ whenever $\widehat{I}$ exceeds the right-hand side.

Proof of Proposition 8. Let $x(u) = \theta + (1-u)U + u\Delta$, so that $x'(u) = d$. The fundamental theorem of calculus gives $I = \int_0^1 G(x(u))^\top d\, du$. Since $\|G(x(u)) - \widehat{G}\| \le e + \beta_{\text{int}}\|(1-u)U + u\Delta\|$,

$$|I - \widehat{G}^\top d| \le e\|d\| + \beta_{\text{int}}\|d\| \int_0^1 \|(1-u)U + u\Delta\|\, du,$$

and convexity of the norm bounds the integral by $(\|U\| + \|\Delta\|)/2$. If $\widehat{I}$ exceeds the bound, then $I \ge \widehat{I} - |I - \widehat{I}| > 0$. □

Estimates used to compute the correction. The bound holds even when d is computed from $\widehat{G}$, because the assumption $\|G(\theta) - \widehat{G}\| \le e$ controls the whole vector. For a cached estimate $\widehat{G}_s$, e can combine the sampling error at θ_s with the drift term $\beta_{\text{temp}}\|\theta - \theta_s\|$, and the error along the correction decomposes exactly as

$$I - \widehat{G}_s^\top d = (G(\theta_s) - \widehat{G}_s)^\top d + (G(\theta) - G(\theta_s))^\top d + \{I - G(\theta)^\top d\},$$

in parallel with Eq. (3). For an exactly quadratic loss with gradient G at θ, $I = G^\top d + \frac{1}{2} d^\top H(\Delta + U)$. The second-order terms of the two endpoints, $\frac{1}{2}\Delta^\top H\Delta$ and $\frac{1}{2}U^\top HU$, cancel up to $\frac{1}{2}d^\top H(\Delta + U)$, which is of order $\|d\|(\|\Delta\| + \|U\|)$.

Proof of Corollary 6. For $\widehat{R}_t > \tau$, Eq. (14) gives $d_t = \lambda \widehat{G}_t$ with $\lambda = (\widehat{R}_t - \tau)/\|\widehat{G}_t\|^2 > 0$. As in the proof of Proposition 8, $I = G(\theta_t)^\top d_t + r$ with $|r| \le \frac{1}{2}\beta_{\text{int}}(\|\Delta_t\| + \|U_t^\star\|)\|d_t\| \le \beta_{\text{int}}\|\Delta_t\|\|d_t\|$, using $\|U_t^\star\| \le \|\Delta_t\|$ for $\tau \ge 0$ (Appendix A.8). Since $G(\theta_t)^\top d_t = \lambda\langle G(\theta_t), \widehat{G}_t\rangle = \|d_t\|\langle G(\theta_t), \widehat{G}_t\rangle/\|\widehat{G}_t\|$, Eq. (15) follows. If $\|G(\theta_t) - \widehat{G}_t\| \le e$, Proposition 8 gives $|I - \widehat{I}| \le (e + \beta_{\text{int}}\|\Delta_t\|)\|d_t\|$, and $\widehat{I} = \|\widehat{G}_t\|\|d_t\|$ yields the relative bound. Finally, $\langle G, \widehat{G}_t\rangle = \|\widehat{G}_t\|^2 + \langle G - \widehat{G}_t, \widehat{G}_t\rangle \ge \|\widehat{G}_t\|(\|\widehat{G}_t\| - e)$, so the full-error condition implies the alignment condition. □

For the step itself, when $R_t \ne 0$, $|\Delta\mathcal{L}_t - R_t| \le \beta_{\text{step}}\|\Delta_t\|^2/2$ and $|R_t| = \|G\|\|\Delta_t\|\,|\cos\angle(G, \Delta_t)|$, so $|\Delta\mathcal{L}_t - R_t|/|R_t| \le \beta_{\text{step}}\|\Delta_t\|/(2\|G\|\,|\cos\angle(G, \Delta_t)|)$, which, unlike Eq. (15), grows without bound as the proposal approaches orthogonality to G.

A.6 Detection–correction asymmetry

Proof of Theorem 5. Detection. Fix μ_0 with $\mu_0^\top \Delta = 0$ and let $P_\pm$ be Gaussian with covariance Σ and means $\mu_\pm = \mu_0 \pm m\,\Sigma\Delta/(\Delta^\top\Sigma\Delta)$, so that $\mu_\pm^\top\Delta = \pm m$. Since $\mu_+ - \mu_- = 2m\Sigma\Delta/(\Delta^\top\Sigma\Delta)$ lies in the range of Σ, $\mathrm{KL}(P_+\|P_-) = \frac{1}{2}(\mu_+ - \mu_-)^\top\Sigma^+(\mu_+ - \mu_-) = 2m^2/(\Delta^\top\Sigma\Delta)$, and for N independent gradients the divergence is N times larger. For any procedure, let A be the event that it reports a negative sign. The Bretagnolle–Huber inequality (Bretagnolle & Huber, 1979; Tsybakov, 2009) gives $P_+^N(A) + P_-^N(A^c) \ge \frac{1}{2}\exp(-2Nm^2/\Delta^\top\Sigma\Delta)$. If both error probabilities are at most δ, then $4\delta \ge \exp(-2Nm^2/\Delta^\top\Sigma\Delta)$, which rearranges to $N \ge \Delta^\top\Sigma\Delta\log(1/(4\delta))/(2m^2)$; substituting $m = \|G\|\|\Delta\|\cos\theta|$ and $\Delta = \|\Delta\|v$ gives Eq. (11).

Correction. Write $\widehat{G}_N = G + \xi$ with $\mathbb{E}\xi = 0$ and $\mathrm{Cov}\,\xi = \Sigma/N$. Then $\langle G, \widehat{G}_N\rangle = \|G\|^2 + G^\top\xi$ is nonpositive only if $|G^\top\xi| \ge \|G\|^2$, and $\mathrm{Var}(G^\top\xi) = \|G\|^2 u^\top\Sigma u/N$, so Chebyshev's inequality gives the first bound of Eq. (12). For Gaussian gradients $G^\top\xi$ is Gaussian with this variance, and the Gaussian tail bound gives the second. For any $\lambda > 0$, the first-order benefit of $d = \lambda\widehat{G}_N$ is $G^\top d = \lambda\langle G, \widehat{G}_N\rangle$, whose sign does not depend on Δ. □

The plug-in sign test attains the detection rate up to a constant: for Gaussian gradients its error is $\Phi(-\sqrt{N}m/\sqrt{\Delta^\top\Sigma\Delta}) \le \frac{1}{2}\exp(-Nm^2/(2\Delta^\top\Sigma\Delta))$, so $N \ge 2\Delta^\top\Sigma\Delta\log(1/(2\delta))/m^2$ suffices, within a factor of about four of the lower bound for small δ. Without the Gaussian assumption, Lemma 3 at $\tau = 0$ gives the Chebyshev version.

For the realized benefit, $\mathbb{E}\langle G, \widehat{G}_N\rangle = \|G\|^2$ and $\mathbb{E}\|\widehat{G}_N\|^2 = \|G\|^2 + \mathrm{tr}\,\Sigma/N$, so the first-order ratio $\langle G, \widehat{G}_N\rangle/\|\widehat{G}_N\|^2$ between realized and predicted benefit is typically close to $\|G\|^2/(\|G\|^2 + \mathrm{tr}\,\Sigma/N)$, which is below one whenever the full gradient is estimated imprecisely. As in Appendix A.2, the probability statements treat the state as independent of the reference sample.

A.7 TRAJECTORY-LEVEL GUARANTEE

Proof of Theorem 7. Fix an objective j and drop its index. Smoothness gives $\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \le G(\theta_t)^\top U_t + \frac{\beta}{2}\|U_t\|^2$. Let $s \le t$ be the last refresh. DFRC-MC executes U_t only if $\langle \widehat{G}_s, U_t\rangle \le \tau$, so

$$G(\theta_t)^\top U_t \le \tau + \langle G(\theta_t) - G(\theta_s), U_t\rangle - \langle \xi_s, U_t\rangle \le \tau + \beta\|\theta_t - \theta_s\|\|U_t\| - \langle \xi_s, U_t\rangle.$$

Because corrections never lengthen a step, $\|\theta_t - \theta_s\| \le \sum_{i=s}^{t-1}\|U_i\| \le K\bar{u}$. Summing over $t < T$, the terms $-\langle \xi_s, U_t\rangle$ of one interval add up to $-\langle \xi_s, D_s\rangle \le |\langle \xi_s, D_s\rangle|$, which gives Eq. (18). DFRC-Avg enforces only $\langle \sum_j w_j \widehat{G}_{j,s}, U_t\rangle \le \sum_j w_j \tau_j$, so the same argument applies to $\sum_j w_j \mathcal{L}_j$, with $\beta = \sum_j w_j \beta_j$ and $\xi_s = \sum_j w_j \xi_{j,s}$, but not to an individual objective. For the plasticity statement, Eq. (14) gives $\|d_t\| = (\widehat{R}_t - \tau)_+ / \|\widehat{G}_t\| \le \widehat{R}_t / \|\widehat{G}_t\| = \|\Delta_t\| \cos\angle(\Delta_t, \widehat{G}_t)$ whenever a correction occurs, and $|a^\top d_t| \le B\|d_t\|$ (Appendix A.8). □

The sampling term is large when the displacement of an interval aligns with the error of its reference estimate. Corrections move along $\widehat{G}_s$ and therefore contribute to this alignment; it is the same effect that makes the realized benefit of a correction smaller than the predicted one (Remark 5).

A.8 PROJECTION, ADAPTATION DISTORTION, AND MULTIPLE OBJECTIVES

For the single constraint, the Lagrangian $\frac{1}{2}\|U - \Delta\|^2 + \lambda(\widehat{G}^\top U - \tau)$ with $\lambda \ge 0$ gives $U = \Delta - \lambda\widehat{G}$ at stationarity. A feasible proposal has $\lambda = 0$; otherwise complementary slackness gives $\lambda = (\widehat{G}^\top \Delta - \tau)/\|\widehat{G}\|^2$, which is Eq. (14).

Norm reduction. Euclidean projection $\Pi_{\mathcal{F}}$ onto a closed convex set $\mathcal{F}$ is nonexpansive: $\|\Pi_{\mathcal{F}}(x) - \Pi_{\mathcal{F}}(y)\| \le \|x - y\|$. If $0 \in \mathcal{F}$, which holds for Eqs. (13) and (17) whenever all tolerances are nonnegative, then $\Pi_{\mathcal{F}}(0) = 0$ and $\|U^\star\| = \|\Pi_{\mathcal{F}}(\Delta)\| \le \|\Delta\|$. For the single constraint the reduction is explicit: with λ as above, $\|\Delta\|^2 - \|U^\star\|^2 = \lambda(\widehat{G}^\top\Delta + \tau)$, which at $\tau = 0$ is the Pythagorean identity $\|\Delta\|^2 = \|U^\star\|^2 + \|\Delta - U^\star\|^2$.

Adaptation distortion. A correction changes the adaptation loss by $a^\top(U - \Delta)$ to first order, for an adaptation gradient a, so the minimum-distance rule minimizes the worst-case first-order disturbance to adaptation among feasible updates; in LoRA adaptation, distances are measured in the trainable A/B coordinates. For any $B > 0$, the Cauchy–Schwarz inequality gives $|a^\top(U - \Delta)| \le B\|U - \Delta\|$ whenever $\|a\| \le B$, with equality for a parallel to $U - \Delta$. The statement is first order in the adaptation objective. Euclidean distances, and hence minimality, depend on the parameterization; for instance, rescaling LoRA factors while preserving their product changes them.

Multiple objectives. Let A have rows $\widehat{G}_j^\top$, set $K = AA^\top$, and collect the tolerances in $\boldsymbol{\tau} \ge 0$. The feasible set is nonempty, and strict convexity and coercivity of the objective give a unique primal solution. Eliminating U from the Lagrangian gives the dual

$$\min_{\lambda \ge 0} \frac{1}{2}\lambda^\top K\lambda - \lambda^\top(A\Delta - \boldsymbol{\tau}), \qquad U = \Delta - A^\top\lambda. \tag{22}$$

Polyhedral constraints admit optimal multipliers without a strict-feasibility condition; if K is singular, the multipliers may be nonunique while the primal update remains unique. For each active constraint, the predicted benefit $\widehat{G}_j^\top(\Delta - U) = \widehat{G}_j^\top\Delta - \tau_j$ is positive exactly when Δ violated that constraint; for the remaining objectives it can have either sign.

Averaged and individual constraints. Let $\mathcal{F}_{\mathrm{MC}}$ be the feasible set of Eq. (17) and $\mathcal{F}_{\mathrm{Avg}} = \{U : \langle \widehat{G}_{\mathrm{avg},t}, U\rangle \le \tau_{\mathrm{avg}}\}$, with $\widehat{G}_{\mathrm{avg},t} = \sum_j w_j \widehat{G}_{j,t}$ and $\tau_{\mathrm{avg}} = \sum_j w_j\tau_j$ for weights $w_j > 0$ summing to one. An update that satisfies every individual constraint also satisfies their weighted average, so $\mathcal{F}_{\mathrm{MC}} \subseteq \mathcal{F}_{\mathrm{Avg}}$, and minimizing the distance to Δ over nested sets gives $\|U^{\mathrm{MC}} - \Delta\| \ge \|U^{\mathrm{Avg}} - \Delta\|$. The two variants differ when objectives conflict. With zero tolerances and two objectives with gradients G and $-G$, the equal-weight average is zero and imposes no constraint, while DFRC-MC requires $G^\top U = 0$; for positively collinear gradients the two feasible sets coincide. Separate constraints thus prevent estimated violations from cancelling, at the price of larger corrections; the experiments compare this trade-off.

A.9 CACHE AND OPTIMIZER-STATE SEMANTICS

Algorithm 1 Dynamic Functional Risk Control (DFRC)

Require: θ_0; optimizer $\mathcal{O}$ with state ω_0; reference sets; tolerances $\tau_j \geq 0$; interval K

for $t = 0, 1, \ldots$ **do**
  Compute the adaptation gradient g_t at θ_t
  $(\Delta_t, \omega_{t+1}) \leftarrow \mathcal{O}(\theta_t, g_t, \omega_t)$ ▷ proposal only; not applied
  $U_t \leftarrow \Delta_t$
  if protected objectives are present **then**
    if the cache is empty, the protected objectives have changed, or $t - s \geq K$ **then**
      Recompute $\widehat{G}_j$ on the fixed reference sets at θ_t; set $s \leftarrow t$
    end if
    if $\langle \widehat{G}_j, \Delta_t \rangle > \tau_j$ for some j **then**
      $U_t \leftarrow$ the solution of Eq. (14) or Eq. (17)
    end if
  end if
  $\theta_{t+1} \leftarrow \theta_t + U_t$
end for

Within an interval of unchanged protected objectives and parameterization, the cache stores the evaluation state θ_s, the fixed reference sets, and the per-objective gradients. A refresh occurs when the cache is empty, when the protected objectives, reference sets, or parameterization change, and when $t - s \geq K$. At most one refresh occurs per proposal, and the resulting gradients, refreshed or reused, define that proposal's constraints. Without protected objectives, $U_t = \Delta_t$.

The optimizer map in Algorithm 1 advances its state once from the adaptation gradient and returns the full proposal, including weight decay when used, without applying it. Computing protected gradients neither overwrites the adaptation gradient nor advances the optimizer moments, and applying U_t does not recompute the moments from the corrected step. Later moments differ only because the executed trajectory differs. For the projection in LoRA coordinates, we represent the A/B increments and cached gradients in FP32 and form and solve the $m \times m$ Gram system in FP64, checking primal feasibility and dual optimality residuals after each solve.

A.10 Connection between analysis and control

Table 2: Results of the functional-risk analysis and how DFRC uses them.

Analysis	Result	Use in DFRC
Optimizer proposal (Sec. 3.1)	Gradient and update tests agree for all gradients only if $P^\top G = \alpha G$, $\alpha > 0$ (Prop. 2)	Test and correct the proposal Δ_t.
Sampling error (Sec. 3.2)	Decision error is governed by the gradient noise along the proposal relative to the margin (Lemma 3)	Score each proposal directly; this needs fewer examples than an accurate gradient.
Staleness (Sec. 3.3)	A cached decision is unchanged while the projected drift is below its margin (Cor. 4)	Refresh every K updates and at task boundaries; the staleness error is at most $\beta_{\text{temp}} K \bar{u} \|\Delta_t\|$ (Eq. (16)).
Higher-order effects (Sec. 3.4)	κ_t and γ_t limit how well first-order scores order proposals (Eq. (10))	Use first-order scores; the remainder was small except after a task switch (Remark 4).
Asymmetry (Sec. 3.5)	Any detector needs reference data growing with $1/\cos^2\theta$; a reliable correction direction does not (Thm. 5)	Act on noisy detections and rely on the alignment of the correction.
Intervention (Sec. 4.1)	A correction along $\widehat{G}_t$ helps whenever $\widehat{G}_t$ is aligned with G beyond a curvature term (Cor. 6)	Use the minimum-distance correction along $\widehat{G}_t$.
Trajectory (Sec. 4.4)	Forgetting is bounded by sampling, staleness, and curvature terms, with no term for detection errors (Thm. 7)	Refresh every K updates and bound each objective separately.
Multiple objectives (Sec. 4.3)	The individual feasible set is contained in the averaged one (App. A.8)	Constrain each objective separately to prevent cancellation.

A.11 Drift-triggered refresh

A refresh rule based on Corollary 4 estimates the rate of gradient change from the two most recent refreshes $s^- < s$ on the same fixed reference set $\mathcal{B}_j$ of objective j and defines

$$\widehat{\beta}_{j,s} = \frac{\|\widehat{G}_{j,s} - \widehat{G}_{j,s^-}\|}{\|\theta_s - \theta_{s^-}\| + \epsilon_d}, \qquad Q_t = \max_{1 \le j \le m} \frac{\widehat{\beta}_{j,s} \|\theta_t - \theta_s\| \|\Delta_t\|}{|\langle \widehat{G}_{j,s}, \Delta_t \rangle - \tau_j| + \epsilon_m}. \tag{23}$$

The numerator estimates b_t and the denominator is the distance of the cached score from its threshold, so Q_t estimates the drift-to-margin ratio of Corollary 4; the rule refreshes when $Q_t \ge \rho$, with a bootstrap refresh after K_{boot} updates and a maximum cache age $K_{\max}$. In Study A this rule was not selective. At step 200 with a cache from step 100, the estimate of b_t exceeded the actual projected drift by a median factor of 85 to 300, $Q_t \ge 17$ for all 48 proposals, and with $\rho = 1$ the rule refreshed for every proposal while catching the 2 decisions that changed.

The looseness comes from the Cauchy–Schwarz step, which replaces the component of the gradient change along Δ_t by its full norm. A tighter bound follows from the exact form $R_t - R_{t;s} = \Delta_t^\top \overline{H} x$ with $x = \theta_t - \theta_s$ (Eq. (8)): for any symmetric $\overline{H}$ with absolute value $|\overline{H}|$ (same eigenvectors, absolute eigenvalues),

$$|\Delta_t^\top \overline{H} x| \le \sqrt{\Delta_t^\top |\overline{H}| \Delta_t} \sqrt{x^\top |\overline{H}| x} \le \beta_{\text{temp}} \|\Delta_t\| \|x\|.$$

Hessian spectra of neural networks concentrate in a few large eigenvalues (Ghorbani et al., 2019), so the curvature along a typical update can lie far below the maximum. In Study A, the curvature along the proposals estimated from step-size scaling was 11 to 213 times smaller than $\widehat{\beta}\|\Delta\|^2$ for six proposals at step 200. Even exact curvatures leave the angle between $\overline{H}x$ and Δ_t unaccounted for, so a selective rule likely requires measuring the drift directly, for example with directional derivatives on a small probe subset of the reference set.

B Mechanism Study Details

B.1 Study A

Trajectory and measurements. A fresh rank-8, $\alpha = 16$ LoRA on the `q_proj` and `v_proj` modules (restricted to full-attention layers in Qwen3.5-2B and to independently parameterized projections in Gemma-4-E2B) is trained on DBpedia for 1,000 AdamW updates (learning rate 10^{-4}, $\beta = (0.9, 0.999)$, no weight decay); the AdamW state is then reset and the model adapts to Amazon for 200 updates. At steps 0, 100, and 200, 16 exact AdamW proposals are reconstructed from different minibatches without applying them; probes may not change parameters, gradients, optimizer state, or the training random state. Protected gradients use a 256-example control set (BF16 backbone, FP32 adapters); protected-loss changes use a disjoint 2,048-example evaluation set in FP32. The preregistered follow-up uses 4,096 further disjoint examples as a gold proxy, independent 256-example subsets, and FP32 gradients throughout. Its criteria were fixed before the analysis: the sampling explanation of the negative correlations is falsified for a key cell if the correlation with a 2,048-example reference is nonpositive, and a direction control fails if random directions reach 80% of the control-set benefit. The first criterion was met for Qwen at step 200 and failed for Gemma at step 0; the second was never met.

Table 3: Study A by model and state. Corr.: Pearson correlation between the score and the measured protected-loss change with the 256-example control set and a 2,048-example gold subset; κ: same-sample remainder ratio; SNR: median over proposals of the score's mean divided by its standard deviation across 512 random 256-example reference sets; $e/\|\widehat{G}\|$: full error of the control-set estimate relative to its norm; $I/\widehat{I}$: realized over predicted benefit; Trig.: triggered proposals out of 16.

Model	Step	Corr. 256	Corr. 2048	κ	SNR	$e/\|\widehat{G}\|$	$I/\widehat{I}$	Trig.
Qwen3.5-2B	0	0.52	0.99	0.03	1.37	0.81	0.78	4
	100	0.79	0.93	0.01	0.57	0.93	0.57	7
	200	−0.53	0.73	0.02	0.86	0.99	0.42	16
SmolLM3-3B	0	0.87	0.95	0.10	0.72	1.30	0.83	3
	100	0.75	0.94	0.03	2.99	0.72	0.71	16
	200	0.80	0.92	0.01	0.63	0.70	0.67	5
Gemma-4-E2B	0	−0.47	−0.05	0.67	0.35	0.96	0.27	16
	100	0.74	−0.29	0.15	0.55	1.28	0.34	16
	200	0.87	0.87	0.08	1.03	0.87	0.57	1

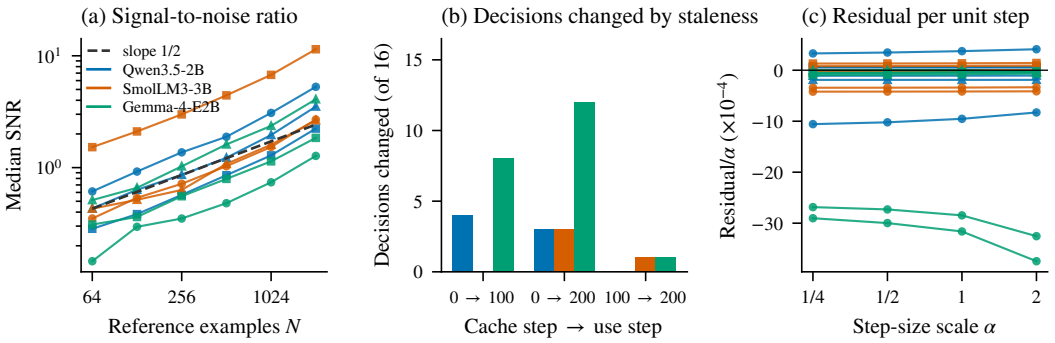


Figure 2: Mechanism measurements in Study A. (a) Median signal-to-noise ratio of the score across random reference sets for the nine model–state cells (markers: step 0 circles, 100 squares, 200 triangles), with a slope-1/2 reference line. (b) Decisions changed, out of 16 proposals per model, when a protected gradient cached at one step is used at a later step. (c) First-order residual divided by the step-size scale α for 18 proposals: a flat line means that the residual is linear in α and therefore comes from estimation error, whereas curvature appears as a slope.

Additional observations. Figure 2 shows the growth of the signal-to-noise ratio with N, the decision changes caused by stale caches, and the step-size scaling of the first-order residual; at the executed step size ($\alpha = 1$), the term linear in α accounts for more than 80% of the residual in 15 of 18 proposals and for about half or more in the other three, so estimation error dominates the residual. Per-example directional contributions are heavy-tailed, with a median excess kurtosis of 142 (76 to 1,832 across proposals), and the 10 largest of 4,096 examples carry 12% to 39% of the absolute projection mass. For Gemma at step 0, two halves of the evaluation set agree on the harm label of only 6 of 16 proposals, so the effects lie below the resolution of the available data. In Gemma at step 100, the gold-proxy scores disagree with the evaluation-set labels because of four extreme examples: removing them raises the correlation between the two from -0.18 to 0.75. These examples belong to the protected objective and are kept. Adding decoupled weight decay with $\lambda \leq 0.1$ to the saved proposals changes none of the 144 decisions.

B.2 Study B

Study B follows three models for 200 AdamW updates from a DBpedia-trained checkpoint toward Amazon with LoRA and probes states 0, 100, and 200 with independent reference blocks for estimation, population geometry, and held-out evaluation. Table 4 reports finite-reference decisions at the initial state, where 128 candidates are generated from different minibatches with fixed model and moments; balanced accuracy is reported only when both classes occur. Table 5 reports state dependence: cross-time population cosines use off-diagonal products of disjoint reference blocks, with 95% jackknife intervals for $\rho_{0,200}$ of $[0.802, 0.826]$, $[0.758, 0.836]$, and $[0.769, 0.855]$, and the static-minus-fresh normalized MAE over 200 executed updates has 95% block-bootstrap intervals of $[-0.030, 0.053]$, $[0.006, 0.084]$, and $[0.156, 0.957]$. In the paired one-step interventions (32 candidates per state, $N = 256$, 10,000 paired bootstrap resamples), protected damage reduction is significantly positive at all three states for SmolLM and at states 0 and 100 for Ministral, and Qwen intervals include zero; no pair shows a significant increase, and target-progress intervals include zero in seven of nine states. At later mixed-class states, decision balanced accuracy is about 0.59 to 0.65 even at $N = 4096$. A larger-reference correction direction at the initial SmolLM state has mean protected-loss reduction -1.51×10^{-4} with interval $[-2.20, -0.88] \times 10^{-4}$. Same-objective first-order correlations on the BF16 evaluation path are high at the initial state (0.96, 0.91, 0.99) but lower later (0.18 to 0.84), and they cannot separate curvature from finite-precision effects.

Table 4: Study B: finite-reference decisions at the initial state (medians across reference partitions). bAcc: balanced accuracy of safe/harmful decisions against a 4,096-example proxy; cos: cosine between the estimated and proxy gradients.

	$N = 256$		$N = 4096$	
Model	bAcc	cos	bAcc	cos
SmolLM3-3B	0.807	0.710	0.932	0.926
Qwen3-8B	0.847	0.405	0.953	0.650
Ministral3-8B	0.802	0.306	0.881	0.614

Table 5: Study B: state dependence of the protected gradient and stale prediction. Lower normalized MAE is better; a zero predictor scores one.

Model	$\rho_{0,100}$	$\rho_{0,200}$	$\rho_{100,200}$	nMAE fresh	nMAE static
SmolLM3-3B	0.839	0.814	0.931	0.969	0.986
Qwen3-8B	0.834	0.797	0.910	1.006	1.052
Ministral3-8B	0.896	0.812	0.906	1.204	1.667

C Sequential Adaptation Protocol

The task orders are O1: DBpedia→Amazon→Yahoo→AG News; O2: DBpedia→Amazon→AG News→Yahoo; and O3: Yahoo→Amazon→AG News→DBpedia. All methods share rank-8 LoRA

targets, optimizer, steps, and evaluation, and all use the same candidate-label scorer. DOC uses the CCIPCA-style component update of its official implementation, and MBF uses orthogonal initialization and continual merging. For T tasks and accuracy A_{ij} on task i after learning task j,

$$\mathrm{AA} = \frac{1}{T}\sum_{i=1}^{T} A_{iT}, \qquad \mathrm{BWT} = \frac{1}{T-1}\sum_{i<T}(A_{iT} - A_{ii}), \qquad \mathrm{F} = \frac{1}{T-1}\sum_{i<T}\Big(\max_{i\le j\le T} A_{ij} - A_{iT}\Big),$$

and current-task acquisition is $T^{-1}\sum_i A_{ii}$. For DFRC, the total reference budget $N = 256$ is divided approximately evenly over the protected tasks. Protected gradients are accumulated chunk by chunk with `torch.autograd.grad`, so the protected backward pass does not overwrite the adaptation gradient or the AdamW moment update, and the multi-constraint controller solves an FP64 Gram-matrix active-set problem and verifies KKT feasibility after each correction.

Table 6: Change relative to SeqLoRA, averaged over the six model–order pairs of the seed-42 matrix.

Method	ΔAA (points) ↑	ΔF (points) ↓
DFRC-Avg	+0.91	−1.07
DFRC-MC	+1.06	−1.45
DOC	−0.83	+0.82
MBF	+1.56	−2.08
MBF+DFRC	+2.07	−2.45

On the same 18,000 raw updates of the DFRC-MC runs, the averaged constraint would trigger on 59.8% of steps yet leave at least one protected task with positive predicted risk on 55.5% of steps, and the per-task constraints intervene on 13.4% of steps that the averaged constraint leaves untouched. These statistics diagnose cancellation in the local constraints. Current-task accuracy is 0.790 for SeqLoRA and 0.788 for MBF+DFRC on Qwen, and 0.784 for both on SmolLM. All entries come from one seed across three task orders, so the variation across orders is descriptive. DFRC computes $N/K \approx 5$ protected per-example gradients per step on average; wall-clock times were recorded under different hardware loads and are therefore omitted.

D Extended Related Work

Forgetting and its mitigation. Catastrophic forgetting was first described for connectionist networks (McCloskey & Cohen, 1989; French, 1999), and De Lange et al. (2022) and Wang et al. (2024) survey the methods developed since. Regularization methods penalize changes to parameters estimated to be important for earlier tasks (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018). Replay methods rehearse stored examples or stored outputs (Rebuffi et al., 2017; Rolnick et al., 2019; Buzzega et al., 2020). DFRC uses its stored examples only to estimate protected gradients, so the same memory could also be replayed.

Gradient projection and null spaces. GEM and A-GEM constrain gradients with stored examples (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019). OGD, GPM, TRGP, Adam-NSCL, and KeepLoRA restrict updates relative to subspaces computed from sampled gradients, representations, or features of earlier tasks (Farajtabar et al., 2020; Saha et al., 2021; Lin et al., 2022; Wang et al., 2021; Luo et al., 2026), and O-LoRA uses LoRA subspaces learned on earlier data (Wang et al., 2023). In knowledge editing, MEMIT preserves unrelated knowledge through key statistics estimated on a text sample, and AlphaEdit projects edits onto the null space of preserved knowledge (Meng et al., 2023; Fang et al., 2025). These methods estimate the protected quantity from a finite sample and usually keep it fixed; our analysis quantifies the sampling and staleness errors of such estimates and explains why a projection along a noisy estimate still reduces the protected loss.

Continual learning of language models. The four text-classification tasks of Section 5 come from the benchmark of de Masson d'Autume et al. (2019), and the three task orders follow prompt- and LoRA-based continual learning (Razdaibiedina et al., 2023; Wang et al., 2023). Rehearsing a small fraction of earlier data lets instruction-tuned models learn new tasks sequentially (Scialom et al.,

2022), but fine-tuning still erodes earlier abilities of large language models, including their safety behavior (Kotha et al., 2024; Qi et al., 2024), and LoRA forgets less than full fine-tuning while also learning less (Biderman et al., 2024).

Theory of forgetting. Theoretical analyses of forgetting cover linear regression (Evron et al., 2022), overparameterized linear models with varying task similarity (Lin et al., 2023), and the regime of the neural tangent kernel (Jacot et al., 2018), where forgetting depends on the kernel overlap between tasks (Doan et al., 2021), including the dynamics of forgetting and generalization at intermediate stages of training (Zheng et al., 2026). These works characterize forgetting over training on whole tasks; we analyze individual optimizer steps and the error of estimating their effect from finite and stale samples.

First-order effects of training. Influence functions and TracIn estimate the effect of training examples on a test loss from gradients (Koh & Liang, 2017; Pruthi et al., 2020), and LESS adapts such estimates to Adam by using preconditioned gradient features (Xia et al., 2024), in the same spirit as our choice to score the preconditioned proposal. With several objectives, rules that combine gradients to resolve conflicts are standard in multi-task learning (Sener & Koltun, 2018; Yu et al., 2020); DFRC-MC imposes one constraint per protected objective on the optimizer proposal and leaves the adaptation objective unconstrained.