AMORTIZING EXACT ORTHOGONALITY: DRIFT-TRIGGERED RESTORATION FOR ORTHOGONAL FINE-TUNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Orthogonal finetuning (OFT) preserves the geometry of pretrained weights, but existing methods construct an exact or approximate orthogonal transform at every optimization step. We show that the construction parameter used for this purpose does not reliably control the orthogonality realized during training: under fixed truncation settings, checkpoints from the reference OFTv2 implementation exhibit defects spanning five orders of magnitude. We introduce *Restarted Tangent Orthogonal Finetuning* (RTOFT), which replaces the per-step orthogonal construction with the tangent update $W(I + S)$, where $S^\top = -S$, and restores exact orthogonality only when the measured drift reaches a tolerance τ. For this update, $\delta(I + S) = \|S\|_2^2$ exactly, so monitoring $\|S\|_2$ and restoring before the next forward pass guarantees $\delta \leq \tau^2$ independently of the loss, data, model, or optimizer; exact orthogonality is therefore guaranteed at the returned model, while every forward pass remains within the prescribed defect budget. Across block-diagonal, subspace, and butterfly parameterizations, RTOFT reduces training time from 1.10–1.58× LoRA's wall-clock to 0.97–1.12×. The block variant runs at 0.97× LoRA's wall-clock while exceeding LoRA by 1.98 ± 0.39 points on the six-task GLUE average over five shared seeds. The same drift tolerance transfers across DeBERTaV3, Qwen, and diffusion models, whereas a fixed restoration period realizes substantially different geometry. Matched controls further show that restoration along the optimization path matters, while restoring at every step provides no measurable benefit over the drift-triggered schedule. RTOFT therefore makes realized geometric drift, rather than elapsed optimization steps, the criterion for restoring exact orthogonality.

"A carpenter and a geometer investigate the right angle in different ways; the former does so in so far as the right angle is useful for his work, while the latter inquires what it is or what sort of thing it is; for he is a spectator of the truth."

— Aristotle, *Nicomachean Ethics* I.7, trans. W. D. Ross

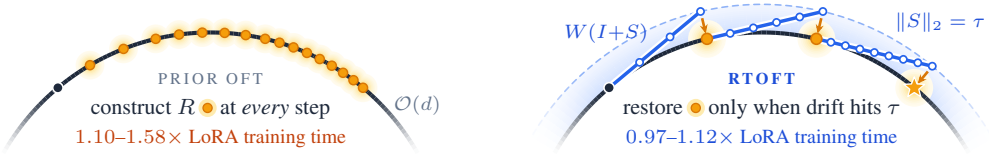


Figure 1: **RTOFT amortizes exact orthogonality.** Tangent steps $W(I+S)$ have defect exactly $\|S\|_2^2$, so exactness is restored only when the drift $\|S\|_2$ reaches τ (exact 2×2 geometry, τ enlarged).

1 INTRODUCTION

Parameter-efficient finetuning (PEFT) adapts pretrained models with a small number of trainable parameters. Low-rank methods such as LoRA (Hu et al., 2022) use additive updates, whereas orthogonal finetuning (OFT) adapts a pretrained weight matrix through $W^+ = WR$ with $R \in \mathcal{O}(d)$, preserving its singular-value spectrum and pairwise row geometry (Qiu et al., 2023). This structure can improve the target–retention trade-off, preserving more of the pretrained model's general capability while adapting to the target task (Huang et al., 2026). Recent structured and low-dimensional

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

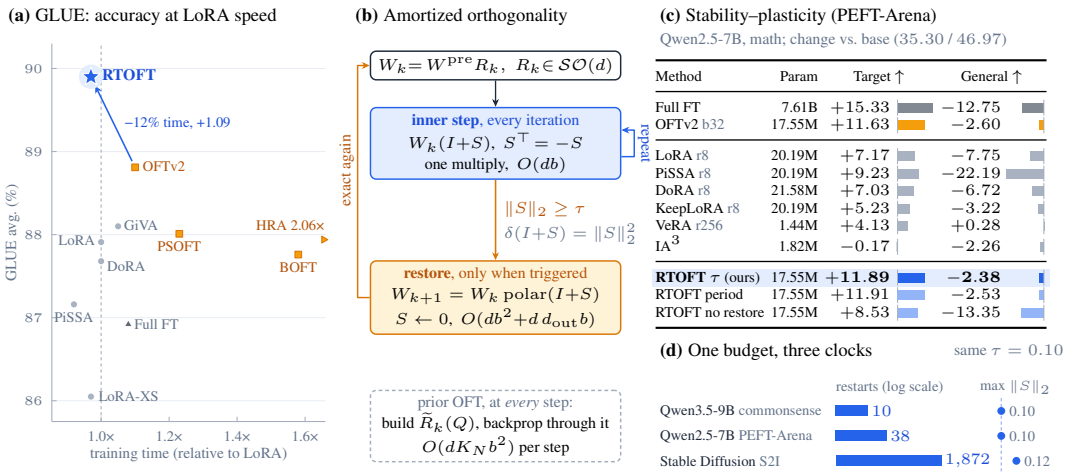


Figure 2: **Exactness is a property to maintain, not a construction to repeat at every step. (a)** On GLUE, RTOFT drops OFTv2's per-step construction at matched parameters and improves its average. **(b)** Between exact restorations a step costs one structured multiplication; the drift $\|S\|_2$ measures the violation exactly. **(c)** On PEFT-Arena, restoration matches or improves the matched orthogonal baseline on both axes; removing it costs 10.82 points of retention. **(d)** One $\tau = 0.10$ yields restart counts $187\times$ apart at realized drift 0.10–0.12.

orthogonal variants have also largely closed the trainable-parameter gap to low-rank PEFT (Liu et al., 2024; Qiu et al., 2025b; Wu et al., 2026). A prominent remaining premium is computational: current methods still construct, and differentiate through, an exact or approximate orthogonal transform inside the optimization loop.

Existing OFT methods typically reconstruct this transform at every optimization step. OFTv2 already relaxes exactness through a truncated Neumann approximation while retaining competitive downstream performance (Qiu et al., 2025b). This raises a more basic question: *why enforce exact orthogonality at every step at all?* Orthogonality is the target constraint on the adapted transform; enforcing it at every intermediate iterate is a separate algorithmic choice. Moreover, truncation order does not directly control the realized constraint violation: under a fixed OFTv2 setting, orthogonality defects vary by orders of magnitude across blocks and tasks (Section 3; per-block detail in Table 21).

We introduce *Restarted Tangent Orthogonal Finetuning* (RTOFT; Figure 1), which replaces per-step orthogonal construction with tangent-space updates and event-triggered restoration. Between restorations it uses

$$W^+ = W(I + S), \qquad S^\top = -S,$$

requiring only one structured multiplication. Its orthogonality defect is available exactly:

$$\delta(I + S) := \|(I + S)^\top (I + S) - I\|_2 = \|S\|_2^2.$$

RTOFT restores exact orthogonality when the measured drift $\|S\|_2$ reaches a tolerance τ, using a polar restoration followed by $S \leftarrow 0$. Monitoring drift after each update and restoring before the next forward pass guarantees $\delta \le \tau^2$ with no assumption on the task or the optimizer.

Our results establish three points, previewed in Figure 2:

- **Approximation order does not control realized orthogonality.** Reference OFTv2 checkpoints exhibit defects spanning five orders of magnitude, and a cheaper truncation can have a smaller realized defect than the default while matching its downstream accuracy.
- **Restoration along the path is necessary, and measured drift says when.** Omitting restoration is consistently worse, while in matched controls per-step restoration offers no measurable advantage over the drift-triggered schedule. A single tolerance keeps realized maximum drift within a narrow range across DeBERTaV3, Qwen, and diffusion models, even as restart schedules differ by orders of magnitude.
- **Most of the per-step computational premium can be removed.** Moving the construction off the optimization step reduces training time from 1.10–$1.58\times$ LoRA's wall-clock to

0.97–$1.12\times$ across three orthogonal parameterizations, spanning the truncated (OFTv2) and exact (PSOFT, BOFT) halves of the family (Table 6).

RTOFT treats the constraint the adapted model should satisfy and the tightness of its enforcement along the optimization path as two separate requirements. The existing family misses one requirement on each side: truncated constructions leave the realized violation uncontrolled (Section 3), while exact constructions enforce at every step a tightness the trajectory does not need (Section 7). For orthogonal finetuning, this shifts the design question from constructing an orthogonal transform more cheaply at every step to deciding when a construction is needed.

2 PRELIMINARIES

Let $W \in \mathbb{R}^{d_{\text{out}} \times d}$ denote a pretrained weight matrix. Low-rank PEFT adapts W additively; LoRA (Hu et al., 2022) uses $W^+ = W + \Delta W$ with $\Delta W = BA$, where $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $A \in \mathbb{R}^{r \times d}$, and $r \ll \min(d_{\text{out}}, d)$.

Orthogonal finetuning (OFT) instead uses $W^+ = WR$ with $R \in \mathcal{O}(d) := \{R \in \mathbb{R}^{d \times d} : R^\top R = I\}$, constraining the adapted transform to the orthogonal group (Qiu et al., 2023). The orthogonal PEFT methods considered here construct R from structured skew-symmetric matrices. We write $\text{Skew}(d) := \{S \in \mathbb{R}^{d \times d} : S^\top = -S\}$, $\text{skew}(A) := (A - A^\top)/2$, and denote by $\text{polar}(A)$ the orthogonal polar factor of A. Throughout, $\|\cdot\|_2$ denotes the spectral norm and $\delta(A) := \|A^\top A - I\|_2$ the *orthogonality defect* of $A \in \mathbb{R}^{d \times d}$, so that $\delta(A) = 0$ exactly when $A \in \mathcal{O}(d)$.

We evaluate three structured orthogonal parameterizations: the block-diagonal parameterization of OFT/OFTv2, in which $S = \text{blockdiag}(S_1, \ldots, S_{d/b})$ with $S_i \in \text{Skew}(b)$ for a block size b dividing d, the fixed-subspace parameterization of PSOFT, and the butterfly parameterization of BOFT (Liu et al., 2024; Qiu et al., 2025b; Wu et al., 2026). Appendix A gives the broader related-work discussion.

3 TRUNCATION ORDER DOES NOT CONTROL ORTHOGONALITY

OFTv2 (Qiu et al., 2025b) evaluates at every training step an order-k truncation $\widetilde{R}_k$ of the Cayley–Neumann series $R = (I + Q)(I - Q)^{-1} = I + 2\sum_{i \geq 1} Q^i$, with $Q \in \text{Skew}(d)$. For $Q \neq 0$, $\widetilde{R}_k$ is generally not orthogonal, and its realized defect depends jointly on k and $\|Q\|_2$. Fixing k before training therefore does not fix the orthogonality violation realized during the run.

Lemma 14 (Appendix F) shows that $\delta(\widetilde{R}_k)$ scales as a power of $\|Q\|_2$. Since $\sqrt{1+\delta}$ bounds the stretch applied to any input direction (Corollary 5), $\delta = 10^{-2}$ corresponds to approximately a half-percent stretch. Across the $5{,}184$ adapted blocks in our audit, defects vary sharply even at fixed order: the median block at $k = 6$ on STS-B lies farther from the orthogonal group than at $k = 3$ on CoLA, although the latter requires two fewer block multiplications (Table 1).

If the cross-task ratio in $\|Q\|_2$ is r, a defect proportional to $\|Q\|_2^p$ varies by r^p. All five rows of Table 1 are consistent with a common cross-task norm ratio $r \approx 3.4$, while the exponent p changes with the truncation. The same nominal order can therefore correspond to very different constraint violations across tasks.

k	CoLA	RTE	STS-B	spread
2	5.4×10^{-3}	2.7×10^{-2}	6.4×10^{-2}	12
3	7.9×10^{-9}	9.3×10^{-7}	1.2×10^{-5}	1460
4	7.9×10^{-9}	9.3×10^{-7}	1.2×10^{-5}	1460
5	$\mathbf{6.2 \times 10^{-6}}$	$\mathbf{1.5 \times 10^{-4}}$	$\mathbf{8.1 \times 10^{-4}}$	**132**
6	5.6×10^{-14}	1.5×10^{-10}	9.8×10^{-9}	174631

Table 1: Median block defect on checkpoints returned by the reference OFTv2 implementation. Spread is the largest-to-smallest ratio across tasks; $k = 5$ (bold) is the default. Per-block extremes appear in Table 21.

The dependence on order is also non-monotone. By Lemma 14, the truncated transform equals the exact Cayley transform multiplied by $I - Q^m$, with $m = 3$ for $k \in \{3, 4\}$ and $m = k - 1$ for $k \geq 5$, the map at $k = 2$ being a separate case. Lemma 12 (Appendix B) shows that odd m cancels the leading terms in the orthogonality defect, giving order $\|Q\|_2^{2m}$, whereas even m leaves a defect of order $\|Q\|_2^m$. The measured exponents are 2.00, 6.00, 6.00, 4.00, and 10.00 for $k = 2, \ldots, 6$, matching this odd–even rule. In particular, $k = 4$ returns the same matrix as $k = 3$, $R(I - Q^3)$, at the same cost.

The default is therefore not the most orthogonal setting its own construction offers. On all three tasks of Table 1, $k = 3$ realizes a smaller defect than the default while using one block multiplication fewer; run downstream, it matches the default's six-task GLUE average to $+0.03 \pm 0.05$ over five paired seeds and trains 3.6% faster (Appendix H).

Thus truncation order is neither transferable across tasks nor monotone in realized orthogonality: it determines the construction, and the constraint violation realized during training depends on the task and the dynamics.

4 RTOFT: Tangent Updates with Drift-Triggered Restoration

Section 3 shows that a construction parameter chosen before training does not control the orthogonality defect realized during training. RTOFT instead measures the realized violation directly and restores exact orthogonality only when the measured drift reaches a prescribed tolerance.

4.1 Adaptation in the Tangent Space

When parameterized by the same tangent vector, exponential, Cayley, and polar retractions share the same first-order expansion around the identity (Absil et al., 2008; Boumal, 2023). RTOFT therefore uses their common tangent term between restorations:

$$W^+ = W(I + S), \qquad S \in \mathrm{Skew}(d). \tag{1}$$

Since $W(I + S)x = Wx + W(Sx)$, the inner step requires one structured multiplication and constructs no orthogonal matrix, matrix inverse, or iterative orthogonalization (Figure 2b; Appendix F).

4.2 Drift-Triggered Restoration

For skew S, $(I + S)^\top (I + S) = I - S^2$ and $\delta(I + S) = \|S\|_2^2$ (Theorem 1). At a restoration, RTOFT replaces $I + S$ by its nearest orthogonal matrix, the polar factor (Fan & Hoffman, 1955; Higham, 1986) $\mathrm{polar}(I+S) = (I+S)(I-S^2)^{-1/2}$, folds this factor into the accumulated adapted transform, and resets $S \leftarrow 0$. Equivalently, restored factors may be accumulated in an adapter while the pretrained weight remains frozen (Appendix G).

A fixed period restores according to elapsed optimization time; RTOFT instead restores when the measured drift reaches a tolerance τ. Under the block parameterization, $\|S\|_2 = \max_i \|S_i\|_2$, requiring only a $b \times b$ spectral-norm evaluation per block. Monitoring drift after each update and restoring before the next forward pass guarantees $\delta \le \tau^2$ independently of the training setup (Corollary 2). The restart interval adapts automatically to the rate at which drift accumulates. Algorithm 1 (Appendix B) gives the full training procedure.

4.3 Relation to the Cayley–Neumann Family

The tangent inner map itself appears in prior work. The $k = 2$ Cayley–Neumann truncation of OFTv2 (Qiu et al., 2025b) retains $I + 2Q$ and is equivalent to Equation 1 under $S = 2Q$; correspondingly, Lemma 14 gives defects $4\|Q\|_2^2$ and $\|S\|_2^2$, respectively. The distinction lies in when exact orthogonality is restored: RTOFT separates the inexpensive inner map from exact restoration and triggers the latter using the realized constraint violation.

The same restoration principle can be paired with higher-order Cayley–Neumann inner maps, which we test directly in Table 7. For the tangent parameterization, the tolerance τ directly specifies the forward-pass defect budget $\delta \le \tau^2$ without reference to a truncation order. The same restoration mechanism applies to the block-diagonal, subspace, and butterfly parameterizations considered here; Appendix B.12 discusses variants outside this form.

5 Exact Defect Control and Trajectory Tracking

A fixed restart period does not control how far an adaptive optimizer moves between restorations. RTOFT instead makes the realized orthogonality defect exactly observable, yielding a uniform bound that requires no assumption on the training dynamics.

Let $W^{\mathrm{pre}} \in \mathbb{R}^{d_{\mathrm{out}} \times d}$ be the pretrained weight and $W_k = W^{\mathrm{pre}} R_k$ the weight after the k-th restoration. Within a cycle, let $S^j \in \mathrm{Skew}(d)$ denote the tangent parameter after j inner steps, with

$S^0 = 0$, and define $A_k^j := R_k(I + S^j)$. Algorithm 2 gives the indexed training loop; proofs are in Appendix B.

Theorem 1 (Exact at restoration, quantified in between)**.** *At every restoration, including the returned model,*

$$W_k = W^{\mathrm{pre}} R_k, \quad R_k \in \mathcal{SO}(d); \qquad \|A^\top A - I\|_2 = \|S\|_2^2, \quad A = R_k(I + S)$$

between restorations. Hence restorations exactly preserve the row Gram matrix and singular spectrum of W^{pre}.

At a restored point, $S = 0$. In local coordinates the tangent space is $\mathrm{Skew}(d)$, corresponding to $R_k\mathrm{Skew}(d)$ at the current orthogonal transform. RTOFT follows this inexpensive linearized map between restorations and returns exactly to the group at the next restoration.

Corollary 2 (Uniform defect bound under a drift trigger)**.** *Fix* $\tau > 0$*, evaluate* $\|S\|_2$ *after every optimizer step, and restore before the next forward pass whenever* $\|S\|_2 \geq \tau$*. Then, independently of the loss, data, model, or optimizer, every forward pass satisfies*

$$\|A^\top A - I\|_2 \leq \tau^2, \qquad \|(I + S) - \mathrm{polar}(I + S)\|_2 = \sqrt{1 + \|S\|_2^2} - 1 \leq \tfrac{1}{2}\tau^2.$$

The corollary's one requirement is implementational. The polar factorization and the merge must be formed in single precision or better even when the surrounding model trains in half, as ours are; run in bfloat16 instead, the same restarts leave a residual twenty times the bound the tightest schedule guarantees (Assumption 3, measured in Appendix B). Under the regularity and cycle-length conditions of Lemma 9, a stronger result shows that each restoration cycle ends within $O(\tau^2)$ of the trajectory obtained by retracting after every optimization step. The bound reflects that the tangent map departs from the exponential flow at second order, while polar restoration rejoins it at third order (Proposition 8).

6 Experiments

We evaluate whether replacing per-step orthogonal construction with drift-triggered restoration preserves the empirical benefits of orthogonal finetuning while reducing its training cost. Experiments cover DeBERTaV3-base (He et al., 2023) on GLUE, Qwen3.5-9B on commonsense reasoning, Stable Diffusion (Rombach et al., 2022) on controllable generation, and Qwen2.5-7B under PEFT-Arena (Huang et al., 2026) for retention. Unless stated otherwise, RTOFT denotes the block variant with the drift trigger at $\tau = 0.10$; the fixed one-epoch period is reported as its matched schedule control. Within each benchmark, compared methods share one training setup and learning-rate search, and direct OFTv2 comparisons match block size and trainable parameter count. The main GLUE comparison re-runs every baseline in a common pipeline over five seeds. Across this section and its appendices, we report 860 finetuning runs from the same experimental pipeline; the GLUE grid alone accounts for 1,221 V100-hours (Table 8). Full training, evaluation, and runtime protocols are given in Appendix J.

6.1 Accuracy and Efficiency on GLUE

Matched OFTv2 comparison. At matched block size and trainable parameter count, RTOFT-B improves on OFTv2 on all six GLUE tasks while reducing mean wall-clock from $1.10\times$ to $0.97\times$ LoRA's (Table 2). Across the five shared seeds, the paired six-task improvement is 1.09 ± 0.25 points and is positive on every seed. The baseline runs the released default k=5; the cheaper k=3 setting matches that default's six-task average to $+0.03 \pm 0.05$ (Section 3). Relative to LoRA, the block variant runs at essentially the same wall-clock while improving the six-task average by 1.98 ± 0.39 points, again positive on all five shared seeds.

The fixed one-epoch period has essentially identical runtime (133:11 versus 133:07) but reaches 89.45 rather than 89.90. The paired advantage of the drift trigger is 0.45 ± 0.13 points and is positive on every seed. The two schedules have matched computational cost and differ only in when restoration is applied. The larger difference is geometric: one frozen τ realizes comparable geometric control wherever it is carried, while a fixed period realizes defects spanning a factor of twenty-two across settings at the same block size (Section 7.3); only the tolerance acts as a control variable for the constraint. The causal role of restoration and of the trigger itself is isolated in Section 7.

Table 2: GLUE test-split results on DeBERTaV3-base under the test-split protocol of Wu et al. (2026), every method re-run in one pipeline; cells are five-seed means with the seed deviation as subscript, Time is mean wall-clock per task on one V100 and Rel. Time is normalized to LoRA (seed deviations and protocol in Appendix J). Best in **bold**, second-best underlined; tied seconds unmarked.

	Efficiency					GLUE performance						
Method	#Param	Adapter	Mem. (GB)	Time ↓	Rel. Time ↓	CoLA	STS-B	RTE	MRPC	SST-2	QNLI	Avg.
Full finetuning	184M	184M	5.9	149:03	1.08×	$68.20_{3.56}$	$91.34_{0.40}$	$83.35_{2.10}$	$90.48_{1.28}$	$95.18_{0.71}$	$92.94_{0.35}$	86.92
Low-rank adaptation												
$\text{LoRA}_{r=8}$	1.33M	1.33M	4.5	137:54	1.00×	$71.98_{3.60}$	$91.42_{0.57}$	$84.53_{1.85}$	$90.63_{0.51}$	$95.01_{1.62}$	$93.92_{0.62}$	87.91
$\text{PiSSA}_{r=8}$	1.33M	1.33M	4.5	**127:10**	**0.92**×	$67.52_{2.48}$	$91.43_{1.16}$	$86.17_{1.45}$	$90.25_{2.42}$	$94.69_{1.47}$	$92.90_{0.86}$	87.16
$\text{DoRA}_{r=8}$	1.41M	1.41M	5.8	138:28	1.00×	$67.99_{1.15}$	$91.92_{0.94}$	$87.68_{3.26}$	$89.24_{1.75}$	$95.25_{2.31}$	$93.99_{0.37}$	87.68
Orthogonal finetuning												
$\text{PSOFT}_{r=46}$	**0.08M**	7.71M	**4.1**	169:52	1.23×	$71.60_{1.98}$	$91.42_{0.39}$	$85.75_{3.98}$	$90.53_{1.40}$	$95.18_{0.70}$	$93.60_{1.24}$	88.01
$\text{HRA}_{r=16}$	1.33M	1.33M	15.2	283:35	2.06×	$70.05_{1.69}$	$90.47_{1.01}$	$86.46_{0.82}$	$\mathbf{92.82}_{1.32}$	$94.71_{0.81}$	$93.15_{0.75}$	87.94
$\text{BOFT}_{b=8,m=2}$	1.41M	1.41M	6.3	218:01	1.58×	$68.85_{2.09}$	$91.85_{0.72}$	$86.78_{2.82}$	$89.91_{1.17}$	$95.33_{1.56}$	$93.85_{0.41}$	87.76
$\text{OFTv2}_{b=32}$ (k=5)	1.29M	**1.29M**	4.6	151:03	1.10×	$71.64_{1.53}$	$91.60_{0.46}$	$88.48_{2.01}$	$92.15_{3.90}$	$94.90_{0.75}$	$94.06_{1.00}$	88.81
$\text{RTOFT-B}_{b=32}$ (one-epoch period)	1.29M	**1.29M**	4.6	133:11	0.97×	$\underline{72.52}_{1.08}$	$\underline{92.30}_{0.36}$	$\underline{88.55}_{1.62}$	$\underline{92.78}_{0.84}$	$\underline{95.88}_{0.39}$	$\underline{94.65}_{0.37}$	$\underline{89.45}$
$\text{RTOFT-B}_{b=32}$ (ours)	1.29M	**1.29M**	4.6	<u>133:07</u>	0.97×	$\mathbf{73.55}_{0.84}$	$\mathbf{92.44}_{0.14}$	$\mathbf{89.20}_{0.92}$	$92.65_{0.31}$	$\mathbf{96.75}_{0.16}$	$\mathbf{94.80}_{0.14}$	**89.90**

Table 3: GLUE under the standard validation-score protocol on DeBERTaV3-base. Literature baselines are from Liu et al. (2024); OFTv2 and both RTOFT schedules are re-run over the same five seeds. Subscripts denote seed standard deviations; among the three re-run rows, the smallest standard deviation for each task is boldfaced in the subscript. Best performance is in **bold** and second-best is underlined.

Method	#Param (M)	Mem. (GB)	MNLI	SST-2	CoLA	QQP	QNLI	RTE	MRPC	STS-B	All
Full finetuning	184	N/A	89.90	95.63	69.19	**92.40**	94.03	83.75	89.46	91.60	88.25
BitFit	0.10	N/A	89.37	94.84	66.96	88.41	92.24	78.70	87.75	91.35	86.20
H-Adapter	1.22	N/A	90.13	95.53	68.64	91.91	94.11	84.48	89.95	91.48	88.28
P-Adapter	1.18	N/A	90.33	95.61	68.77	92.04	94.29	85.20	89.46	91.54	88.41
$\text{LoRA}_{r=8}$	1.33	N/A	<u>90.65</u>	94.95	69.82	91.99	93.87	85.20	89.95	91.60	88.50
AdaLoRA	1.27	N/A	**90.76**	96.10	71.45	<u>92.23</u>	**94.55**	88.09	90.69	91.84	89.46
$\text{BOFT}^{b=8}_{m=2}$	0.75	N/A	90.25	**96.44**	72.95	92.10	94.23	88.81	92.40	<u>91.92</u>	89.89
$\text{OFTv2}_{b=16}$	0.62	3.28	$90.38_{0.29}$	$95.95_{0.11}$	$73.47_{1.82}$	$92.06_{0.37}$	$94.11_{0.68}$	$88.70_{2.01}$	$\underline{92.75}_{1.21}$	$91.69_{0.28}$	$89.89_{0.85}$
$\text{RTOFT}_{b=16}$ (one-epoch period)	0.62	3.28	$90.40_{0.15}$	$96.10_{\mathbf{0.07}}$	$\underline{73.55}_{1.59}$	$91.96_{0.12}$	$\underline{94.30}_{0.52}$	$\underline{89.19}_{1.36}$	$92.21_{0.71}$	$91.89_{\mathbf{0.13}}$	$\underline{89.95}_{0.58}$
$\text{RTOFT}_{b=16}$ (ours)	0.62	3.28	$90.54_{\mathbf{0.13}}$	$\underline{96.30}_{\mathbf{0.07}}$	$\mathbf{74.65}_{\mathbf{1.43}}$	$92.05_{\mathbf{0.11}}$	$\mathbf{94.55}_{\mathbf{0.46}}$	$\mathbf{90.69}_{\mathbf{1.23}}$	$\mathbf{92.95}_{\mathbf{0.64}}$	$\mathbf{92.12}_{\mathbf{0.13}}$	$\mathbf{90.48}_{\mathbf{0.52}}$

Protocol robustness. The same ordering holds under the standard eight-task validation-score protocol used in prior PEFT comparisons. With literature baselines reported under that protocol and OFTv2 and both RTOFT schedules re-run over the same five seeds, the drift trigger reaches an average of 90.48, compared with 89.95 for the one-epoch period and 89.89 for OFTv2 (Table 3). The trigger improves on the period by a paired 0.53 ± 0.02 points, positive on every seed and every task.

Efficiency across parameterizations. The reduction in training cost holds across parameterizations. Using the same fixed one-epoch restoration schedule, the wall-clock premium over LoRA falls from $1.10\times$ to $0.97\times$ for the block parameterization, from $1.23\times$ to $1.03\times$ for the fixed-subspace parameterization, and from $1.58\times$ to $1.12\times$ for the butterfly parameterization (Tables 2, 12, and 13). HRA, whose Householder-chain construction is not covered by this tangent-and-restoration formulation, remains above twice LoRA's wall-clock (Tables 2 and 14).

The measured saving is consistent with removing the per-step orthogonal construction: the reference implementation constructs and differentiates through the transform at every optimization step, whereas RTOFT pays for exact restoration only intermittently. Detailed operation counts, kernel-level accounting, and sequence-length scaling are reported in Appendices F, D, H, and C.

6.2 Retention Under PEFT-Arena

PEFT-Arena (Huang et al., 2026) evaluates target adaptation together with retained general capability. At matched parameter count, RTOFT loses 2.38 points of general capability, compared with 2.60 for the PEFT-Arena OFT block-32 baseline, while achieving a larger target gain (Figure 2c and Table 16). Across three runs of each method, the weakest RTOFT retention result remains above the strongest of those OFT results. RTOFT therefore improves the target–retention trade-off.

The broader comparison shows the same trade-off. Methods that retain more general capability than RTOFT produce substantially smaller target gains, whereas the baselines that improve further on the target sacrifice considerably more general capability while using several times the parameter budget. Thus the reduction in per-step orthogonal construction preserves the retention behavior for which orthogonal finetuning is used.

Table 4: Commonsense reasoning with Qwen3.5-9B under the Commonsense-15K protocol (Appendix J). "Adapter" is the total stored adapter state, including frozen factors; time is end-to-end wall-clock training time in *mm:ss*, and accuracy cells are five-seed means with seed standard deviation as subscript. RTOFT-B uses $\tau = 0.10$; the one-epoch row differs only in its restoration schedule. Best parameter-efficient results are in **bold** and second-best are underlined; full finetuning and the no-restoration ablation are shown for reference.

	Efficiency				Commonsense reasoning								
Method	#Param	Adapter	Mem. (GB)	Time ↓	BoolQ	PIQA	SIQA	HellaSwag	WinoG.	ARC-e	ARC-c	OBQA	Avg.
Full finetuning	8.95B	8.95B	52.09	117:01	$68.11_{0.58}$	$83.25_{0.09}$	$74.81_{0.16}$	$84.86_{0.09}$	$74.94_{0.42}$	$92.26_{0.20}$	$82.01_{0.28}$	$85.87_{0.59}$	$80.76_{0.11}$
Low-rank adaptation													
LoRA$_{r=8}$	12.19M	12.19M	26.69	140:00	$73.09_{0.25}$	$90.53_{0.26}$	$78.40_{0.28}$	$94.34_{0.15}$	$76.48_{0.31}$	$98.32_{0.05}$	$93.77_{0.22}$	$93.00_{0.24}$	$87.24_{0.10}$
PiSSA$_{r=8}$	12.19M	12.19M	26.69	**133:58**	$72.11_{0.23}$	$91.41_{0.16}$	$\mathbf{79.38}_{0.25}$	$94.98_{0.10}$	$78.92_{0.61}$	$97.94_{0.18}$	$93.79_{0.27}$	$93.41_{0.43}$	$87.74_{0.18}$
DoRA$_{r=8}$	12.99M	12.99M	29.78	153:53	$72.75_{0.40}$	$90.59_{0.37}$	$78.20_{0.31}$	$94.44_{0.16}$	$76.56_{0.58}$	$98.27_{0.15}$	$93.77_{0.31}$	$93.00_{0.25}$	$87.20_{0.16}$
LoRA-XS$_{r=316}$	11.18M	492.68M	26.66	134:09	$72.97_{0.33}$	$90.32_{0.23}$	$79.17_{0.36}$	$94.55_{0.11}$	$\mathbf{79.95}_{0.31}$	$97.47_{0.16}$	$93.17_{0.12}$	$93.00_{0.23}$	$87.58_{0.08}$
Orthogonal finetuning													
HRA$_{r=16}$	11.53M	11.53M	OOM	N/A	N/A								N/A
OFTv2$_{b=32}$	**11.17M**	**11.17M**	26.95	174:50	$71.59_{0.51}$	$91.02_{0.26}$	$77.89_{0.31}$	$94.23_{0.09}$	$76.09_{0.36}$	$98.23_{0.16}$	$93.69_{0.41}$	$91.80_{0.57}$	$86.82_{0.18}$
RTOFT-B$_{b=32}$ (one-epoch period)	**11.17M**	**11.17M**	**26.58**	134:05	$68.38_{0.32}$	$89.77_{0.27}$	$78.71_{0.24}$	$94.92_{0.14}$	$78.93_{0.51}$	$98.06_{0.14}$	$93.69_{0.41}$	$\mathbf{95.40}_{0.48}$	$87.23_{0.22}$
RTOFT-B$_{b=32}$ (ours)	**11.17M**	**11.17M**	**26.58**	134:18	$\mathbf{73.61}_{0.22}$	$\mathbf{91.57}_{0.24}$	$79.27_{0.40}$	$\mathbf{95.03}_{0.15}$	$\mathbf{79.95}_{0.67}$	$\mathbf{98.34}_{0.23}$	$\mathbf{93.86}_{0.20}$	$\mathbf{95.40}_{0.50}$	$\mathbf{88.38}_{0.25}$
RTOFT-B$_{b=32}$ w/o restart	11.17M	11.17M	26.58	134:04	$71.65_{0.37}$	$88.84_{0.16}$	$77.85_{0.28}$	$93.94_{0.15}$	$77.47_{0.89}$	$97.52_{0.12}$	$92.74_{0.07}$	$94.74_{0.25}$	$86.84_{0.14}$

6.3 Transfer Across Scale and Modality

Qwen3.5-9B. The accuracy–efficiency improvement persists at nine-billion-parameter scale. Under the Commonsense-15K protocol, parameter-matched OFTv2 requires 174:50 of training time and reaches an eight-task average of 86.82, whereas RTOFT requires 134:18 and reaches 88.38 (Table 4). The fixed one-epoch control has essentially the same wall-clock (134:05) but reaches 87.23, with the gap to the trigger carried mainly by BoolQ (68.38 against 73.61): the periodic arm drifts toward the majority class, answering "true" on up to 90.9% of items against a 62.2% base rate and 66.3% for the tolerance arm (Appendix J.2).

Stable Diffusion. The result also transfers beyond language models. On Stable Diffusion, the one-epoch RTOFT control trains at 1.01× LoRA's wall-clock, faster than every orthogonal baseline in Table 5, while improving matched-parameter OFTv2 by 1.66 mIoU on segmentation-to-image. Replacing the fixed period with the same $\tau = 0.10$ drift trigger increases mIoU from 29.45 to 30.18, mAcc from 42.26 to 42.78, and aAcc from 67.97 to 69.19, with wall-clock changing only from 1.01× to 1.02× LoRA's. The trigger also gives the lowest landmark-to-face error, 5.31, compared with 5.36 for the fixed-period control. The same tolerance is therefore used unchanged across the language and diffusion experiments. Absolute runtimes, additional baselines, and qualitative samples are reported in Tables 17 and 15 and Appendix I.

Method	#P. (M)	Time	S2I ↑			L2F ↓
		×LoRA	mIoU	mAcc	aAcc	Error
LoRA$_{r=16}$	2.50	1.00	24.58	37.74	60.31	8.01
OFT$_{b=4}$	26.40	1.11	27.18	40.23	62.58	7.03
COFT$_{b=4}$	26.40	1.11	26.81	39.95	62.82	6.97
OFTv2$_{c=2}$	26.34	1.09	27.79	40.78	64.36	6.61
BOFT$^{m=4}_{bn=16}$	26.40	1.44	29.01	41.43	67.94	5.60
HRA$_{r=8,\lambda=10^{-3}}$	0.89	1.53	29.19	40.80	67.73	5.51
RTOFT-B$_{b=2}$ w/o restart	26.35	1.01	26.32	40.91	64.90	6.86
RTOFT-B$_{b=2}$ (one-epoch period)	26.35	1.01	29.45	42.26	67.97	5.36
RTOFT-B$_{b=2}$ (ours)	26.35	1.02	**30.18**	**42.78**	**69.19**	**5.31**

Table 5: Controllable generation with Stable Diffusion (S2I: ADE20K; L2F: CelebA-HQ). Each method is trained and timed at its best trainable-parameter budget over a 1–26M sweep; Time is S2I wall-clock relative to LoRA. RTOFT-B uses $\tau = 0.10$, and the one-epoch row differs only in its restoration schedule.

7 Why Restoration Is Needed, and When to Apply It

Corollary 6 converts a stretch budget ε into $\tau(\varepsilon) = \sqrt{(1+\varepsilon)^2 - 1}$. Setting $\varepsilon = 0.5\%$ gives $\tau = 0.10012$; we use $\tau = 0.10$, which guarantees that every forward-pass tangent factor stretches any direction by at most 0.499%, equivalently $\delta \leq 10^{-2}$.

7.1 The Necessity of Restoration

In five of the seven settings in Table 6, tangent adaptation without restoration falls below the method it replaces; landmark-to-face and commonsense are the exceptions. Restoration gives the best cell in

Table 6: **Restoration ablation across seven settings.** The restarted row uses the fixed one-epoch period throughout; the drift trigger is evaluated separately in the main experiments.

Variant	OFTv2			OFT		PSOFT	BOFT
	GLUE 6-task avg. ↑	Common. 8-task avg. ↑	Arena retention ↑	S2I mIoU ↑	L2F error ↓	Subspace 6-task avg. ↑	Butterfly 6-task avg. ↑
Reference	88.81	86.82	−2.60	27.18	7.03	88.01	87.76
Tangent only	87.77	86.84	−13.35	26.32	6.86	87.03	87.07
Restarted tangent (one-epoch period)	**89.45**	**87.23**	**−2.53**	**29.45**	**5.36**	**88.19**	**88.31**

all seven. On GLUE it adds 1.68 ± 0.37 points, positive on all five shared seeds; on commonsense it adds 0.39 points, while on PEFT-Arena it recovers 10.82 points of retained capability (Table 16).

The PEFT-Arena result also separates path maintenance from endpoint feasibility. A single polar projection after training recovers only 0.24 points of the 10.82-point retention loss, and projection along the no-restoration trajectory does not repair it either (Appendix H). Restoration thus works by limiting geometric departure during adaptation, beyond making the returned transform orthogonal.

The exception is a parameterization whose support the run itself relocates. On RTOFT-L, restoration changes the factor through which subsequent gradients move the support and reduces the six-task average by 0.35 ± 0.22 points. On the fixed-support parameterizations, resetting the tangent parameter does not erase the first-order gradient component within that support (Appendix G, Table 10). The analysis below therefore concerns fixed-support parameterizations.

7.2 Restoration Changes the Preferred Inner Map

The 2×2 intervention in Table 7 separates the inner map from restoration. Without restoration, the default $k = 5$ Cayley–Neumann map outperforms the tangent-equivalent $k = 2$ map by 1.04 points. With the same restoration schedule, the preference reverses: $k = 2$ leads by 0.86 points while using three fewer block multiplications per step.

Table 7: **Restoration reverses the preferred inner map.** Six-task GLUE averages over five seeds.

Inner map	No restoration	With restoration	Restoration gain
Tangent, k=2	87.77	**89.45**	+1.68
Cayley–Neumann, k=5	**88.81**	88.59	−0.22
Higher-order gain	+1.04	**−0.86**	–

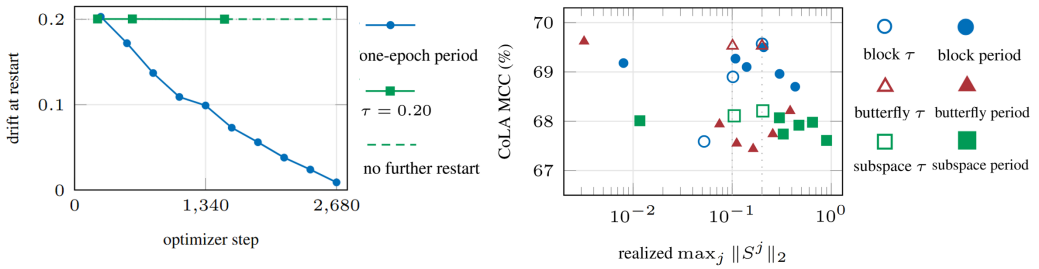
At the $\|Q\|_2$ reached on CoLA, the realized defect of the $k = 5$ map is $78\times$ smaller than that of $k = 2$ (Appendix F), leaving correspondingly less for restoration to correct. A higher-order inner map is therefore not intrinsically preferable once exactness is restored along the optimization path.

7.3 Drift Transfer Across Dynamics

A fixed period controls elapsed optimization time but not the geometry reached between restorations. A drift tolerance instead fixes the geometric budget while allowing the restoration interval to adapt to the run.

Figure 3a shows the distinction within a single run: as training slows, we observe the one-epoch period restoring at the same spacing while the drift it realizes falls by more than an order of magnitude. The tolerance instead keeps realized geometry near its prescribed budget and allows the interval to lengthen. Across the three fixed-support parameterizations, Figure 3b shows the same distinction: the drift realized after a fixed number of steps differs widely across the three parameterizations, while a frozen tolerance keeps it near the prescribed budget.

The distinction persists across models. With one $\tau = 0.10$, post-update monitored maxima remain near the same budget across DeBERTaV3, Qwen, and diffusion models while restart counts differ by orders of magnitude (Figure 2d). A fixed one-epoch period instead realizes defects spanning a factor of twenty-two across settings sharing block size $b = 32$ (Appendix H). Proposition 11 formalizes why the trigger count can depend on spectral path length relative to τ, whereas a period's count is fixed by the step budget.



(a) fixed period, adaptive drift

(b) three supports, one drift coordinate

Figure 3: **Realized drift provides a control variable that is common across settings. (a)** On CoLA, a fixed period keeps its spacing while realized drift falls from 0.203 to 0.009; a fixed tolerance instead maintains the geometric budget and lets restart intervals lengthen. **(b)** Across block, subspace, and butterfly parameterizations, fixed tolerances align realized drift whereas fixed periods do not. RTE reproduces the comparison in Appendix E.

The crossing step's post-update reading can exceed τ; it is restored before any forward pass uses it. On segmentation-to-image, for example, the $\tau = 0.10$ trigger fires 1,872 times over 27,820 steps—about once every fifteen steps—while every forward pass remains within the certified budget (Table 5).

Finally, a high-learning-rate stress test separates restoration placement from restoration count. From $0.5\times$ through $1.5\times$ the tuned learning rate, the fixed period and drift trigger do not separate on SST-2. At $2\times$, however, the trigger leads by 1.69 points, with every paired seed positive and the period's best seed below the trigger's worst (Appendix H). Per-step restoration matches the trigger, and giving the fixed-period control the trigger's restart count recovers less than half of the gap. The remaining factor is where restoration is placed along the realized drift trajectory.

Together, these interventions identify the transferable control variable. The stretch budget determines τ, and realized drift determines when restoration is required. A period or a truncation order leaves the realized defect unknown until the run ends, and the tolerance fixes it in advance.

8 Conclusion

RTOFT separates exact orthogonal restoration from per-step optimization. Tangent updates make constraint violation exactly observable through $\delta(I+S) = \|S\|_2^2$, allowing exact orthogonality to be restored when the measured drift reaches a prescribed tolerance rather than reconstructing an orthogonal transform at every iteration. Across fixed-support orthogonal adapters, this brings the family's training cost from 1.10–$1.58\times$ LoRA's wall-clock to 0.97–$1.12\times$ while preserving or improving target performance and retention. Matched controls show that removing restoration degrades retention, and restoring at every step gives no measurable benefit over restoring at the trigger.

This distinction also defines the scope of the method. Between restorations, the transform is only approximately orthogonal, so guarantees requiring exact feasibility at every iterate do not apply; when the adaptation support itself moves, resetting the tangent parameter can also alter the subsequent optimization signal. More broadly, when constraint violation is directly measurable, feasibility of intermediate iterates and feasibility of the returned solution can be treated as separate algorithmic design choices. Extending the schedule to supports that move during training, and to constraints whose defect is only estimated, is left open.

Reproducibility Statement

The construction is stated in full in Section 4: a first-order tangent step $W(I + S)$ with S skew, a polar restart that folds $\mathrm{polar}(I + S)$ into the frozen weight and resets S, and the drift trigger of Corollary 2. Assumptions and proofs are in Appendix B.

Every method in Table 2 is re-run in one pipeline over the same five seeds, on one machine, with wall-clock timed under one instrument; the seed-level spread of every method in the extended Table 14 are in Table 11, and the per-task protocol, including the learning-rate grid searched for each method, is in Appendix J. Wall-clock differences below one percent are treated as unresolved and no conclusion is drawn from them (Appendix D).

The truncation-order audit of Section 3 is computed on the reference implementation's released checkpoints, so it can be repeated without retraining. Where a drift tolerance is reported, we state how the drift was monitored: the thirty GLUE runs are read after every optimizer step and carry the per-forward-pass bound of Corollary 2. The one arm that is read every twenty steps is marked as such, and the value quoted for that arm is the largest reading rather than a certificate. Appendix H gives the codebase, the schedule sweep's selection rule, the per-task controls, and the high-learning-rate controls in full.

REFERENCES

Pierre Ablin and Gabriel Peyré. Fast and accurate optimization on the orthogonal manifold without retraction. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Pierre Ablin, Simon Vary, Bin Gao, and P.-A. Absil. Infeasible deterministic, stochastic, and variance-reduction algorithms for optimization under orthogonality constraints. *Journal of Machine Learning Research*, 25(389):1–38, 2024.

P.-A. Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters. In *ECAI 2025: 28th European Conference on Artificial Intelligence*, pp. 3194–3201, 2025. arXiv:2405.17604.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, 1997.

Massimo Bini, Leander Girrbach, and Zeynep Akata. Decoupling angles and strength in low-rank adaptation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2503.18225.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, Cambridge, 2023. ISBN 978-1-009-16615-7. doi: 10.1017/9781009166164.

Ky Fan and Alan J. Hoffman. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6(1):111–116, 1955.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2111.09543.

Desmond J. Higham. Time-stepping and preserving orthonormality. *BIT Numerical Mathematics*, 37(1):24–36, 1997.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

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 (ICLR)*, 2022. arXiv:2106.09685.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. arXiv:2304.01933.

Yangyi Huang, Ruotian Peng, Zeju Qiu, Jiale Kang, Yandong Wen, Bernhard Schölkopf, and Weiyang Liu. PEFT-Arena: Understanding parameter-efficient finetuning from a stability-plasticity perspective. *arXiv preprint arXiv:2605.28819*, 2026.

Adrián Javaloy and Antonio Vergari. An embarrassingly simple way to optimize orthogonal matrices at scale. In *International Conference on Machine Learning*, 2026.

Kui Jia, Dacheng Tao, Shenghua Gao, and Xiangmin Xu. Improving training of deep neural networks via singular value bounding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. MuonBP: Faster Muon via block-periodic orthogonalization. In *International Conference on Learning Representations (ICLR)*, 2026.

Mario Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1909.09501.

Vijay Lingam, Atula Tejaswi, Aditya Vavre, Aneesh Shetty, Gautham Krishna Gudur, Joydeep Ghosh, Alex Dimakis, Eunsol Choi, Aleksandar Bojchevski, and Sujay Sanghavi. SVFT: Parameter-efficient fine-tuning with singular vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.19597.

Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. Parameter-efficient orthogonal finetuning via butterfly factorization. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.06243.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226. arXiv:2302.08453.

An Nguyen, Jaesik Choi, and Anh Tong. LoCO: Low-rank compositional rotation fine-tuning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2026. arXiv:2605.15916.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. arXiv:2306.07280.

Zeju Qiu, Simon Buchholz, Tim Z. Xiao, Maximilian Dax, Bernhard Schölkopf, and Weiyang Liu. Reparameterized LLM training via orthogonal equivalence transformation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. arXiv:2506.08001.

Zeju Qiu, Weiyang Liu, Adrian Weller, and Bernhard Schölkopf. Orthogonal finetuning made scalable. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31946–31963, Suzhou, China, November 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1627. arXiv:2506.19847.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2112.10752.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023. arXiv:2208.12242.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning (ICML)*, 2017.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Efficient orthogonal fine-tuning with principal subspace adaptation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.11235.

Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Advancing parameter efficiency in fine-tuning via representation editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13445–13464, 2024a. arXiv:2402.15179.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024b. arXiv:2404.03592.

Shen Yuan, Haotian Liu, and Hongteng Xu. Bridging the gap between low-rank and orthogonal adaptation via householder reflection adaptation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2405.17484.

Vasily Zadorozhnyy, Edison Mucllari, Cole Pospisil, Duc Nguyen, and Qiang Ye. Orthogonal gated recurrent unit with Neumann–Cayley transformation. *Neural Computation*, 36(12), 2024. arXiv:2208.06496.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023a. arXiv:2302.05543.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023b. arXiv:2303.10512.

Lanxin Zhao, Bamdev Mishra, Pratik Jawanpuria, Lequan Lin, Dai Shi, Junbin Gao, and Andi Han. LOFT: Low-rank orthogonal fine-tuning via task-aware support selection. In *Advances in Neural Information Processing Systems 39 (NeurIPS 2026)*, 2026.

AI USE STATEMENT

Two tools were used. Claude (Anthropic) helped with LaTeX and checked and fixed the mathematics, reading the derivations and proofs for mistakes and correcting them; ChatGPT (OpenAI) helped with formatting and commented on wording. Every claim, proof, experimental result and conclusion was then verified by the authors, who are responsible for the entire paper.

Algorithm 1 RTOFT: tangent orthogonal finetuning with restart, at a fixed period M; the shipped drift trigger of Corollary 2 replaces the test "M divides t" by $\|S_t\|_2 \geq \tau$

Require: pretrained weights W^{pre}, step size $\eta > 0$, restart period $M \in \mathbb{N} \cup \{\infty\}$, iterations T.

Initialize $W_0 \leftarrow W^{\mathrm{pre}}$, $S_0 \leftarrow 0$.
for $t = 1, \ldots, T$ **do**
  $S_t \leftarrow S_{t-1} - \eta\,\mathrm{skew}\big(\nabla_S \mathcal{L}\big(W_{t-1}(I + S_{t-1})\big)\big)$
  if M divides t **then**
    $W_t \leftarrow W_{t-1}\mathrm{polar}(I + S_t)$, $S_t \leftarrow 0$ {restart}
  else
    $W_t \leftarrow W_{t-1}$
  end if
end for
return $W_T\,\mathrm{polar}(I + S_T)$

Algorithm 2 RTOFT in double-loop form

Require: pretrained weights W^{pre}, step size $\eta > 0$, restart period M, cycles K.

Initialize $W_0 \leftarrow W^{\mathrm{pre}}$.
for $k = 0, 1, \ldots, K-1$ **do**
  $S_k^0 \leftarrow 0$
  for $j = 0, 1, \ldots, M-1$ **do**
    $S_k^{j+1} \leftarrow S_k^j - \eta\,\mathrm{skew}\Big(\nabla_S \mathcal{L}\big(W_k(I + S_k^j)\big)\Big)$
  end for
  $W_{k+1} \leftarrow W_k\,\mathrm{polar}(I + S_k^M)$
end for
return W_K

A Related Work

Existing parameterizations vary in what is built and never in when: each evaluates its transformation, and the backward pass through it, at every training step (Table 9). OFT (Qiu et al., 2023) maps block-diagonal skew-symmetric parameters onto the group through the Cayley transform; OFTv2 (Qiu et al., 2025b) replaces the matrix inverse by a truncated Neumann expansion, cheaper at the price of a nonzero orthogonality defect; HRA (Yuan et al., 2024) chains r Householder reflections, exactly orthogonal but evaluated sequentially, its regularization strength λ trading adaptation capacity against regularity; BOFT (Liu et al., 2024) composes sparse butterfly factors into dense transformations with few trainable parameters; PSOFT (Wu et al., 2026) restricts adaptation to an r-dimensional subspace, $R = I + P_r^\top (R_r - I) P_r$, and LOFT (Zhao et al., 2026) selects that subspace in a task-aware manner and freezes it.

The individual technical ingredients are familiar. Dynamic trivialization restarts a local chart at the current manifold iterate, while every update within that chart still evaluates an exact map onto the group (Lezcano-Casado, 2019); merge-and-reinitialize strategies in orthogonal model training measure the realized error but merge on a preset step count (Qiu et al., 2025a); MuonBP amortizes the orthogonalization of optimizer updates (Khaled et al., 2026); Landing-style infeasible-iterate methods allow off-manifold steps with bounded deviation (Ablin & Peyré, 2022; Ablin et al., 2024), while POGO (Javaloy & Vergari, 2026) addresses that lack of restoration; periodic restoration of near-orthogonal weights predates PEFT (Higham, 1997; Jia et al., 2017), and Neumann-series approximations to Cayley maps also appear in orthogonal recurrent-network parameterizations (Zadorozhnyy et al., 2024). The closest PEFT relative, LoCO (Nguyen et al., 2026), composes rotations through a parallel approximation whose deviation from the group is bounded a priori. None of these methods monitors the adapter's realized orthogonality defect and uses it as the signal for scheduling restoration.

B Assumptions and Proofs

B.1 Decomposition of Theorem 1, and the polar restart

The theorem combines the following two propositions.

Proposition 3 (Exactness at restart)**.** *In exact arithmetic,*

$$W_k = W^{\mathrm{pre}} R_k, \qquad R_k \in \mathcal{SO}(d)$$

for every restart k. Consequently,

$$W_k W_k^\top = W^{\mathrm{pre}} (W^{\mathrm{pre}})^\top,$$

and the restart leaves every singular value, row norm, and pairwise row inner product of the pretrained weight untouched, hence every pairwise row angle wherever it is defined.

Thus the geometric preservation enjoyed by OFT holds verbatim at every RTOFT restart and at the final returned model: the returned weight lies exactly on the pretrained orthogonal orbit.

Between restarts, the deviation can be characterized without approximation.

Proposition 4 (In-flight geometry)**.** *Let $S \in \mathrm{Skew}(d)$ and $W_k = W^{\mathrm{pre}} R_k$ with $R_k \in \mathcal{SO}(d)$. Then:*

(i)

$$(I+S)^\top (I+S) = I - S^2,$$

and therefore, for $A = R_k(I+S)$,

$$\|A^\top A - I\|_2 = \|S\|_2^2.$$

Moreover,

$$\mathrm{dist}_2(I+S, \mathcal{O}(d)) = \sqrt{1+\|S\|_2^2} - 1,$$

with the minimum attained by $\mathrm{polar}(I+S)$*, and*

$$\big\| W_k (I+S)(I+S)^\top W_k^\top - W^{\mathrm{pre}} (W^{\mathrm{pre}})^\top \big\|_2 \le \sigma_{\max}(W^{\mathrm{pre}})^2 \|S\|_2^2.$$

(ii) For every singular-value index i,

$$\sigma_i(W^{\mathrm{pre}}) \le \sigma_i\big(W_k(I+S)\big) \le \sqrt{1+\|S\|_2^2}\, \sigma_i(W^{\mathrm{pre}}).$$

For every row vector w_i of W_k,

$$\|w_i(I+S)\|_2^2 = \|w_i\|_2^2 + \|w_i S\|_2^2.$$

If W^{pre} has full column rank, then

$$\kappa_2\big(W_k(I+S)\big) \le \sqrt{1+\|S\|_2^2}\, \kappa_2(W^{\mathrm{pre}}).$$

Proposition 4 makes the role of restart explicit. The quantity reset to zero is exactly $\|S\|_2^2$. It also shows what the result does *not* imply. Although the ordered singular values and row norms cannot decrease, individual input directions need not be amplified: when $(W^{\mathrm{pre}})^\top W^{\mathrm{pre}}$ is anisotropic, $\|W_k(I+S)x\|$ may still be smaller than $\|W^{\mathrm{pre}}x\|$ for particular x.

Corollary 5 (The tolerance is a stretch budget)**.** *For any transform A, $\sigma_{\max}(A)^2 = \|A^\top A\|_2 \le 1 + \|A^\top A - I\|_2$, so a defect δ admits a stretch of at most $\sqrt{1+\delta}$. The singular values of $I+S$ are $\sqrt{1+\theta_j^2}$ over the eigenvalues $\pm i\theta_j$ of S (Lemma 10), so under the trigger above every singular value of the tangent factor lies in $[1, \sqrt{1+\tau^2}\,]$ at every forward pass, and each restart returns all of them to exactly one.*

Corollary 6 (The default is the tolerance a stretch budget derives)**.** *Fix a stretch budget ε on the tangent factor: no forward pass may apply an $I+S$ that scales any direction by more than $1+\varepsilon$. By Corollary 5 the largest factor the tangent map applies is $\sqrt{1+\tau^2}$, so the largest admissible tolerance is*

$$\tau(\varepsilon) = \sqrt{(1+\varepsilon)^2 - 1} = \sqrt{2\varepsilon + \varepsilon^2}.$$

At the half-percent budget $\varepsilon = 1/200$ this gives $\tau = 0.10012$; the shipped default $\tau = 0.10$ is that number, and conversely certifies a stretch of at most 0.499%. No schedule enters the statement: restarts return every singular value to one, so the budget is per pass and does not compound across cycles.

Corollary 7 (The drift bounds how far the restart moves the frame)**.** *Let $Q = \mathrm{polar}(I + S)$ for any $S \in \mathrm{Skew}(d)$. Then Q rotates each eigenplane of S by $\arctan\theta_j$, so*

$$\|Q - I\|_2 = 2\sin\!\big(\tfrac{1}{2}\arctan\|S\|_2\big) \;\le\; \arctan\|S\|_2 \;\le\; \|S\|_2.$$

The tangent gradient at the restarted point is $\mathrm{skew}(Q^\top W^\top \nabla_W \mathcal{L})$ against $\mathrm{skew}(W^\top \nabla_W \mathcal{L})$ before it; since skew is 1-Lipschitz in the spectral norm, at a fixed loss gradient the tangent gradient the carried optimizer moments track changes by at most $\|Q - I\|_2 \, \|W^\top \nabla_W \mathcal{L}\|_2 \le \arctan\|S\|_2 \, \|W^\top \nabla_W \mathcal{L}\|_2$, evaluated at the drift the restart clears rather than at the tolerance that called it.

B.2 Accuracy of the Polar Restart

The polar restart is the nearest orthogonal matrix to $I + S$ in Frobenius norm (Fan & Hoffman, 1955; Higham, 1986). It also agrees closely with the exact exponential flow generated by the same skew-symmetric parameter.

Proposition 8 (Third-order agreement of the restart)**.** *For every $S \in \mathrm{Skew}(d)$,*

$$\big\|\mathrm{polar}(I + S) - \exp(S)\big\|_2 \le \frac{1}{3}\|S\|_2^3,$$

and

$$\big\|(I + S) - \exp(S)\big\|_2 \le \frac{1}{2}\|S\|_2^2 + \frac{1}{3}\|S\|_2^3.$$

The constant $1/3$ in the first inequality is sharp. For comparison, the Cayley retraction

$$\Omega \mapsto (I - \Omega/2)^{-1}(I + \Omega/2)$$

admits the corresponding third-order constant $1/12$.

The proposition separates the two approximations in RTOFT. The inner map $I + S$ differs from the exponential at second order, whereas the polar restart agrees with the same exponential to third order. The Cayley retraction has the smaller constant; the polar factor gives the nearest orthogonal restart in closed form and is paid for only periodically.

B.3 Standing assumptions

We use three structural or implementational assumptions; the loss, the gradients, the optimizer and the realized trajectory are unconstrained.

Assumption 1 (Layout and placement)**.** *$W \in \mathbb{R}^{d_{\mathrm{out}} \times d}$ acts as $y = Wx$, so neurons are the rows of W, and the adapter multiplies on the right, $W \mapsto W(I + S)$ with $S \in \mathrm{Skew}(d) = \{A \in \mathbb{R}^{d\times d} : A^\top = -A\}$; equivalently, the orthogonal factor acts on the contracted index.*

This is OFT's placement transposed: OFT writes $z = W^\top x = (RW^0)^\top x$ with column-neurons, which is Assumption 1 after transposition. The assumption is load-bearing. With column-neurons and right multiplication $WW^\top$ is still preserved, but the Gram matrix of the neurons is $W^\top W$, which right multiplication does not preserve, so the neuron norms and angles of Proposition 3 are lost.

Assumption 2 (Full column rank)**.** *$\mathrm{rank}(W^{\mathrm{pre}}) = d$, i.e. $d_{\mathrm{out}} \ge d$ and $\sigma_d(W^{\mathrm{pre}}) > 0$.*

Used *only* for the rank and condition-number clauses of Proposition 4(ii). The singular-value sandwich and the neuron-norm identity hold without it. Down-projections, for which $d_{\mathrm{out}} < d$, violate it: there the sandwich still applies to the $\min(d, d_{\mathrm{out}})$ nonzero singular values, but the condition-number clause is vacuous, since such a layer is rank-deficient before any adaptation.

Assumption 3 (Exact restart arithmetic)**.** *At each restart, $\mathrm{polar}(I + S^M)$ is formed in closed form (blockwise via the eigendecomposition of the skew block, or equivalently by an SVD) and composed into the anchor in at least single precision, so that the restart is exact up to the unit roundoff of that precision.*

Assumption 3 is what separates the word "exactly" in Proposition 3 from an asymptotic statement: an iterative approximation of the polar factor would reintroduce the defect of Lemma 14, and repeated merging in half precision does worse than accumulate a drift of the same order as the in-flight gap it is meant to remove: it swamps it. Composing the 46 merges of the τ=0.05 schedule at $d = 768$ and $b = 32$, each folding a factor with $\|S\|_2 = 0.05$, and measuring the deviation of the accumulated transform from orthogonality gives 1.5×10^{-5} in float32 and, to the digit, the same in float64: the residual is the intrinsic error of a single float32 polar factorization rather than anything the composition accumulates, and it sits a factor of 170 below the $\tau^2 = 2.5 \times 10^{-3}$ the schedule guarantees at its tightest setting. In bfloat16 the same composition gives 5.1×10^{-2}, twenty times the bound, and in float16 6.0×10^{-3}, still over it. The practical form of Assumption 3 is therefore narrow and worth stating: the polar factorization and the merge must be carried out in single precision or better even when the surrounding model trains in half, which is what our implementation does.

The displacement bound the identity replaces. A per-inner-step displacement bound $\|S^{j+1} - S^j\|_F \le \eta D$ gives $\|S^j\|_2 \le \eta M D$ uniformly and a restart period rule $M \le \sqrt{\varepsilon}/(\eta D)$. Such a bound is unnecessary, because Proposition 4(i) makes the gap an exactly computable function of S^j rather than something to be bounded. The correct practice is to monitor $\max_{j \le M} \|S^j\|_2$, which is not the same as its end-of-epoch value. The feasibility guarantee needs no such rule at any point. Where a rule of this kind does reappear, deliberately, is Lemma 9: certifying that a cycle *tracks* per-step retraction, a stronger statement than feasibility, costs the cycle-length condition $\eta G M \le 2\tau$, and the lemma states it as a hypothesis.

B.4 One cycle of the schedule tracks the retracted flow

Corollary 2 certifies feasibility and asks nothing of the loss. The following lemma asserts more, that a cycle of tangent updates *tracks* per-step polar-retracted optimization, and the stronger conclusion has a price, stated as a hypothesis: the cycle must be short relative to the gradient scale, $\eta G M \le 2\tau$. The condition is necessary in general: when the skew part of the gradient is small against G, the drift trigger fires late and the discrepancy below is genuinely first order in τ. It holds automatically whenever drift consumes its budget at the worst-case rate, and can be enforced outright by capping the period at $2\tau/(\eta G)$.

Lemma 9 (One-cycle tracking)**.** *Let F satisfy $\|\nabla F\|_2 \le G$ with L-Lipschitz gradient on a convex neighborhood of $\mathcal{O}(d)$ containing all iterates below. Fix $R \in \mathcal{O}(d)$, a step size η, and a cycle of M steps, and compare, from the common start,* (a) *the scheduled tangent dynamics $S_0 = 0$,*

$$S_{t+1} = S_t - \eta\,\mathrm{skew}\big(R^\top \nabla F(R(I + S_t))\big), \qquad R' = R\,\mathrm{polar}(I + S_M),$$

and (b) *the per-step polar-retracted dynamics $Q_0 = R$,*

$$Q_{t+1} = Q_t\,\mathrm{polar}\big(I - \eta\,H_t\big), \qquad H_t = \mathrm{skew}\big(Q_t^\top \nabla F(Q_t)\big).$$

Suppose $\|S_t\|_2 \le \tau$ throughout, with $\eta G \le \tau \le \min\{\frac{1}{2},\, G/(L + 2G)\}$ and $\eta G M \le 2\tau$. Then

$$\big\|R\,\mathrm{polar}(I + S_M) - Q_M\big\|_2 \;\le\; 8\,\tau^2\, e^{(2G+L)M\eta} \;\le\; 8e^2\tau^2.$$

The two dynamics agree to first order in the drift; what separates them is second order, and the budget bounds it.

Proof. Write $A_t = R(I + S_t)$, $U_t = \mathrm{polar}(I + S_t)$, $\tilde{Q}_t = RU_t$, $s_t = \|S_t\|_2$, and let $\mathcal{M}(Q) = Q\,\mathrm{skew}(Q^\top \nabla F(Q))$ denote the retracted dynamics' displacement field. Four facts are used throughout. *(F1)* skew is linear and 1-Lipschitz in the spectral norm, and $\mathrm{skew}(0) = 0$. *(F2)* By Lemma 10, $\|(I + X) - \mathrm{polar}(I + X)\|_2 = \sqrt{1 + \|X\|_2^2} - 1 \le \|X\|_2^2/2$ for skew X: applied to $X = -\eta H_t$, and using that left multiplication by the orthogonal Q_t preserves the norm, it bounds the retraction remainder of one step of (b) by $\eta^2 G^2/2$; applied to $X = S_t$ it gives $\|A_t - \tilde{Q}_t\|_2 \le s_t^2/2$. *(F3)* The eigenvalues of U_t are $e^{i \arctan\theta_j}$ over the eigenvalues $\pm i\theta_j$ of S_t (Lemma 10), so $\|U_t - I\|_2 \le \max_j \arctan\theta_j \le s_t$. *(F4)* On the group, $\mathcal{M}$ is Lipschitz with constant at most $2G + L$: telescoping its three dependencies on Q, the outer factor and the transpose inside contribute at most G each by (F1) and $\|Q\|_2 = 1$, and the argument of ∇F contributes at most L.

Step growth. By (F1), $\|S_{t+1}\|_2 \le \|S_t\|_2 + \eta G$, so $s_t \le \min\{t\eta G, \tau\}$.

One-step recursion. Let $E_t = \|A_t - Q_t\|_2$, $E_0 = 0$. By the two update rules and (F2),

$$E_{t+1} \le E_t + \eta \left\| R\,\mathrm{skew}(R^\top \nabla F(A_t)) - \mathcal{M}(Q_t) \right\|_2 + \tfrac{1}{2}\eta^2 G^2.$$

Insert $\mathcal{M}(\tilde{Q}_t)$. For the first half, with $g = R^\top \nabla F(\tilde{Q}_t)$, $\|g\|_2 \le G$, and $U_t = I + \Delta$, $\|\Delta\|_2 \le s_t$ by (F3), linearity of skew gives the exact expansion

$$\mathcal{M}(\tilde{Q}_t) = R\big[\mathrm{skew}(g) + \mathrm{skew}(\Delta^\top g) + \Delta\,\mathrm{skew}(g) + \Delta\,\mathrm{skew}(\Delta^\top g)\big],$$

so, adding and subtracting $\nabla F(A_t)$ in the first slot and using (F1), (F2),

$$\left\| R\,\mathrm{skew}(R^\top \nabla F(A_t)) - \mathcal{M}(\tilde{Q}_t) \right\|_2 \le L\,\tfrac{s_t^2}{2} + 2G\,s_t + G\,s_t^2.$$

For the second half, (F4) and (F2) give $\|\mathcal{M}(\tilde{Q}_t) - \mathcal{M}(Q_t)\|_2 \le (2G + L)(E_t + \frac{s_t^2}{2})$. Combining, with $C_2 = L + 2G$,

$$E_{t+1} \le \big(1 + \eta(2G + L)\big)E_t + \eta\big(2G\,s_t + C_2\,s_t^2\big) + \tfrac{1}{2}\eta^2 G^2.$$

Unrolling. Discrete Grönwall gives $E_M \le e^{(2G+L)M\eta}\,\Sigma$ with $\Sigma = \sum_{t<M}\big[\eta(2Gs_t + C_2 s_t^2) + \frac{1}{2}\eta^2 G^2\big]$. The restart enters here: because the cycle starts at $S_0 = 0$, the frame misalignment $2Gs_t$, first order pointwise, integrates to second order,

$$\sum_{t<M} \eta\,2G\,s_t \le 2G\eta \cdot \eta G\,\tfrac{M^2}{2} = (\eta G M)^2 \le 4\tau^2$$

by $s_t \le t\eta G$ and the hypothesis $\eta G M \le 2\tau$. The remaining sums are lower order: $\sum_{t<M} \eta C_2 s_t^2 \le C_2\tau^2\,\eta M \le 2C_2\tau^3/G \le 2\tau^2$ by $\tau \le G/C_2$, and $\frac{1}{2}M\eta^2G^2 = \frac{1}{2}(\eta G M)(\eta G) \le \tau^2$ by $\eta G \le \tau$. Hence $E_M \le 7\tau^2 e^{(2G+L)M\eta}$.

Cycle-end restoration. By (F2), $\|R\,\mathrm{polar}(I+S_M) - A_M\|_2 \le \tau^2/2$, so $\|R' - Q_M\|_2 \le E_M + \frac{1}{2}\tau^2 \le 8\tau^2 e^{(2G+L)M\eta}$. The absolute form follows from $(2G + L)M\eta \le (2G + L)\,2\tau/G \le 2$ under $\tau \le G/(L + 2G)$. □ □

B.5 Spectral data of $I + S$

Lemma 10. *Let $S \in \mathrm{Skew}(d)$ have nonzero eigenvalues $\pm i\theta_1, \ldots, \pm i\theta_m$ with $\theta_j > 0$ and a zero eigenvalue of multiplicity $d - 2m$. Then*

(a) $(I + S)^\top (I + S) = (I + S)(I + S)^\top = I - S^2 = I + S^\top S \succeq I$;

(b) the singular values of $I + S$ are $\sqrt{1 + \theta_j^2}$ (each twice) and 1 with multiplicity $d - 2m$; in particular $\|I + S\|_2 = \sqrt{1 + \|S\|_2^2}$, $\sigma_{\min}(I + S) \ge 1$, and $I + S$ is invertible with $\|(I + S)^{-1}\|_2 \le 1$;

(c) $\det(I + S) = \prod_{j=1}^m (1 + \theta_j^2) > 0$, *and $I - S$ is invertible;*

(d) $\mathrm{polar}(I + S) = (I + S)(I - S^2)^{-1/2}$ *is a function of S, hence normal and simultaneously unitarily diagonalizable with S and $\exp(S)$, and it maps the eigenvector of S for $i\theta$ to eigenvalue $e^{i\arctan\theta}$.*

Proof. S is real skew, hence normal, hence unitarily diagonalizable with purely imaginary eigenvalues, and $S^\top = -S$ gives $(I + S)^\top (I + S) = I + S + S^\top + S^\top S = I + S^\top S = I - S^2$, which proves (a). The same computation with the factors reversed gives the second equality. Eigenvalue-wise, $I - S^2$ has eigenvalue $1 + \theta^2$ on the eigenvector of S for $i\theta$, so the singular values of $I + S$ are $\sqrt{1 + \theta^2}$, which is (b). The largest is $\sqrt{1 + \max_j \theta_j^2} = \sqrt{1 + \|S\|_2^2}$ because $\|S\|_2 = \max_j \theta_j$. For (c), $\det(I + S) = \prod_j (1 + i\theta_j)(1 - i\theta_j) \cdot 1^{d-2m} = \prod_j (1 + \theta_j^2) > 0$, and $I - S$ is invertible because its eigenvalues $1 \mp i\theta$ are nonzero. For (d), $(I + S)^\top (I + S) = I - S^2 \succ 0$, so the polar factor is $(I + S)\big((I + S)^\top (I + S)\big)^{-1/2} = (I + S)(I - S^2)^{-1/2}$. Both factors are functions of S and therefore commute with it and with each other, and eigenvalue-wise the map is $i\theta \mapsto (1 + i\theta)/\sqrt{1 + \theta^2} = e^{i\arctan\theta}$. □

B.6 Proof of Proposition 3

Proof. Induction on k. For $k = 0$, $W_0 = W^{\mathrm{pre}}$ and $R_0 = I \in \mathcal{SO}(d)$. Assume $W_k = W^{\mathrm{pre}} R_k$ with $R_k \in \mathcal{SO}(d)$ and let $Q_k = \mathrm{polar}(I + S^M)$. Then Q_k is orthogonal by construction, and by Lemma 10(c), $\det Q_k = \det(I + S^M)/|\det(I + S^M)| = +1$, so $Q_k \in \mathcal{SO}(d)$. Hence $W_{k+1} = W_k Q_k = W^{\mathrm{pre}}(R_k Q_k)$ with $R_{k+1} = R_k Q_k \in \mathcal{SO}(d)$. Finally $W_k W_k^\top = W^{\mathrm{pre}} R_k R_k^\top (W^{\mathrm{pre}})^\top = W^{\mathrm{pre}}(W^{\mathrm{pre}})^\top$. Under Assumption 3 the two displayed equalities are exact up to the unit roundoff of the arithmetic used.

The consequences quoted in the main text are the known invariance of the row Gram matrix under right orthogonal multiplication (Qiu et al., 2023): writing w_i for the i-th neuron (the i-th row of W^{pre} transposed), the i-th neuron of $W^{\mathrm{pre}} R$ is $R^\top w_i$, so $\|R^\top w_i\| = \|w_i\|$, $\langle R^\top w_i, R^\top w_j\rangle = \langle w_i, w_j\rangle$, the normalized neurons satisfy $\widehat{R^\top w_i} = R^\top \hat{w}_i$ and hence $\|\widehat{R^\top w_i} - \widehat{R^\top w_j}\| = \|\hat{w}_i - \hat{w}_j\|$. Summing the last identity over $i \neq j$ leaves the hyperspherical energy $\mathrm{HE}(W) = \sum_{i\neq j} \|\hat{w}_i - \hat{w}_j\|^{-1}$ (Liu et al., 2018) unchanged. Equality of the singular values follows from equality of $WW^\top$. □

B.7 Proof of Proposition 4

Proof. *(i)* The identity $(I + S)^\top (I + S) = I - S^2$ is Lemma 10(a). For the cumulative adapter $A = R_k(I + S)$ with R_k orthogonal, $A^\top A - I = (I + S)^\top R_k^\top R_k (I + S) - I = -S^2$, so $\|A^\top A - I\|_2 = \|S^2\|_2 = \|S\|_2^2$, the last step because S is normal. Note that this is where exactness of the restart is load-bearing: the cumulative gap equals the gap of the current chart alone, with no cross-epoch accumulation, and it is 0 at $S = 0$.

For the distance, it is known that for nonsingular X and any unitarily invariant norm the minimum of $\|X - R\|$ over orthogonal R is attained at the orthogonal polar factor of X and equals $\|\Sigma - I\|$ for Σ the diagonal of singular values (Fan & Hoffman, 1955; Higham, 1986). By Lemma 10(b) all singular values of $I+S$ are at least 1, so in the spectral norm the minimum is $\max_i |\sigma_i(I+S)-1| = \sqrt{1 + \|S\|_2^2} - 1$, and by Lemma 10(c) the minimizer lies in $\mathcal{SO}(d)$.

At the weight level, with $W_t = W_k(I + S) = W^{\mathrm{pre}} R_k (I + S)$,

$$W_t W_t^\top - W^{\mathrm{pre}}(W^{\mathrm{pre}})^\top = W^{\mathrm{pre}} R_k \big[(I+S)(I+S)^\top - I\big] R_k^\top (W^{\mathrm{pre}})^\top = -\,W^{\mathrm{pre}} R_k S^2 R_k^\top (W^{\mathrm{pre}})^\top,$$

whose spectral norm is at most $\|W^{\mathrm{pre}}\|_2^2 \|S^2\|_2 = \sigma_{\max}(W^{\mathrm{pre}})^2 \|S\|_2^2$.

(ii) Upper bound: by the multiplicative singular-value inequality $\sigma_{i+j-1}(AB) \le \sigma_i(A)\sigma_j(B)$ at $j = 1$ (Horn & Johnson, 1991; Bhatia, 1997), $\sigma_i(W_k(I+S)) \le \sigma_i(W_k)\|I+S\|_2 = \sqrt{1 + \|S\|_2^2}\,\sigma_i(W_k)$. Lower bound: $I + S$ is invertible by Lemma 10(b), so applying the same inequality to $W_k = \big(W_k(I + S)\big)(I + S)^{-1}$ gives $\sigma_i(W_k) \le \sigma_i(W_k(I + S))\,\|(I + S)^{-1}\|_2 \le \sigma_i(W_k(I + S))$, since $\|(I+S)^{-1}\|_2 = 1/\sigma_{\min}(I+S) \le 1$. By Proposition 3, $\sigma_i(W_k) = \sigma_i(W^{\mathrm{pre}})$ for every i, which gives the stated sandwich. Under Assumption 2, $\sigma_d(W_k(I+S)) \ge \sigma_d(W^{\mathrm{pre}}) > 0$, so $W_k(I+S)$ has full column rank, and dividing the upper bound on σ_1 by the lower bound on σ_d gives $\kappa_2(W_k(I+S)) \le \sqrt{1 + \|S\|_2^2}\,\kappa_2(W^{\mathrm{pre}})$.

Neuron norms: the i-th row of $W_k(I + S)$ is $w_i^\top(I + S)$, so by Lemma 10(a), $\|w_i^\top(I + S)\|^2 = w_i^\top(I + S)(I + S)^\top w_i = w_i^\top(I - S^2)w_i = \|w_i\|^2 + \|Sw_i\|^2$, with equality to $\|w_i\|^2$ exactly when $w_i \in \ker S$.

Finally, the negative clause. The inequality $\|W_k(I + S)x\| \ge \|W^{\mathrm{pre}}x\|$ for all x is equivalent to $(I + S)^\top M(I + S) \succeq M$ with $M = R_k^\top (W^{\mathrm{pre}})^\top W^{\mathrm{pre}} R_k$, which fails in general: taking $M = \mathrm{diag}(1, 10^{-2})$ and $S = \left(\begin{smallmatrix} 0 & -0.3 \\ 0.3 & 0 \end{smallmatrix}\right)$, the matrix $(I + S)^\top M(I + S) - M$ has eigenvalues of both signs. The claim is therefore restricted to the ordered singular values and the neuron norms, as stated. □

Structured parameterizations. Both statements are in the spectral norm, which is what makes the structured cases tighter rather than merely equal. For block-diagonal $S = \mathrm{diag}(S_1, \ldots, S_{d/b})$ we have $\|S\|_2 = \max_i \|S_i\|_2$, so the in-flight gap is governed by the worst block and not by the aggregate. The corresponding Frobenius statement gains nothing, since $\|S\|_F^2 = \sum_i \|S_i\|_F^2$. Moreover $\mathrm{polar}(I + S) = \mathrm{diag}(\mathrm{polar}(I + S_i))$, so the restart is d/b independent $b \times b$ polar factors

at total cost $O(db^2)$. For the low-rank parameterization $S = UV^\top - VU^\top$ of rank at most $2r$, Lemma 10(b) gives $\sigma_i(I+S) = 1$ for $i > 2r$ and $\mathrm{polar}(I+S)$ acts as the identity on $\ker S$, so both the in-flight distortion and the epoch update are confined to a $2r$-dimensional subspace and the restart costs $O(dr^2)$.

B.8 Proof of Corollary 2

Let j index the inner steps of the current cycle and let $\mathcal{J}$ be the set of iterates at which $\|S^j\|_2$ is evaluated, which the hypothesis makes every optimizer step. The schedule restarts at the first $j \in \mathcal{J}$ with $\|S^j\|_2 \geq \tau$, and S is reset to 0 there, so $\|S^j\|_2 < \tau$ at every $j \in \mathcal{J}$ that does not trigger, and $\|S^j\|_2 = 0$ once the reset at a triggering j has been applied. Every forward pass is evaluated either at a monitored iterate that did not trigger, where $\|S\|_2 < \tau$, or at a freshly reset one, where $S = 0$: by hypothesis the restart fires before the next forward pass, so the crossing reading itself is never evaluated. Logged maxima can therefore exceed τ by one step's growth; no forward pass sees them. Substituting into Proposition 4(i) gives $\|A^\top A - I\|_2 = \|S^j\|_2^2 \leq \tau^2$. For the second inequality, Lemma 10(b) gives $\sigma_{\max}(I+S) = \sqrt{1+\|S\|_2^2}$ and $\sigma_{\min}(I+S) \geq 1$, and $\mathrm{polar}(I+S)$ replaces every singular value by 1, so $\|(I+S) - \mathrm{polar}(I+S)\|_2 = \sqrt{1+\|S\|_2^2} - 1$. Concavity of $u \mapsto \sqrt{1+u}$ at $u = 0$ gives $\sqrt{1+u} \leq 1 + u/2$ for $u \geq 0$, whence the bound is at most $\|S\|_2^2/2 \leq \tau^2/2$. No step uses a property of the loss, the data, the model or the optimizer. □

B.9 Restart Count under the Trigger and under a Period

Proposition 11 (Path length bounds the trigger's count; the step budget fixes the period's)**.** *Let S_t^- be the tangent parameter the t-th optimizer update produces and S_t its value after any restart at that step, so that $S_t = 0$ if a restart fired at t and $S_t = S_t^-$ otherwise, and let $L_T = \sum_{t<T} \|S_{t+1}^- - S_t\|_2$ be the spectral path length the optimizer moves the parameter over a run of T steps, the resets excluded. A monitor that restarts only on a reading $\|S_t^-\|_2 \geq \tau$ fires at most $\lfloor L_T/\tau \rfloor$ times, whatever the schedule on which it is read; read every m steps it also fires at most $\lfloor T/m \rfloor$ times, so its count is at most the smaller of the two. A fixed period M fires $\lfloor T/M \rfloor$ times, the merge at the terminal step counted, whatever L_T is.*

Proof. A cycle starts at $S = 0$ and, if it ends with a firing at step s, has $\|S_s^-\|_2 \geq \tau$; by the triangle inequality the path length the optimizer accumulates over that cycle is at least $\|S_s^-\|_2 \geq \tau$. Completed cycles are disjoint and the trailing incomplete one contributes a nonnegative length, so $N_R\,\tau \leq L_T$ for the number N_R of firings, whence $N_R \leq \lfloor L_T/\tau \rfloor$. Nothing in the argument uses when the norm is read, only that a firing follows a reading of norm at least τ; a monitor read every m steps has at most $\lfloor T/m \rfloor$ readings. The period's count is its definition. □ □

The bound is one-sided: it says how many restarts a path can force, not how many it will. Read that way, the 10 and 1,872 restarts of Figure 2d certify spectral path lengths of at least 1.0 and 187 at $\tau = 0.10$, while the 38-restart arm is read every twenty steps and fires on thirty-eight of thirty-nine readings, so its count is set by the reading granularity rather than by L_T. A period reports $\lfloor T/M \rfloor$ whatever the path. The lengthening intervals of Appendix H are consistent with a path that slows as the learning rate decays, which is what fewer restarts per window is forced by; the converse is not implied.

B.10 Proof of Proposition 8

Proof. By Lemma 10(d), $\mathrm{polar}(I+S)$ and $\exp(S)$ are both functions of the normal matrix S, hence simultaneously unitarily diagonalizable, and their difference is normal, so its spectral norm equals its spectral radius. On the eigenvector of S for $i\theta$ the two maps take values $e^{i\arctan\theta}$ and $e^{i\theta}$, whence

$$\big\|\mathrm{polar}(I+S) - \exp(S)\big\|_2 = \max_j \big|e^{i\arctan\theta_j} - e^{i\theta_j}\big| \leq \max_j \big|\theta_j - \arctan\theta_j\big|,$$

using $|e^{ia} - e^{ib}| = 2|\sin\frac{a-b}{2}| \leq |a-b|$. Let $f(\theta) = \theta^3/3 - \theta + \arctan\theta$ for $\theta \geq 0$. Then $f(0) = 0$ and $f'(\theta) = \theta^2 - 1 + \frac{1}{1+\theta^2} = \frac{\theta^4}{1+\theta^2} \geq 0$, so $f \geq 0$, i.e. $\theta - \arctan\theta \leq \theta^3/3$ for all $\theta \geq 0$, and the bound follows with $\max_j \theta_j = \|S\|_2$. It is not improvable, since $(\theta - \arctan\theta)/(\theta^3/3) \to 1$

as $\theta \to 0$, matching the Taylor expansions $\mathrm{polar}(I+S) = I + S + \frac{1}{2}S^2 + \frac{1}{2}S^3 + O(S^4)$ and $\exp(S) = I + S + \frac{1}{2}S^2 + \frac{1}{6}S^3 + O(S^4)$, whose leading discrepancy is $\frac{1}{3}S^3$.

For the surrogate, on the same eigenvector $(1+i\theta) - e^{i\theta} = \sqrt{1+\theta^2}\, e^{i \arctan\theta} - e^{i\theta}$, so $|(1+i\theta) - e^{i\theta}| \le (\sqrt{1+\theta^2} - 1) + |e^{i\arctan\theta} - e^{i\theta}| \le \frac{\theta^2}{2} + \frac{\theta^3}{3}$.

For the Cayley retraction, $(I - \Omega/2)^{-1}(I + \Omega/2)$ has eigenvalue $(1 + i\theta/2)/(1 - i\theta/2) = e^{2i\arctan(\theta/2)}$, so the same argument bounds the difference from $e^{i\theta}$ by $\max_j |\theta_j - 2\arctan(\theta_j/2)|$. With $g(\theta) = \theta^3/12 - \theta + 2\arctan(\theta/2)$ one has $g(0) = 0$ and $g'(\theta) = \frac{\theta^2}{4} - 1 + \frac{4}{4+\theta^2} = \frac{\theta^4}{4(4+\theta^2)} \ge 0$, giving the constant $\frac{1}{12}$. □

On the retraction identity quoted in the main text. For nonsingular X, $\arg\min_{R \in \mathcal{O}(d)} \|X - R\|_F = \mathrm{polar}(X)$ (Fan & Hoffman, 1955; Higham, 1986), and polar is left-equivariant, $\mathrm{polar}(R_k X) = R_k \,\mathrm{polar}(X)$ for orthogonal R_k. Hence $R_{k+1} = R_k \,\mathrm{polar}(I + S^M) = P_{\mathcal{O}(d)}\big(R_k(I + S^M)\big)$, the projection retraction at the tangent vector $R_k S^M$, which is second order (Absil & Malick, 2012; Boumal, 2023). This is a statement in adapter coordinates with the embedded Frobenius metric on $\mathcal{O}(d)$. The corresponding weight-space statement, $P_{W^{\mathrm{pre}}\mathcal{O}(d)}(W_k + W_k S) = W_k \,\mathrm{polar}(I+S)$, is *false* for general W^{pre}. Minimizing $\|W^{\mathrm{pre}}(A - R)\|_F$ over $R \in \mathcal{O}(d)$ is a weighted orthogonal Procrustes problem. Writing $M = (W^{\mathrm{pre}})^\top W^{\mathrm{pre}}$, the quadratic term is constant on the group, $\mathrm{tr}(R^\top M R) = \mathrm{tr}(MRR^\top) = \mathrm{tr}(M)$, so the problem reduces to maximizing $\langle MA, R\rangle$ and its minimizer is $\mathrm{polar}(MA)$. With $A = R_k(I + S)$ that equals $R_k \,\mathrm{polar}(I + S)$ when $M \propto I$ and not otherwise: polar is equivariant under an orthogonal left factor, not under the symmetric weight M. The adapter-space retraction is therefore the one that does not see W^{pre}: $\mathrm{polar}(I + S)$ is the same operation at every layer, which is what makes the defect bound of Corollary 2 hold without reference to the pretrained weight.

B.11 The odd–even rule

Lemma 12 (Defect of an approximated orthogonal map)**.** *Let $Q \in \mathrm{Skew}(d)$ with $\theta = \|Q\|_2$, let R be any exactly orthogonal function of Q, and let $\widetilde{R} = R\, e(Q)$, where e is real-analytic at the origin with $e(0) = 1$ and coefficients $a_0 = 1, a_1, a_2, \ldots$. Then*

$$\widetilde{R}^\top \widetilde{R} - I \;=\; e(-Q)\, e(Q) - I,$$

so the defect depends on the error factor e alone and not on R. It is an even function of the eigenvalue magnitudes of Q, and

$$\big\|\widetilde{R}^\top \widetilde{R} - I\big\|_2 \;=\; |c_{2r}|\, \theta^{2r} + O(\theta^{2r+2}), \qquad c_{2s} \;=\; \sum_{j+k=2s} (-1)^{(j-k)/2}\, a_j a_k,$$

where $2r$ is the least even index at which c_{2s} is nonzero.

Corollary 13 (Odd–even rule)**.** *Write $e(z) = 1 + a_n z^n + O(z^{n+1})$ with $a_n \neq 0$. If n is even, the defect is of order n with leading coefficient $2|a_n|$. If n is odd, that order cancels identically; for $n \ge 3$ the defect is of order $n+1$ with coefficient $2|a_{n+1}|$ when $a_{n+1} \neq 0$, and of order $2n$ with coefficient a_n^2 when $a_{n+1} = \cdots = a_{2n} = 0$. For $n = 1$ the order-two coefficient is $a_1^2 - 2a_2$, which can itself vanish, and for the exponential series it vanishes at every order, as it must. An approximation whose leading error sits at an odd power pays no orthogonality defect at that order.*

Proof of Lemma 12. Since $R^\top R = I$, $\widetilde{R}^\top \widetilde{R} = e(Q)^\top R^\top R\, e(Q) = e(Q)^\top e(Q)$, and $e(Q)^\top = e(Q^\top) = e(-Q)$ because e has real coefficients; $e(-Q)$ and $e(Q)$ are functions of the same matrix and commute, which gives the display and removes R. Set $g(z) = e(-z)e(z) - 1$. Then $g(-z) = g(z)$, so g carries only even powers. Over the eigenvalue pair $\pm i\theta_\ell$ of Q the defect's eigenvalue is $g(i\theta_\ell) = |e(i\theta_\ell)|^2 - 1$, real, and expanding $e(i\theta)e(-i\theta)$ gives the coefficient $(-1)^s \sum_{j+k=2s}(-1)^k a_j a_k$ at order $2s$, which equals c_{2s} because $(j+k)/2 + k$ and $(j-k)/2$ differ by the even number $2k$. The spectral norm is the largest eigenvalue magnitude, attained to leading order at $\theta_\ell = \theta$, giving $|c_{2r}|\theta^{2r} + O(\theta^{2r+2})$. □

Proof of Corollary 13. With $a_1 = \cdots = a_{n-1} = 0$, a pair (j,k) with $j + k = 2s < n$ and both indices in $\{0\} \cup \{n, n+1, \ldots\}$ requires $j = k = 0$, so $c_{2s} = 0$ below order n. If n is even,

the pairs $(0, n)$ and $(n, 0)$ give $c_n = 2(-1)^{n/2}a_n$. If n is odd, n is not an even index; for $n \geq 3$ only $(0, n+1)$ and $(n+1, 0)$ reach order $n+1$, giving $c_{n+1} = 2(-1)^{(n+1)/2}a_{n+1}$, and when $a_{n+1} = \cdots = a_{2n} = 0$ the pairs $(0, 2n)$ and $(2n, 0)$ drop out as well, so the first surviving pair is (n, n), giving $c_{2n} = a_n^2$. For $n = 1$ the pairs $(0, 2)$, $(2, 0)$ and $(1, 1)$ combine to $c_2 = a_1^2 - 2a_2$. □

B.12 Where the Cayley–Neumann family's moves stop

This subsection carries the comparison that Section 4.3 summarizes.

The family's moves aim at the size of the defect a step leaves, a cheaper Cayley evaluation and a deeper truncation, and neither reaches zero (Section 3). The free variable is the frequency at which exactness is paid for: a restart sets it and clears the defect exactly (Theorem 1), which is what makes the $k = 2$ truncation viable. A single polar chart holds only rotations whose angles stay inside $(-\pi/2, \pi/2)$; three restarts reach every rotation in the block group (Proposition 15), so what the chart excludes the schedule reaches.

The construction is not this method's alone to drop. The identity $\|A^\top A - I\|_2 = \|S\|_2^2$, the tolerance it supports and the restart are agnostic to the support's shape, so any adapter whose support is fixed by construction, rather than moved by the run, gets one structured multiply per step, a single merge into the frozen weight at the end, and a restart computed from the generator it already ships. Where the merged factor stays inside the support, as it does on the block-diagonal and the fixed-subspace ones, the pretrained weight is read and never written during training, so the adapter accumulates the restored factors and is stored and merged the way a low-rank one is; on butterfly the factor leaves the support and is folded into the weight instead (Appendix G). HRA, whose Householder chain is not of this form, is where the reach of the argument ends.

B.13 Proof of Lemma 14

Proof. We work with the map as implemented rather than with the product form $(I+Q)\sum_{i=0}^{m} Q^i$ or with the plain expanded truncation; all three approach the same limit but differ at finite order, and it is the default accumulation that matters here. That accumulation places coefficient two on every power up to $m-1$ and coefficient one on the last, and

$$C(z)\,(1-z^m) = (1+z)\big(1+z+\cdots+z^{m-1}\big) = 1+2z+2z^2+\cdots+2z^{m-1}+z^m, \qquad C(z) = \frac{1+z}{1-z},$$

so $\widetilde{R}_k = R\,(I - Q^m)$ with $R = C(Q)$, which is well defined because $I - Q$ is invertible for every skew Q by Lemma 10(c). The index map follows from the loop bounds. The loop over $\{3, \ldots, k-2\}$ is empty for $k \in \{3, 4\}$ while the final unit-coefficient multiplication is performed regardless, so both settings return the same matrix, with $m = 3$; for $k \geq 5$ the loop contributes powers through $k-2$ and the final term raises this to $m = k-1$. The case $k = 2$ returns before the tail term is added.

For the defect, $R^\top R = I$ and $(Q^m)^\top = (-1)^m Q^m$, so

$$\widetilde{R}_k^\top \widetilde{R}_k = \big(I - (-1)^m Q^m\big)\big(I - Q^m\big) = \begin{cases} I - Q^{2m}, & m \text{ odd}, \\ (I - Q^m)^2, & m \text{ even}. \end{cases}$$

Every factor is a polynomial in Q, hence normal and simultaneously diagonalizable with it, so the spectral norm is read off the eigenvalues $\pm i\theta_j$ of Q. For m odd, $(i\theta)^{2m} = -\theta^{2m}$ and the eigenvalue is $1 + \theta^{2m}$. For m even, $(i\theta)^m = (-1)^{m/2}\theta^m$ and the eigenvalue is $(1 \mp \theta^m)^2$, giving $2\theta^m \mp \theta^{2m}$. For m odd and $m \equiv 2 \pmod 4$ each deviation is increasing in θ, so the maximum over j is attained at $\theta = \|Q\|_2$; for $m \equiv 0 \pmod 4$ the deviation is $2t - t^2$ in $t = \theta_j^m$, increasing on $t \leq 1$, so the same holds for $\theta \leq 1$. For $k = 2$, $(I - 2Q)(I + 2Q) = I - 4Q^2$ has eigenvalue $1 + 4\theta^2$. □

Odd against even, and the k=2 case. The defect is not monotone in k. For odd m the deviation of $I - Q^m$ from orthogonality cancels to order $2m$; for even m it does not, and survives at order m. Since $m = k - 1$ away from the boundary cases, odd k land on the even-m branch and even k on the odd-m branch, so incrementing k alternately raises the order from m to $2m + 2$ and lowers it from $2m$ to $m + 1$. The default $k = 5$ sits on the worse branch: at order four it is less orthogonal than $k = 3$ at order six, which computes one term fewer. And $k = 3$ and $k = 4$ return the identical

Table 8: Per-task training time on GLUE in *mm:ss*, under the same setting as Table 2. The average is computed in seconds and converted back to *mm:ss*. Every method is timed with the block-diagonal skew matrix built as its own source writes it; the constant-basis replacement of Appendix D is applied to none of them. Lower is better; best is **bold** and second-best is underlined.

Method	#Param	Mem. (GB)	CoLA	STS-B	MRPC	RTE	SST-2	QNLI	Avg.
Full finetuning	184M	5.9	25:24	32:35	100:12	65:02	215:40	455:24	149:03
LoRA$_{r=8}$	1.33M	4.5	23:19	**29:25**	88:42	59:34	208:06	418:17	137:54
PiSSA$_{r=8}$	1.33M	4.5	24:03	32:22	**81:07**	**54:56**	**174:48**	**395:44**	**127:10**
DoRA$_{r=8}$	1.41M	5.8	27:25	33:56	87:04	61:02	201:43	419:38	138:28
LoRA-GA$_{r=8}$	1.33M	4.5	23:56	29:47	81:32	55:34	176:56	398:03	127:38
LoRA-XS$_{r=136}$	1.33M	4.2	23:33	32:40	86:26	58:55	185:32	412:56	133:20
GiVA$_{r=8}$	0.09M	4.7	23:46	32:31	87:20	60:18	189:47	476:28	145:02
PSOFT$_{r=46}$	0.08M	4.1	54:11	48:04	96:41	65:19	259:52	495:07	169:52
HRA$_{r=16}$	1.33M	15.2	144:40	107:18	129:32	88:32	616:48	614:37	283:35
BOFT$^{b=8}_{m=2}$	1.41M	6.3	74:09	65:06	114:10	73:12	368:58	612:32	218:01
OFTv2$_{b=32}$	1.29M	4.6	35:42	36:43	95:38	63:52	218:40	455:45	151:03
RTOFT-B$_{b=32}$ w/o restart	1.29M	4.6	**23:08**	32:22	85:58	58:18	185:41	410:56	132:44
RTOFT-B$_{b=32}$ (one-epoch period)	1.29M	4.6	23:36	32:50	86:40	59:00	185:55	411:03	133:11
RTOFT-B$_{b=32}$ (ours)	1.29M	4.6	23:11	33:28	86:38	58:32	185:48	411:04	133:07
RTOFT-L$_{r=6}$ w/o restart	0.99M	4.5	23:10	32:28	85:16	58:04	183:58	403:27	131:04
RTOFT-L$_{r=6}$	0.99M	4.5	23:33	32:33	85:25	58:11	184:23	403:28	131:16

matrix at the same cost: the fourth term is never formed. Second, and more useful here, the $k = 2$ case evaluates to $4\|Q\|_2^2$, which is exactly the defect $\|S\|_2^2$ of Theorem 1 under $S = 2Q$. The tangent map of Section 4.1 is therefore exactly the $k{=}2$ truncated Cayley–Neumann map, and Section 4.3 draws the consequence. Approximation bounds for the truncated Cayley map have appeared in that line of work under the condition $\|Q\|_2 < 1$; Lemma 14 differs in needing no convergence condition (its odd and $m \equiv 2 \pmod 4$ branches hold for every $\|Q\|_2$, and the $m \equiv 0 \pmod 4$ branch for $\|Q\|_2 \leq 1$, an order of magnitude above the measured range) and in describing the orthogonality defect rather than the approximation error.

C Per-task Runtime on GLUE

Table 8 resolves the average wall-clock column of Table 2 into its six per-task components. All runs share hardware, epoch budget and task-level hyperparameters. The ordering holds on every task: all three block RTOFT rows are faster than every orthogonal baseline on all six tasks, and the margin grows with the per-step orthogonalization that the baseline's parameterization performs. On CoLA, OFTv2 takes 35:42 against the one-epoch period's 23:36, a 34% reduction, while BOFT takes 74:09 and HRA 144:40 on the same task. The restart is nearly free. Enabling the one-epoch period moves the block variant from 23:08 to 23:36 on CoLA, twenty-eight seconds on a run of roughly twenty-three minutes. On the six-task average it moves from 132:44 to 133:11, an increase of 0.3%; the trigger, at 23:11 and 133:07, sits between the two. This is the amortization made concrete: the polar factorization is performed once per restart, on d/b blocks of size b, rather than on every step. The low-rank variant is the fastest entry in the table apart from PiSSA (127:10) and LoRA-GA (127:38), at 131:04 without restart and 131:16 with it. PSOFT makes the same point from the opposite direction: it carries the smallest trainable-parameter count in the table and is nonetheless among the slowest methods in it, at 169:52. Wall-clock in orthogonal finetuning is set by the per-step orthogonalization cost and by what the backward pass must unroll, not by the number of trainable parameters.

D Where the Wall-Clock Goes

Total wall-clock mixes training with evaluation and checkpointing, so we also time the optimizer step in isolation, with evaluation and saving disabled.

Table 9: Comparison of orthogonal finetuning parameterizations for a pretrained $W \in \mathbb{R}^{d_{\text{out}} \times d}$. b is the block size, m the number of butterfly components, r the number of reflections or the skew rank, and K_N the Neumann truncation order. "Per-step orthogonalization" is the work needed to turn the trainable parameters into an orthogonal matrix on *every* forward pass; "unrolled backward" indicates that this work is an iteration whose computation graph the backward pass must differentiate through. RTOFT is the only entry whose per-step cost is a single multiply, and the only one that restores exactly, on a schedule the measured drift itself can set.

	OFT	BOFT	HRA	OFTv2	**RTOFT (ours)**
Update	$W\,\mathrm{diag}(R_i)$	$W\prod_{i=1}^{m}\widetilde{B}_i$	$W\prod_{i=1}^{r}(I-2u_iu_i^\top)$	$W\,\mathrm{diag}(R_i)$	$W(I+S)$
Orthogonal map	Cayley	Cayley $\times\, m$	reflection chain	Cayley–Neumann	tangent, $R=I+S$
Per-step orthogonalization	$(I-S)^{-1}$	m inverses	r rank-1 updates	K_N block matmuls	**none**
Unrolled backward	✓	✓	✓	✓	–
# Parameters	$\frac{d(b-1)}{2}$	$\frac{dm(b-1)}{2}$	rd	$\frac{d(b-1)}{2}$	$\frac{d(b-1)}{2}$ or $2dr$
Per-step cost	$O(d(b^2{+}b{+}d_{\text{out}}))$	$O(d((b^2{+}b)m{+}d_{\text{out}}))$	$O(d(r{+}d_{\text{out}}))$	$O(d(K_Nb^2{+}d_{\text{out}}))$	$O(d(b{+}d_{\text{out}}))$
Amortized restart cost	N/A	N/A	N/A	N/A	$O(db^2)$ or $O(dr^2)$
Exact orthogonality	every step	every step	every step	*never* (truncated)	**at every restart**
Full-capacity setting	$b=d$	$m{=}\log d$, $b{=}2$	$r=d-1$	$b=d,\ K_N\to\infty$	$b=d$, restart

Table 10: All four RTOFT variants on GLUE under the test-split protocol of Table 2. "B" is the block-diagonal skew parameterization with block size b and "L" the low-rank skew parameterization with rank r. The restart column indicates whether the tangent parameters are periodically projected back onto $\mathcal{O}(d)$. OFTv2 is repeated from Table 2 as the reference point. Subscripts are standard deviations over the five seeds. The restart does not benefit the low-rank variant, and Sections 7 and 8 carry that loss as the limit of the construction.

Variant	Restart	#Param	Mem. (GB)	Time ↓	CoLA	STS-B	RTE	MRPC	SST-2	QNLI	Avg.
OFTv2$_{b=32}$ (k=5)	✗	1.29M	4.6	151:03	$71.64_{1.53}$	$91.60_{0.46}$	$88.48_{2.01}$	$92.15_{3.90}$	$94.90_{0.75}$	$94.06_{1.00}$	88.81
OFTv2$_{b=32}$ (k=5)	✓	1.29M	4.6	N/A	72.04	92.12	85.45	92.31	95.28	94.36	88.59
RTOFT-B$_{b=32}$	✗	1.29M	4.6	132:44	$71.72_{1.48}$	$91.54_{0.64}$	$84.83_{2.05}$	$90.85_{1.07}$	$93.62_{1.19}$	$94.03_{0.41}$	87.77
RTOFT-B$_{b=32}$ (one-epoch period)	✓	1.29M	4.6	133:11	$\mathbf{72.52}_{1.08}$	$\mathbf{92.30}_{0.36}$	$\mathbf{88.55}_{1.62}$	$\mathbf{92.78}_{0.84}$	$\mathbf{95.88}_{0.39}$	$\mathbf{94.65}_{0.37}$	**89.45**
RTOFT-L$_{r=6}$	✗	0.99M	4.5	131:04	$70.44_{1.18}$	$91.68_{0.61}$	$87.05_{2.44}$	$90.20_{1.25}$	$95.27_{0.66}$	$93.99_{0.45}$	88.11
RTOFT-L$_{r=6}$	✓	0.99M	4.5	131:16	$71.59_{0.80}$	$91.71_{0.58}$	$83.74_{2.37}$	$89.71_{1.99}$	$95.18_{0.33}$	$94.59_{0.54}$	87.75

Measurement resolution. Rerunning STS-B unchanged returns a bitwise identical score and a training time differing by 0.94%, so the difference is scheduling and hardware alone. We therefore treat wall-clock differences below one percent as unresolved and draw no conclusion from them.

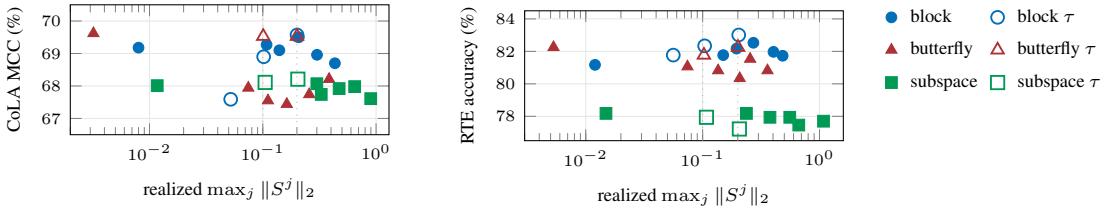
The grid. Three implementation switches are available: applying the transform to the weight rather than to the activations, materializing $I-\alpha A$ rather than subtracting a scaled product, and building the skew matrix by a constant-basis multiplication rather than a scatter. Over the resulting 2^3 grid, on three sequence lengths, the constant-basis switch is the only one that moves the wall-clock. It is bitwise identical (three seeds agree to the last digit of Matthews correlation) and removes 935 kernel launches per step, a count that does not move with sequence length (935.3, 936.5 and 935.5 at $2{,}048$, $4{,}096$ and $8{,}192$ tokens), while making device time slightly *worse*. Wall-clock nonetheless falls by 11.5%, 8.4% and 3.1% at those three lengths. Applying the transform to the weight instead cuts device time by 25 to 34% and moves wall-clock by -0.4%, $+0.2\%$ and -0.4%. Wall-clock here follows the number of kernels launched and is insensitive to the arithmetic they perform.

The launch count is also the one quantity that does not vary with sequence length, so a comparison that leaves the two constructions unequal reports a near-constant difference, and what is constant in it belongs to the implementations rather than to the methods. Removed from both, what remains is the truncation itself.

Whole runs against isolated steps. Table 8 times whole runs, and its ratios are the larger ones: 1.10 to 1.18 on the three longest tasks against 1.02 per training step. Evaluation is the reason, and it widens the gap rather than diluting it. The reference construction is rebuilt on every forward pass, so it is paid once per evaluation batch as well, against a pass that has no backward computation to spread it over; on QNLI the reference spends 25.5 seconds in evaluation against our 19.7.

Table 11: Standard deviation across the five seeds for every method of Table 14, in the same column order, with both RTOFT-B schedules of Table 2. Bold marks the smallest deviation among the orthogonal methods and underline the smallest overall where that is a different entry. The last column averages the six tasks. The trigger is the least variable entry in the table on five of the seven columns, and less variable than the period it replaces on every column.

Method	CoLA	STS-B	RTE	MRPC	SST-2	QNLI	Mean
Full finetuning	3.56	0.40	2.10	1.28	0.71	0.35	1.40
LoRA$_{r=8}$	3.60	0.57	1.85	0.51	1.62	0.62	1.46
PiSSA$_{r=8}$	2.48	1.16	1.45	2.42	1.47	0.86	1.64
DoRA$_{r=8}$	1.15	0.94	3.26	1.75	2.31	0.37	1.63
LoRA-XS$_{r=136}$	2.86	0.65	2.58	1.97	0.34	0.28	1.45
GiVA$_{r=8}$	<u>0.30</u>	0.71	1.54	1.44	0.59	0.74	0.89
PSOFT$_{r=46}$	1.98	0.39	3.98	1.40	0.70	1.24	1.62
HRA$_{r=16}$	1.69	1.01	**0.82**	1.32	0.81	0.75	1.07
BOFT$_{b=8,m=2}$	2.09	0.72	2.82	1.17	1.56	0.41	1.46
OFTv2$_{b=32}$	1.53	0.46	2.01	3.90	0.75	1.00	1.61
RTOFT-B$_{b=32}$ (one-epoch period)	1.08	0.36	1.62	0.84	0.39	0.37	0.78
RTOFT-B$_{b=32}$ (ours)	**0.84**	**0.14**	0.92	**0.31**	**0.16**	**0.14**	**0.42**



(a) CoLA, 5 seeds per point

(b) RTE, 3 seeds per point

Figure 4: One maintenance picture across three parameterizations. Fixed periods (closed markers; the rightmost point of each family is no restart) and drift tolerances (open markers) at a fixed ten-epoch budget, plotted against the drift they realize. Each family moves by at most 2.2 points across its full range, and the two tolerances pin the same two drift values on every support. The horizontal axis is the same operator-level drift for all three supports, $\max_j \|S^j\|_2$, equivalently the square root of the spectral orthogonality defect $\|A^\top A - I\|_2$.

E Two Further Parameterizations

The drift axis of Figure 3 is the same operator quantity in all three families. For the block parameterization $A = I + S$ with S skew, $\|A^\top A - I\|_2 = \|S\|_2^2$ (Lemma 10). For the subspace parameterization $A = I + P^\top S_r P$ with P row-orthonormal and S_r skew, the linear terms cancel and $A^\top A - I = -P^\top S_r^2 P$, so $\|A^\top A - I\|_2 = \|S_r\|_2^2$ again. For the butterfly parameterization the factors collapse to first order into a single skew generator, and the monitor reads that merged generator, the one the forward pass uses. In every case the plotted drift is $\max_j \|S^j\|_2$ over adapted modules: the horizontal axis of Figure 3 is the square root of the whole transform's spectral defect.

BOFT composes m orthogonal factors with permutations between them, and what it gains over a single rotation comes from that depth. Linearization destroys the depth: to first order $(I + S_1)(I + S_2) = I + S_1 + S_2$, and a sum of skew matrices is again a single skew matrix, so a tangent butterfly of depth m is a tangent butterfly of depth one. The restart buys it back, because folding $\mathrm{polar}(I+S)$ into W begins the next cycle in a new frame and the frames compose. The subspace construction writes $R = I + P_r^\top (R_r - I) P_r$ with P_r a frozen projection; replacing R_r by $I_r + S_r$ gives the forward pass $W(I + P_r^\top S_r P_r)$, and since $P_r P_r^\top = I_r$ the restart is an $r \times r$ polar rather than anything of the ambient dimension.

Table 12: The tangent–restart construction applied to a subspace-restricted parameterization, at the PSOFT configuration of Table 2 and five seeds throughout. The tangent version matches the method it replaces and trains $1.19\times$ faster; removing the restart costs 1.16 points and drops it below the original.

Method	Time ↓	CoLA	STS-B	RTE	MRPC	SST-2	QNLI	Avg.
PSOFT$_{r=46}$	169:52	71.60	91.42	85.75	90.53	**95.18**	**93.60**	88.01
tangent, no restart	**141:44**	70.54	89.88	84.27	90.50	94.61	92.36	87.03
tangent + restart	142:20	**72.05**	**91.43**	**86.02**	**91.22**	94.86	93.56	**88.19**

Table 13: The tangent–restart construction applied to butterfly factorization, five seeds throughout, with seed standard deviations subscripted. Linearizing the product drops it below the method it replaces; the restart recovers the depth and passes it, at $1.41\times$ the training speed.

Method	Time ↓	CoLA	STS-B	RTE	MRPC	SST-2	QNLI	Avg.
BOFT$_{b=8,m=2}$	218:01	68.85	91.85	86.78	89.91	**95.33**	**93.85**	87.76
tangent, no restart	**153:01**	$68.35_{2.58}$	$90.91_{1.25}$	$84.41_{3.14}$	$90.72_{1.76}$	$94.88_{1.30}$	$93.16_{2.28}$	87.07
tangent + restart	154:10	$\mathbf{70.23}_{1.49}$	$\mathbf{92.25}_{0.43}$	$\mathbf{87.13}_{2.63}$	$\mathbf{91.27}_{1.42}$	$95.21_{0.96}$	$93.74_{1.29}$	**88.31**

F Cost Accounting and Reachability

Efficient parameterizations. A dense skew-symmetric matrix contains $d(d-1)/2$ free parameters. We consider two structured forms that keep the tangent update parameter-efficient.

Block parameterization. Let b divide d and $S = \mathrm{blockdiag}(S_1, \ldots, S_{d/b})$ with $S_i \in \mathrm{Skew}(b)$: $d(b-1)/2$ trainable parameters, the same intrinsic count as block OFT and OFTv2 at block size b. The tangent forward pass costs $O(d(b + d_{\mathrm{out}}))$ ($O(db)$ for the block product Sx, $O(dd_{\mathrm{out}})$ for W), with no per-step $O(db^2)$ orthogonal construction. A restart is block-separable, $\mathrm{polar}(I+S) = \mathrm{blockdiag}(\mathrm{polar}(I+S_1), \ldots, \mathrm{polar}(I+S_{d/b}))$: the d/b independent $b \times b$ SVDs cost $O(db^2)$ in total, merging the rotation into $W \in \mathbb{R}^{d_{\mathrm{out}} \times d}$ costs $O(dd_{\mathrm{out}}b)$, and the amortized restart surcharge is $O\big(d(b^2 + d_{\mathrm{out}}b)/M\big)$. With one restart per epoch this surcharge is 0.3% of total GLUE training time, and it falls below the one-percent measurement resolution on the three longest tasks, where there is most to amortize it over (Table 8).

Optimizer state. A restart rewrites the parameters, not the optimizer: Adam's moment estimates are carried across untouched, so the tangent parameter resumes from zero with its accumulated momentum. Folding $R = \mathrm{polar}(I+S)$ into W replaces $W^\top \nabla\mathcal{L}$ by $R^\top W^\top \nabla\mathcal{L}$ ahead of the skew projection, so at a fixed loss gradient the tangent gradient the carried moments track moves by at most $\arctan \|S\|_2 \, \|W^\top \nabla\mathcal{L}\|_2$, and the misalignment is governed by the same drift the schedule controls (Corollary 7), and no optimizer reset is needed.

Extension to subspace orthogonal parameterizations. The construction extends beyond the full or block-diagonal generator. An orthogonal PEFT method acting in a fixed r-dimensional support, $R = I + P_r^\top (R_r - I_r) P_r$ with $R_r \in \mathcal{O}(r)$, has the tangent form $R \approx I + P_r^\top S_r P_r$ with $S_r \in \mathrm{Skew}(r)$, and its restart needs only $\mathrm{polar}(I_r + S_r)$ in the r-dimensional support: the amortization extends to subspace-restricted orthogonal adapters unchanged. This is the fixed-support condition of Section 7; a low-rank generator and a warm-restart variant are given in Appendix G.

For a block size b, OFTv2 (Qiu et al., 2025b) approximates the Cayley inverse by a Neumann polynomial of order K_N, so every forward pass performs K_N block-matrix multiplications beyond the frozen linear map, contributing $O(dK_N b^2)$, and the backward pass differentiates through the same polynomial. RTOFT replaces that per-step construction by the tangent product $I + S$, which costs $O(db)$ on top of the common $O(dd_{\mathrm{out}})$. A polar restart costs $O(db^2)$ for the block polar decompositions, together with the $O(dd_{\mathrm{out}}b)$ merge of Section 4, so spread over a period of M

steps orthogonalization contributes

$$O\left(\frac{d(b^2 + d_{\text{out}}b)}{M}\right)$$

per inner step. This is the precise sense in which RTOFT amortizes orthogonality. Wall-clock is measured separately in Section 7 (Appendix D); the accounting identifies which method has a fixed per-step cost to amortize.

What an approximation to an orthogonal map costs. Truncation is the standard way to make an orthogonal parameterization cheap, and the question of what a given truncation costs in orthogonality has a general answer that does not depend on which orthogonal map is being approximated.

Lemma 14 (Defect of the Cayley–Neumann map)**.** *Let $Q \in \mathrm{Skew}(d)$ have eigenvalues $\pm i\theta_j$, write $\theta = \|Q\|_2 = \max_j \theta_j$, and let $R = (I + Q)(I - Q)^{-1}$. The Cayley–Neumann map retaining k terms is, as implemented,*

$$\widetilde{R}_k = R\,(I - Q^m), \qquad m = \begin{cases} 3, & k \in \{3, 4\}, \\ k - 1, & k \geq 5, \end{cases}$$

with $\widetilde{R}_2 = I + 2Q$ a separate case. Its orthogonality defect is

$$\left\|\widetilde{R}_k^\top \widetilde{R}_k - I\right\|_2 = \begin{cases} \theta^{2m}, & m \text{ odd}, \\ 2\theta^m - \theta^{2m}, & m \equiv 0 \pmod 4 \text{ and } \theta \leq 1, \\ 2\theta^m + \theta^{2m}, & m \equiv 2 \pmod 4, \end{cases}$$

and $\|\widetilde{R}_2^\top \widetilde{R}_2 - I\|_2 = 4\theta^2$. For $m \equiv 0 \pmod 4$ and $\theta > 1$ the norm is $\max_j |2\theta_j^m - \theta_j^{2m}|$ over the eigenvalue magnitudes; every setting we measure has $\theta \leq 0.39$. The defect is thus of order $2m$ for odd m and of order m for even m. In particular the default setting $k = 5$ has $m = 4$ and a defect of order four, while the cheaper $k = 3$ has $m = 3$ and a defect of order six.

Both statements follow from Corollary 13: the error factor is $e_m(z) = 1 - z^m$, whose leading coefficient sits at z^m with $a_m = -1$ and all others zero, so the defect is of order m when m is even and of order $2m$ when it is odd. The separate case $k = 2$ is not of that form: there $e(z) = (1 + 2z)(1 - z)/(1 + z) = 1 - 2z^2 + O(z^3)$, giving $n = 2$ and a defect of $2|a_2|\theta^2 = 4\theta^2$, and the general rule covers it without a separate argument.

Near the origin OFTv2's default $k = 5$ pointwise orthogonality defect is fourth order in $\|Q\|_2$, whereas RTOFT's in-flight defect is second order in $\|S\|_2$, these being the $k = 5$ and $k = 2$ members of the same family (Section 4.3). At measured operating points the margin is far narrower than the orders suggest: on the three tasks where we logged both, OFTv2's in-flight gap is smaller by a factor of seventy-eight on CoLA, where its $\|Q\|_2$ only reaches 0.16, and by under three on RTE and STS-B, where it reaches 0.34 and 0.39. The ordering holds throughout; the size of it is set by how far the optimizer happens to travel. Its trade-off is different: the inner computation is cheaper, and the defect is reset to exactly zero at every restart and at the returned model, which no finite truncation attains at any iterate.

By Proposition 15, a single polar chart realizes only rotations whose angles all lie in $(-\pi/2, \pi/2)$, a proper subset of the Cayley image.

Proposition 15 (Restarts cover the chart)**.** *Let $\mathcal{G} = \{\mathrm{polar}(I + S) : S \in \mathcal{S}\}$ be the set of transforms one restart can realize, with $\mathcal{S}$ the block-diagonal skew family defined above, and let $\mathcal{B}$ be the corresponding block-diagonal rotation group. Then $\mathcal{G}$ is exactly the set of elements of $\mathcal{B}$ whose rotation angles all lie in $(-\pi/2, \pi/2)$, and $\mathcal{G}^3 = \mathcal{B}$.*

Proof. For skew S with eigenvalues $\pm i\theta_j$, $\mathrm{polar}(I + S)$ has eigenvalues $e^{\pm i \arctan \theta_j}$, and $\arctan$ maps $\mathbb{R}$ onto $(-\pi/2, \pi/2)$. Conversely, if $R \in \mathcal{B}$ has every angle φ_j in that interval, let S act on the eigenplanes of R with $\theta_j = \tan \varphi_j$; each eigenplane lies inside a block, so S is block-diagonal skew, and $\mathrm{polar}(I + S) = R$ by Lemma 10(d). This gives the first claim. For the second, any $R \in \mathcal{B}$ lies in a maximal torus, where it is fixed by angles $\varphi_j \in (-\pi, \pi]$; the element with angles $\varphi_j/3$ lies in $\mathcal{G}$ since $|\varphi_j/3| < \pi/2$, and its cube is R because the three factors commute. □

Table 14: GLUE test-split results on DeBERTaV3-base under the test-split protocol of Wu et al. (2026); every method, LoRA-XS (Bałazy et al., 2025) included, is re-run in the same pipeline. Cells are means over the same five seeds; Time is mean wall-clock per task on one V100, Rel. Time normalized to LoRA. Per-seed standard deviations and the full protocol are in Appendix J. Best in **bold**, second-best underlined; tied seconds are left unmarked.

	Efficiency					GLUE performance						
Method	#Param	Adapter	Mem. (GB)	Time ↓	Rel. Time ↓	CoLA	STS-B	RTE	MRPC	SST-2	QNLI	Avg.
Full finetuning	184M	184M	5.9	149:03	1.08×	68.20	91.34	83.35	90.48	95.18	92.94	86.92
Low-rank adaptation												
LoRA$_{r=8}$	1.33M	1.33M	4.5	137:54	1.00×	71.98	91.42	84.53	90.63	95.01	93.92	87.91
PiSSA$_{r=8}$	1.33M	1.33M	4.5	**127:10**	**0.92**×	67.52	91.43	86.17	90.25	94.69	92.90	87.16
DoRA$_{r=8}$	1.41M	1.41M	5.8	138:28	1.00×	67.99	91.92	87.68	89.24	95.25	93.99	87.68
LoRA-XS$_{r=136}$	1.33M	23.89M	4.2	133:20	0.97×	64.67	91.22	85.00	87.59	93.69	94.10	86.05
GiVA$_{r=8}$	0.09M	1.41M	4.7	145:02	1.05×	72.18	91.51	85.36	90.58	95.17	93.81	88.10
Orthogonal finetuning												
PSOFT$_{r=46}$	**0.08M**	7.71M	**4.1**	169:52	1.23×	71.60	91.42	85.75	90.53	95.18	93.60	88.01
HRA$_{r=16}$	1.33M	1.33M	15.2	283:35	2.06×	70.05	90.47	86.46	**92.82**	94.71	93.15	87.94
BOFT$_{b=8,m=2}$	1.41M	1.41M	6.3	218:01	1.58×	68.85	91.85	86.78	89.91	95.33	93.85	87.76
OFTv2$_{b=32}$ (k=5)	1.29M	**1.29M**	4.6	151:03	1.10×	71.64	91.60	88.48	92.15	94.90	94.06	88.81
RTOFT-B$_{b=32}$ (one-epoch period)	1.29M	**1.29M**	4.6	133:11	0.97×	72.52	92.30	88.55	92.78	95.88	94.65	89.45
RTOFT-B$_{b=32}$ (ours)	1.29M	**1.29M**	4.6	133:07	0.97×	**73.55**	**92.44**	**89.20**	92.65	**96.75**	**94.80**	**89.90**

Method	#P. (M)	S2I ↑			L2F ↓
		mIoU	mAcc	aAcc	Error
DreamBooth	859.52	7.72	14.40	33.61	146.19
ControlNet	361.30	20.88	30.91	61.42	7.61
T2I-Adapter	77.00	16.38	26.31	51.63	23.75
LoRA$_{r=16}$	2.50	24.58	37.74	60.31	8.01
OFT$_{b=4}$	26.40	27.18	40.23	62.58	7.03
COFT$_{b=4}$	26.40	26.81	39.95	62.82	6.97
OFTv2$_{c=2}$	26.34	27.79	40.78	64.36	6.61
BOFT$^{m=4}_{bn=16}$	26.40	29.01	41.43	67.94	5.60
HRA$_{r=8,\lambda=10^{-3}}$	0.89	29.19	40.80	67.73	5.51
RTOFT-B$_{b=2}$ w/o restart	26.35	26.32	40.91	64.90	6.86
RTOFT-B$_{b=2}$ (one-epoch period)	26.35	29.45	42.26	67.97	5.36
RTOFT-B$_{b=2}$ (ours)	26.35	**30.18**	**42.78**	**69.19**	**5.31**

Table 15: Controllable generation with Stable Diffusion. S2I: ADE20K; L2F: CelebA-HQ. The RTOFT-B (ours) row runs the drift trigger of Corollary 2 at $\tau = 0.10$, evaluated every step; the one-epoch period row differs from it in the schedule alone.

Three restarts therefore realize every transform in the group, and a run under the one-epoch period performs one per epoch. The chart is a per-interval constraint that the schedule removes, the same trade the paper makes with exactness. Relaxed rather than strict orthogonality has been reported preferable in other settings (Vorontsov et al., 2017).

G ALL RTOFT VARIANTS

Low-rank parameterization. Alternatively, let

$$S = UV^\top - VU^\top, \qquad U, V \in \mathbb{R}^{d\times r}. \tag{2}$$

This is skew-symmetric by construction and contains $2dr$ trainable parameters. We initialize U with Kaiming initialization and set $V = 0$, so that the initial adapter is exactly the identity while both factors remain trainable.

The dense matrix S is never materialized. Its action is

$$Sx = U(V^\top x) - V(U^\top x),$$

which costs $O(dr)$, giving total forward complexity

$$O(d(r + d_{\text{out}})).$$

A naive polar decomposition of $I + S$ would cost $O(d^3)$. The low-rank structure avoids this. Let

$$P = [\, U \mid V \,] \in \mathbb{R}^{d \times 2r}, \qquad P = Q_P R_P$$

be a thin QR factorization. Since both the row and column spaces of S lie in $\mathrm{col}(P)$,

$$S = Q_P M_S Q_P^\top, \qquad M_S = Q_P^\top S Q_P \in \mathrm{Skew}(2r).$$

Any orthogonal completion $[\, Q_P \; Q_\perp \,]$ therefore block-diagonalizes $I + S$ into

$$\mathrm{blockdiag}(I_{2r} + M_S, I_{d-2r}).$$

Hence

$$\mathrm{polar}(I + S) = I + Q_P\big(Q_s - I_{2r}\big)Q_P^\top, \qquad Q_s \coloneqq \mathrm{polar}(I_{2r} + M_S). \tag{3}$$

The corresponding merge can be performed without materializing a dense $d \times d$ orthogonal matrix, as $W\,\mathrm{polar}(I+S) = W + (WQ_P)(Q_s - I_{2r})Q_P^\top$. A restart therefore costs $O(dr^2)$ for the QR and projections, $O(r^3)$ for the small polar decomposition, and $O(dd_{\mathrm{out}}r)$ for the merge, an amortized surcharge of $O(d(r^2 + d_{\mathrm{out}}r)/M)$ with no cubic dependence on d.

Warm restart. The block parameterization has no trainable state beyond S, so resetting S after a merge loses nothing. The low-rank parameterization stores its state in two factors. We therefore carry the current U subspace into the new frame while resetting only V:

$$U \leftarrow Q^\top U, \qquad V \leftarrow 0, \qquad Q = \mathrm{polar}(I + S). \tag{4}$$

The new tangent adapter is again the identity, but the selected subspace is retained, and at $V = 0$ its first gradient is $\nabla_V \mathcal{L}|_{V=0} = -2\,\mathrm{skew}(W^\top \nabla \mathcal{L}(W))\,U$, so the retained U immediately determines the directions explored in the next cycle. Proposition 16 states what each kind of reset leaves to the gradient.

Proposition 16 (What a reset leaves to the gradient)**.** *Let $g = \mathrm{skew}\big(W^\top \nabla \mathcal{L}(W)\big)$ be the Riemannian gradient of the loss at the current weight, and consider the tangent adapter $W(I + S)$ at the instant after a restart.*

- *(i)* Fixed support. *If $S = S(\theta)$ is linear in its parameters with injective Jacobian J, as on the block-diagonal, subspace and butterfly supports, then at the reset $\theta = 0$ one has $\nabla_\theta \mathcal{L} = J^\top g$, which vanishes exactly when g has no component in the support. The reset discards no first-order information.*
- *(ii)* Moving support. *If $S = UV^\top - VU^\top$, then $\nabla_U \mathcal{L} = 2\,gV$ and $\nabla_V \mathcal{L} = -2\,gU$. Under the warm reset $V \leftarrow 0$ the loss is stationary in U and the first step after the restart can move V only, along the directions gU that the retained U selects, provided $gU \neq 0$; if the retained U lies in $\ker g$ the warm reset degenerates to the cold one, where $U = V = 0$ and both gradients vanish and the iterate is a critical point of the parameterization.*

Proof. For skew E, $\frac{d}{dt}\mathcal{L}\big(W(I + tE)\big)\big|_{t=0} = \langle \nabla \mathcal{L}(W), WE \rangle = \langle W^\top \nabla \mathcal{L}(W), E \rangle = \langle g, E \rangle$, the symmetric part of $W^\top \nabla \mathcal{L}(W)$ pairing to zero against E. In (i) the differential of S is $J\,d\theta$, so $\nabla_\theta \mathcal{L} = J^\top g$, and $J^\top g = 0$ if and only if $g \perp \mathrm{range}(J)$. In (ii) the differential is $dS = dU\,V^\top + U\,dV^\top - dV\,U^\top - V\,dU^\top$; pairing with g and using $g^\top = -g$ gives $\langle g, dU\,V^\top - V\,dU^\top \rangle = 2\langle gV, dU \rangle$ and $\langle g, U\,dV^\top - dV\,U^\top \rangle = -2\langle gU, dV \rangle$, which are the two gradients. Setting $V = 0$, or $U = V = 0$, in them gives the two reset cases. □ □

The warm restart carries the column space of U into the new frame and resets V. Restart improves the block variant and does not improve the low-rank one.

Adapter form. The restart of Section 4 writes the restored factor into the weight. Equivalently, the pretrained weight stays untouched and the factor accumulates in the adapter, $R \leftarrow R\,\mathrm{polar}(I + S)$ and $S \leftarrow 0$, the deployed weight being $W^{\mathrm{pre}}R$ after the last restart. On the block support every factor is block-diagonal with the same blocks, so R is, and the adapter holds db numbers; on the subspace support every factor has the form $I + P_r^\top(Q_r - I_r)P_r$ with the same frozen P_r and $Q_r \in \mathcal{O}(r)$, so R does too. In either case the pretrained weight is read and never written during training, which is what lets the adapter be stored apart from it and merged once, as a low-rank adapter is. The butterfly support is the exception: the polar factor of its merged generator need not lie in the support, so there the factor is folded into the weight.

G.1 Connection with Low-rank Adaptation

Expanding (2) in the tangent update gives

$$W(I + UV^\top - VU^\top) = W + (WU)V^\top - (WV)U^\top. \tag{5}$$

Thus RTOFT-L has the algebraic form of an additive low-rank update, but its left factors are generated from the pretrained weight rather than trained independently. In particular,

$$\mathrm{col}(W(I + S)) = \mathrm{col}(W),$$

because $I+S$ is invertible for every skew-symmetric S. The tangent update therefore remains inside the column space already spanned by the pretrained layer.

LoRA with rank r stores $(d + d_{\mathrm{out}})r$ trainable parameters, whereas RTOFT-L stores $2dr$. The difference reflects the fact that the d_{out}-side factors above are supplied by W rather than learned freely. The resulting update has rank at most $2r$.

Table 10 reports all four combinations of skew parameterization and restart under the test-split protocol of Table 2, with OFTv2 repeated as the reference point. The block variant is the clearest case: the restart, the one-epoch period throughout this table, lifts the six-task average from 87.77 to 89.45, and it does so on every one of the six tasks, with the largest gains on RTE and SST-2. Without restart, the same parameterization reaches only 87.77, below OFTv2's 88.81. The restart also narrows the seed spread: the block variant's mean standard deviation across the six tasks falls from 1.14 to 0.78, and it is lower on every one of the six tasks. This is the same effect visible against OFTv2 in Table 3, here with the restart as the only change, and it is what periodic re-projection should do: it removes drift that would otherwise accumulate differently in each run. The low-rank variant behaves in the opposite way: restart moves its average from 88.11 down to 87.75, and the loss falls entirely on RTE (87.05 to 83.74); across the other five tasks the restart is a net gain of 1.20 points. The two parameterizations differ in what a reset leaves behind (Proposition 16). For the block variant, $S \leftarrow 0$ leaves a nonzero gradient, $\nabla_S \mathcal{L} = \mathrm{skew}(W^\top \nabla \mathcal{L})$, so the parameter moves on the next step. For the low-rank one, $\nabla_U \mathcal{L} = 2\,\mathrm{skew}(W^\top \nabla \mathcal{L})\, V$ is proportional to V, which the warm restart sets to zero: the retained subspace receives a learning signal scaled by $\|V\|$, and that factor is reset at every restart. The warm restart also rotates U, leaving the carried moments in a frame that has moved, though Corollary 7 bounds that rotation by $\arctan \|S\|_2$. Runtime is essentially unaffected by the restart flag in both parameterizations.

Seed variability on the parameter-matched arms. The three rows the comparison turns on, all at block 32 and 17.55M, were each run three times. The trigger averages 47.18 ± 0.10 on target and 44.58 ± 0.11 on general capability, the fixed period 47.12 ± 0.16 and 44.45 ± 0.13, and OFTv2 46.79 ± 0.50 and 44.33 ± 0.07. The ordering holds on both coordinates in the mean; on target the three ranges overlap, and on general capability the trigger's weakest run stands above OFTv2's strongest. The arm without a restart was also run three times, 43.83 ± 0.35 on target and 33.62 ± 0.42 on general capability; its strongest general run, 34.04, sits 10.29 below the period's weakest. The values printed above are the medians of the three in every case. Every other row is a single run under the benchmark's own protocol.

H Additional Measurement and Ablation Detail

This appendix records the truncation-order measurement protocol of Section 3, the full schedule-selection protocol, and the secondary measurements summarized in Section 7.

Changing the optimizer's dynamics moves the period's geometry and not the budget's. Across six variants of learning-rate schedule, scale and horizon on CoLA, the frozen one-epoch period realizes maximum drifts from 0.12 to 0.32 over eighteen runs; the frozen tolerance holds its peaks between 0.2000 and 0.2022 while its trigger count adapts from one restoration per run to ten. A period fixes the timing and discovers the geometry; a budget fixes the geometry and lets the timing adapt.

The reach of the high-rate separation. The separation appears only at $2\times$ the tuned rate: at tuned rates the three families move by at most 2.2 points out to the largest drift any of them realizes, 1.08 on the subspace support (Figure 4); what fails at $2\times$ is the joint excursion of rate and geometry.

Table 16: Full stability–plasticity comparison on PEFT-Arena (Huang et al., 2026), mathematics domain, Qwen2.5-7B, supervised finetuning; the panel in Figure 2(c) is the budget-matched slice of this table. We follow the same protocol throughout. Bold marks the best value in each column; underlines mark the strongest OFT setting reported by PEFT-Arena, the family RTOFT-B replaces. We retain PEFT-Arena's released "OFT" label for these baseline rows; its released implementation for this baseline corresponds to the OFTv2 construction used in our matched comparison. The first RTOFT-B row runs the fixed one-epoch period and the second the drift trigger at $\tau = 0.10$, and the third removes the restart; all three are medians of three runs, with the spreads below the table.

Method	Setting	Tr. Param	Target (%) ↑	General (%) → 0
Base	N/A	7.61B	35.30 (+0.00)	46.97 (+0.00)
Full finetuning	N/A	7.61B	**50.63** (+15.33)	34.22 (-12.75)
OFT	block 16	8.49M	42.33 (+7.03)	42.58 (-4.39)
OFT	block 32	17.55M	46.93 (+11.63)	<u>44.37</u> (-2.60)
OFT	block 64	35.68M	46.23 (+10.93)	35.97 (-11.00)
OFT	block 128	71.92M	<u>47.77</u> (+12.47)	36.98 (-9.99)
LoRA	r4a8	10.09M	42.33 (+7.03)	41.66 (-5.31)
LoRA	r8a16	20.19M	42.47 (+7.17)	39.22 (-7.75)
LoRA	r16a32	40.37M	44.87 (+9.57)	34.91 (-12.06)
LoRA	r32a64	80.74M	45.37 (+10.07)	38.21 (-8.76)
AdaLoRA	r8a16	30.28M	40.43 (+5.13)	35.41 (-11.56)
PiSSA	r8a16	20.19M	44.53 (+9.23)	24.78 (-22.19)
MiLoRA	r8a16	20.19M	42.60 (+7.30)	37.62 (-9.35)
KeepLoRA	r8	20.19M	40.53 (+5.23)	43.75 (-3.22)
MiSS	r8	11.12M	43.17 (+7.87)	39.12 (-7.85)
MiSS	r64	89.00M	46.93 (+11.63)	32.77 (-14.20)
VeRA	r256	1.44M	39.43 (+4.13)	**47.25** (+0.28)
DoRA	r8a16	21.58M	42.33 (+7.03)	40.25 (-6.72)
IA^3	N/A	1.82M	35.13 (-0.17)	44.71 (-2.26)
RTOFT-B (one-epoch period)	block 32	17.55M	47.21 (+11.91)	44.44 (-2.53)
RTOFT-B (ours)	block 32	17.55M	47.19 (+11.89)	44.59 (-2.38)
RTOFT-B w/o restart	block 32	17.55M	43.83 (+8.53)	33.62 (-13.35)

Table 17: Training cost for controllable generation, in *h:mm:ss*, under the same setting as Table 5, with each time column also given relative to LoRA. Every row is trained here under one pipeline, on the same hardware and step budget. Peak memory is 53GB for every row, spanning 53.14 to 53.35GB.

Method	Config	#Param (M)	S2I time ↓	×LoRA	L2F time ↓	×LoRA
LoRA	$r = 16$	2.50	**21:09:37**	**1.000**	**15:34:26**	**1.000**
OFT	$b = 4$	26.40	23:28:10	1.109	17:57:13	1.153
COFT	$b = 4$	26.40	23:31:53	1.112	18:00:33	1.156
OFTv2	$c = 2$, skew	26.34	22:58:47	1.086	17:54:42	1.150
BOFT	$m = 4, bn = 16$	26.40	30:29:12	1.441	24:29:31	1.573
HRA	$r = 8, \lambda = 10^{-3}$	0.89	32:21:08	1.529	26:20:44	1.692
RTOFT-B (one-epoch period)	$b = 2$	26.35	21:26:06	1.013	15:51:08	1.018
RTOFT-B (one-epoch period)	$b = 4$	13.15	<u>21:11:35</u>	<u>1.002</u>	15:50:56	1.018
RTOFT-B w/o restart	$b = 2$	26.35	21:25:58	1.013	<u>15:50:52</u>	<u>1.018</u>
RTOFT-B (ours)	$b = 2$	26.35	21:35:36	1.020	15:58:10	1.025

Projection along the trajectory. Along matched twenty-epoch trajectories evaluated at five checkpoints, the no-restart arm's drift saturates at 0.48–0.66 while the periodic arm's falls to 0.005. Projecting the no-restart checkpoint onto the orthogonal group at any of the five moments subtracts rather than repairs, averaging -1.36 over the fifteen CoLA evaluations. At the endpoint the same projection moves retention from -13.35 to -13.11.

Controls at the doubled learning rate. At twice the tuned SST-2 rate the period's best seed falls below the trigger's worst. Restoring after every step reaches 95.5 where the drift trigger reaches 95.42, and a fixed period handed the trigger's own restart count reaches 94.44. What fails at the doubled rate is unmaintained geometry rather than the learning rate, and it is where the restarts are placed rather than how many of them there are.

Table 18: Hyperparameters for RTOFT on DeBERTaV3-base under the validation-score protocol of Table 3, at block size $b = 16$. The setup follows Liu et al. (2024) and Yuan et al. (2024).

	MNLI	SST-2	CoLA	QQP	QNLI	RTE	MRPC	STS-B
Batch size	32	32	32	32	32	32	32	32
Epochs	8	10	34	12	12	34	60	11
Learning rate	2E-04	2E-04	9E-04	1E-03	8E-05	1E-03	1E-03	1.5E-03
Warm-up steps	1000	500	100	1000	500	50	50	50
Dropout	0	0	0	0	0	0	0	0
Max. seq. len.	256	128	64	320	512	320	320	128

Table 19: Learning rates for RTOFT on DeBERTaV3-base under the test-split protocol of Table 2, at block size $b = 32$. The split and the remaining settings follow Wu et al. (2026); the learning-rate grid is stated in Appendix J.

	CoLA	STS-B	MRPC	RTE	SST-2	QNLI
Learning rate	1E-04	1E-03	8E-04	5E-04	1E-04	1E-04
Epochs	20	20	30	30	10	5
Max. seq. len.	64	128	256	256	128	256
Batch size	32	32	32	32	32	32

Per-block extremes. Table 21 records the minimum, median and maximum block defect at each truncation order over the same checkpoints; medians repeat Table 1. At the default k=5 the per-block defect spans 1.3×10^{-7} to 1.36×10^{-2}, five orders of magnitude under a single convention.

	CoLA			RTE			STS-B		
k	min	median	max	min	median	max	min	median	max
2	7.7×10^{-4}	5.4×10^{-3}	3.8×10^{-2}	3.9×10^{-3}	2.7×10^{-2}	1.9×10^{-1}	1.6×10^{-2}	6.4×10^{-2}	2.6×10^{-1}
3	2.3×10^{-11}	7.9×10^{-9}	2.7×10^{-6}	2.7×10^{-9}	9.3×10^{-7}	3.2×10^{-4}	1.8×10^{-7}	1.2×10^{-5}	8.1×10^{-4}
4	2.3×10^{-11}	7.9×10^{-9}	2.7×10^{-6}	2.7×10^{-9}	9.3×10^{-7}	3.2×10^{-4}	1.8×10^{-7}	1.2×10^{-5}	8.1×10^{-4}
5	1.3×10^{-7}	6.2×10^{-6}	3.0×10^{-4}	3.1×10^{-6}	1.5×10^{-4}	7.3×10^{-3}	4.8×10^{-5}	8.1×10^{-4}	1.36×10^{-2}
6	3.5×10^{-18}	5.6×10^{-14}	8.9×10^{-10}	9.5×10^{-15}	1.5×10^{-10}	2.4×10^{-6}	8.5×10^{-12}	9.8×10^{-9}	1.1×10^{-5}

Table 21: Per-block orthogonality defect at each truncation order: minimum, median and maximum over all blocks of the reference implementation's own checkpoints.

Measured truncation orders. The reference implementation obeys the odd–even rule at each of its own settings, and the default sits on the wrong branch:

k	2	3	4	5 *(default)*	6
m	N/A	3	3	4	5
measured order of $\|\widetilde{R}_k^\top \widetilde{R}_k - I\|_2$	2.00	6.00	6.00	4.00	10.00
block multiplications	0	2	2	3	4

The orders are read from local slopes along a ladder in $\|Q\|_2$; the exact branch of the same code returns 1.1×10^{-15}, machine precision.[1]

The same two orders predict how the defect ratio evolves over training. Orders 6 and 4 put the ratio of the k=5 to the k=3 defect at $\|Q\|_2^{-2}$, so it should shrink as training grows the parameter. Over the

[1] The two branches are not approximations of one another. The exact branch solves for $(I - Q)(I + Q)^{-1}$ while the series converges to $(I + Q)(I - Q)^{-1}$; the two are transposes, so a checkpoint trained under one setting has every rotation inverted under the other.

Table 20: Configuration for the commonsense reasoning experiments of Table 4, following the Commonsense-15K protocol used by SVFT (Lingam et al., 2024), based on the eight-task LLM-Adapters benchmark (Hu et al., 2023). Every compared method uses this configuration and the same adapter placement.

Setting	Value
Backbone	Qwen3.5-9B (base)
Training data	`commonsense_15k` (15K examples)
Target modules	`q_proj`, `k_proj`, `v_proj`, `in_proj_qkv`, `up_proj`, `down_proj`
Batch size / micro-batch	64 / 2
Epochs	3
Warm-up steps	100
Weight decay	0
Cutoff length	512
Validation set size	120
Evaluation / checkpoint interval	every 233 steps
Train on inputs	yes

thirty-nine checkpoints of Figure 5 it falls from 89.6 to 11.2, an eightfold contraction that puts the growth in $\|Q\|_2$ at 2.83; the two curves read separately, through their own exponents, put it at 2.94 and 2.91. The orders are measured along a ladder in $\|Q\|_2$ at one checkpoint and the ratio along a trajectory over training, and the three estimates agree to 4%.

The cheaper truncation downstream. The ordering above is a statement about the construction; whether it costs anything downstream is a separate question. We train the $k = 3$ inner map under the same protocol, learning-rate grid and five seeds as the default $k = 5$, on the six GLUE tasks of Table 2. The paired six-task difference is $+0.03 \pm 0.05$ points in favour of $k = 3$, and its wall-clock is 3.6% lower, consistent with the one block multiplication it does not perform. The extra multiplication the default performs therefore buys neither a smaller defect nor a better score.

The default construction along its trajectory. Table 1 measures one checkpoint per task. Measuring every epoch of one run gives the same reading a second way and fixes when the crossing happens: the defect rises monotonically in the mean, passes τ^2 before a quarter of training is done, and never returns (Figure 5).

Choosing the schedule. Following the practice of tuning schedules on shorter runs (Khaled et al., 2026), we sweep six fixed periods and three tolerances on CoLA and RTE at a fixed ten-epoch budget, learning rate held at its main-table value, every setting run under one launcher on the same seeds (Figure 9). The selection rule, fixed before the runs, is the one-standard-error rule: keep every setting within one standard error of the task's best, then take the cheapest.

Every schedule that returns lands on one plateau. On CoLA the six periods span 2680 restarts down to none and a fiftyfold range of realized drift, and their scores lie between 68.70 and 69.50, inside a single standard error of each other; restarting at every step, which holds the drift to 0.008, is no better than restarting once an epoch at 0.208, and per-step exactness spends 2680 restarts buying no measurable accuracy. Never returning is the failure: it is the worst arm in every setting of Table 6, and even at this short budget it is excluded on RTE, at 81.73 against a band of 82.69. What the schedule buys is the return, not the tightness.

The rule selects a tolerance rather than a period. On CoLA the band admits everything except $\tau{=}0.05$; on RTE it admits only $\tau{=}0.20$; the intersection is $\tau{=}0.20$ alone, the highest scorer on both tasks, 69.57 and 83.01, reached with 3.8 and 3.0 restarts against the one-epoch period's ten. Theorem 1 prices the difference pointwise: at $\tau{=}0.20$ the in-flight defect is 0.040 on CoLA and 0.041 on RTE, against 0.19 and 0.23 with no restart at all. The same contrast is visible inside the sweep: the one-epoch period realizes drifts of 0.208 and 0.272 on the two tasks, thirty percent apart, while $\tau{=}0.20$ realizes 0.2006 and 0.2037, one and a half percent apart: the period lands near

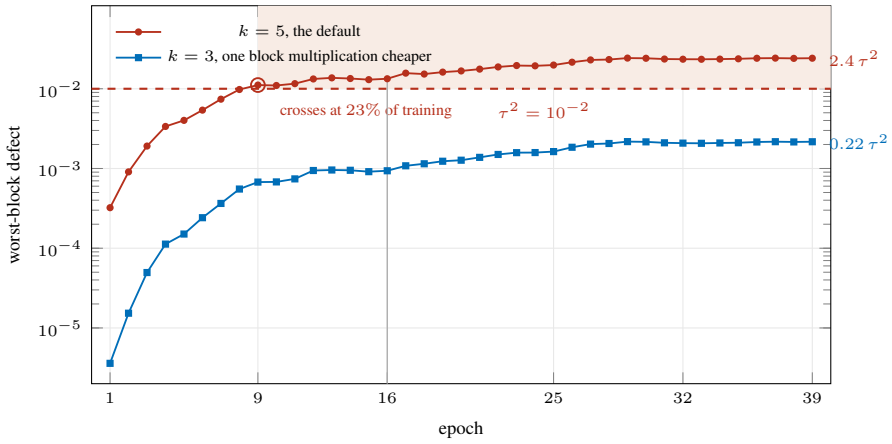


Figure 5: The default construction on its own trajectory. Worst-block defect $\|\widetilde{R}_k^\top \widetilde{R}_k - I\|_2$ at every epoch of an OFTv2 run on STS-B, at the default order and at the order that computes one block multiplication fewer, against the $\tau^2 = 10^{-2}$ a restart at $\tau = 0.10$ permits anywhere. Both curves are read from the same thirty-nine checkpoints and differ only in the truncation. The default order crosses at epoch 9 and never returns below the line; the cheaper one is inside the budget at all thirty-nine. The vertical line marks the checkpoint the run returns, which is the one measured in Table 1.

the right operating point on one task and overshoots on the other, and the tolerance lands on it by construction, which is why the two defects above agree to a third digit.

What the main tables use. The drift trigger at $\tau = 0.10$, with the fixed one-epoch period reported beside it as the open-loop control. Corollary 6 sets the tolerance from the stretch budget; the two values are that corollary read at two budgets, a certified defect of 0.040 at τ=0.20 against $\tau^2 = 10^{-2}$ at the shipped one. The sweep establishes the kind of schedule, a tolerance rather than a period. Shipping the tighter value buys the tighter certificate and pays for it in restarts. The period, the simplest schedule that returns, settles whether per-step exactness is necessary; the tolerance settles which quantity should decide when restoration happens. One epoch establishes that exactness can be amortized; drift is how the amortization should be controlled. On CoLA the period and the selected tolerance land at the same operating point, 0.208 against 0.2006; on RTE the period is the looser of the two. At the twenty epochs the main tables use, the restart is worth 0.80 points on CoLA, 72.52 against 71.72 (Table 10). A longer run without restarts accumulates further from the group, and has more to lose by never returning. The trajectory behind that peak is what a fixed period cannot control. Drift is not a stable operating point within a single run: on CoLA it reaches 0.21 in the first epoch and falls to 0.011 by the tenth, a factor of eighteen, and RTE behaves the same way. A fixed period therefore leaves the iterates least constrained exactly where the tangent parameter is largest, and goes on restarting long after there is anything left to restore. The shape tracks the decaying learning rate more closely than it tracks convergence, which is the reason a period cannot be tuned against it: it moves whenever the schedule does.

Nor does a fixed period transfer between models. At M equal to one epoch, the largest $\|S\|_2$ reached anywhere in a run is 0.170 on commonsense reasoning, 0.208 on CoLA, 0.272 on RTE and 0.799 on the stability–plasticity protocol, all four at $b = 32$; segmentation-to-image reaches 1.65 at $b = 2$, where the maximum runs over far more blocks and is not directly comparable. Across the four settings that share a block size, one identical schedule delivers in-flight defects from 0.029 to 0.638, a factor of twenty-two, and in the order the ablations imply: the setting whose transform wanders furthest from the orthogonal group is the setting where failing to return to it costs the most, PEFT-Arena giving up 10.82 points of retained capability without the restart against commonsense's 0.39 points of accuracy at the lowest drift of the five. At 0.799 the tangent parameter is not small:

$I + S$ stretches some direction by 28% (Corollary 5) and the discarded second-order terms are not negligible, yet this is where the restart gains most, because each restart returns S to zero.

Selecting a tolerance instead of a period inverts the dependency: by Corollary 2 it bounds the defect by τ^2 everywhere it is monitored. Tolerances of 0.05, 0.10 and 0.20 realize drift maxima of 0.0521, 0.1018 and 0.2006 on CoLA (Figure 9). The trigger overshoots by the growth of a single step, 4.2%, 1.8% and 0.3% of τ, since $\|S\|_2$ is read after each step and can cross τ between two reads; the realized guarantee is τ plus one increment, and it tightens as τ grows because that increment does not grow with it. What the tolerance does not control is accuracy: 67.59 ± 1.95, 68.90 ± 2.37 and 69.57 ± 2.10, against a fixed-period arm spanning 68.70 to 69.50 across its six periods, every pairwise difference inside its own noise.

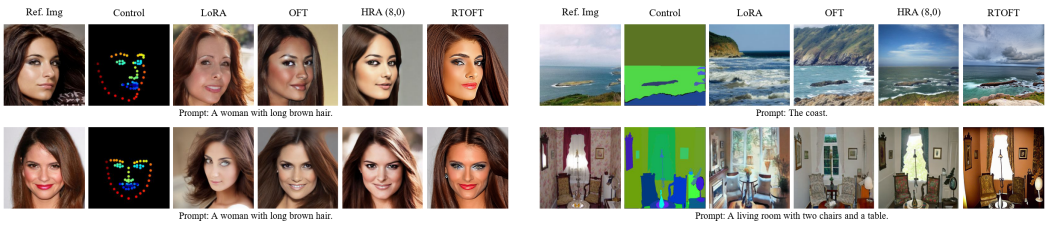
The useful comparison is cost. At $\tau = 0.20$ the trigger holds the drift the fixed period reaches, 0.2006 against 0.208, using 3.8 restarts rather than ten.

It is the period, not the trigger, that tracks the learning-rate schedule. Repeating both under a constant learning rate in place of the decaying one moves the drift a fixed period realizes by 79% on CoLA, from 0.208 to 0.372, and by 69% on RTE, from 0.272 to 0.461, while the drift the tolerance realizes does not move at all. The period therefore fails to transfer in two independent ways, and the tolerance in neither of them: across models the fixed period spans a factor of twenty-two in defect and the tolerance a factor of 1.02 in drift; across learning-rate schedules the fixed period moves by most of a factor of two and the tolerance does not move. A period is an input whose realized guarantee is discovered after the run. A tolerance is the guarantee. A truncation order is the same kind of input as a period: it too fixes something other than the departure, and leaves the departure to be read off afterwards.

What the tolerance costs is two things the period does not pay: more restarts on the settings that drift fastest, and one reading of $\|S\|_2$ per step. Over the six GLUE tasks its wall-clock nonetheless matches the period's, 133:07 against 133:11, a difference below the one-percent resolution of Appendix D, because on the slowest-drifting tasks it restarts far less often than once an epoch: two times on CoLA against the twenty a one-epoch period performs over that task's twenty-epoch budget. On the GLUE test split (Table 2) the frozen tolerance is ahead of the period by a paired 0.45 ± 0.13 points over the five shared seeds, positive on every seed; it takes five of the six tasks and gives up 0.13 on MRPC. Under the validation protocol of Table 3, eight tasks at $b = 16$ on the same five seeds, the tolerance is ahead of the period by a paired 0.53 ± 0.02 points, 90.48 against 89.95, positive on every seed and on every one of the eight tasks, by 0.09 on QQP to 1.50 on RTE; its seed spread is no larger than the period's on any task and is 0.52 against 0.58 on the eight-task average. On the test split it is also the steadier arm, and on every task: its seed deviation runs 0.84, 0.14, 0.92, 0.31, 0.16 and 0.14 against the period 1.08, 0.36, 1.62, 0.84, 0.39 and 0.37, 0.42 against 0.78 in the mean. A schedule that answers to the drift a run realizes absorbs what the seed changes; one that answers to the step count cannot.

The schedule also adapts inside a single run. As the learning rate decays the drift accumulates more slowly, so the gaps between triggers lengthen on their own: on CoLA under the validation protocol of Table 3, a 9112-step budget, a tolerance of 0.15 takes the interval from 140 steps to 1320, a factor of 9.4, and the same widening appears at every tolerance we log, by 3.1 and by 7.4. The tolerance sets the density as well: moving it from 0.15 to 0.40 takes the mean interval from 425 steps to 4140. A fixed period can do neither, because a period is a statement about training time and the quantity being controlled is not.

Order against schedule: codebase and per-task controls. Both $k{=}2$ arms are ours and both $k{=}5$ arms are the reference implementation with our restart added, so each within-row comparison of the order–schedule table uses one codebase. Across rows the codebases differ; running $k{=}2$ through the reference code path moves it by 0.18 points, a fifth of the 0.86 gap. One task carries most of the $k{=}5$ loss: the restart improves $k{=}5$ on five of the six tasks, by 0.16 to 0.52 points, and costs 3.03 on RTE, where the 85.45 it reaches is three points below OFTv2's own score; excluding RTE the restart is worth $+0.35$ at $k{=}5$ against $+1.27$ at $k{=}2$. The three Neumann terms OFTv2 ships are worth $+1.04$ points when nothing replaces the exactness they supply and -0.86 when something does.



(a) Landmark-to-face on CelebA-HQ.

(b) Segmentation-to-image on ADE20K.

Figure 6: Controllable generation with Stable Diffusion, two examples per task; the complete grids are in Figures 7 and 8. All methods use the settings of Table 5. HRA (8,0) denotes $\text{HRA}_{r=8,\lambda=0}$, the configuration these panels were generated under; the tables report HRA at $\lambda = 10^{-3}$, the value used by Yuan et al. (2024) in their controllable-generation experiments. RTOFT denotes RTOFT-$\text{B}_{b=2}$.

The α dial moves the drift as predicted. Retaining a residual $(1-\alpha)S$ predicts drift ratios of 3, 1 and 0.33 across $\alpha \in \{0.25, 0.5, 0.75\}$; the realized ratios are 2.39, 1 and 0.31 on SST-2 and 2.42, 1 and 0.36 on QNLI.

Batch geometry of the stability–plasticity runtime comparison. Sequences of 8192 tokens, a micro-batch of two per device, and sixteen gradient-accumulation passes make one optimizer step cost about a minute; the measured per-step gap is 56.3 ± 2.0 seconds against 55.9 ± 1.9, 0.7%, a factor of five below the 3.5% run-to-run spread.

I Additional Qualitative Results

Figures 7 and 8 give the complete comparison grids for controllable generation, from which the examples in Figure 6 are drawn. The grids contain every generated example.

J Implementation Details

J.1 Natural Language Understanding

The backbone is DeBERTaV3-base (He et al., 2023), and we report two protocols.

The validation-score protocol of Table 3 follows Liu et al. (2024) and Zhang et al. (2023b): it reports the best score on the original validation split, so that our numbers are directly comparable with those already published. Under this protocol the baseline rows are quoted from the literature, while OFTv2 (Qiu et al., 2025b) and RTOFT under both schedules were re-run by us in an identical pipeline at block size $b = 16$ on the same five seeds. The full setup is given in Table 18 and follows the GLUE configuration of Liu et al. (2024) and Yuan et al. (2024).

The test-split protocol of Table 2 follows Wu et al. (2026), which adopts it from earlier work on GLUE evaluation (Wu et al., 2024a;b; Bini et al., 2025): each task's public validation set is cut in two under a fixed seed; the first half chooses the checkpoint and the second, untouched while that choice is made, supplies the number we print. The learning rates we use are given in Table 19, at block size $b = 32$.

CoLA is scored by Matthews correlation, STS-B by Pearson correlation, and the four tasks that remain by accuracy. Each reported average is the arithmetic mean over the tasks in that table.

Two points apply to both protocols. Carried unchanged to commonsense, the frozen tolerance exceeds the period's eight-task average over five seeds, 88.38 against 87.23, and is the only schedule above every baseline in Table 4. The paired six-task margins carry $t(4) = 4.3$ against OFTv2, $t(4) = 6.9$ against PSOFT and $t(4) = 5.1$ against LoRA, positive on every one of the five shared seeds against all three. The learning-rate grid is identical for every compared method: the search range is 5×10^{-5} to 5×10^{-3}, probed first at 10^{-4}, 5×10^{-4} and 10^{-3}, then refined at one or two values between the best two, with each method keeping its own best. Every number we print is a



Figure 7: Complete landmark-to-face comparison on CelebA-HQ.

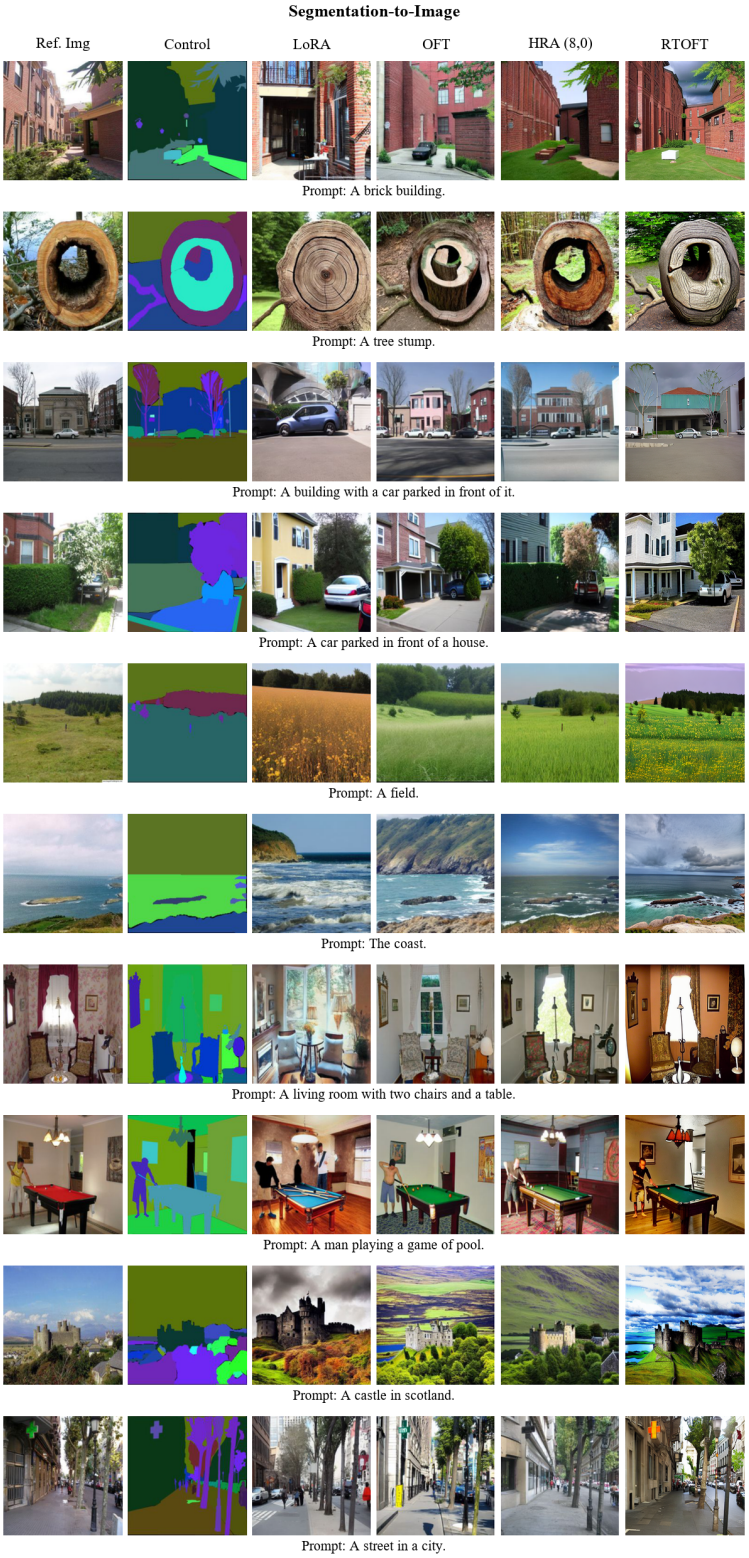
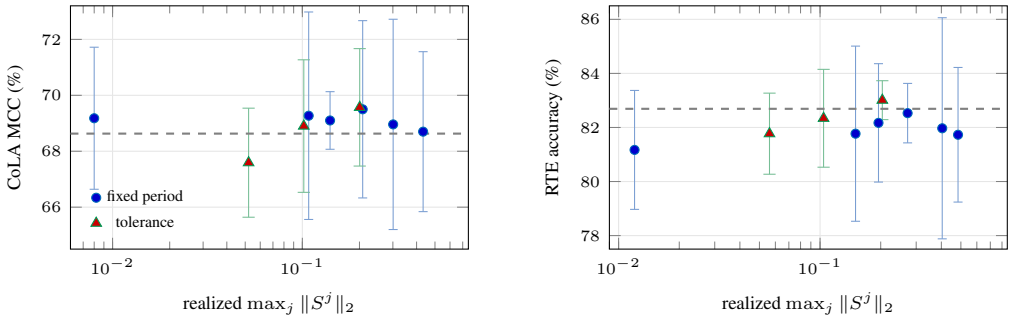


Figure 8: Complete segmentation-to-image comparison on ADE20K.



(a) CoLA, 5 seeds

(b) RTE, 3 seeds

Figure 9: Fixed restart periods and drift tolerances at a fixed ten-epoch budget (RTOFT-B b=32; mean $\pm$ SD; setup in Appendix J). Both controls are plotted against the drift they actually realize, the quantity Theorem 1 prices pointwise; the periods run from 2680 restarts down to none. Dashed lines mark one standard error below each task's best. On CoLA nothing separates: a fiftyfold range of realized drift moves the score by less than its own noise. On RTE the band admits only τ=0.20, which reaches it with three restarts.

mean of five seeds, whose deviations are in Table 11. Wall-clock and peak memory come from one NVIDIA V100.

The restart-period sweep of Figure 9 uses the same machinery at a fixed ten-epoch budget for every period: RTOFT-B at $b = 32$, the learning rate fixed at the main-table value for each task, five seeds on CoLA and three on RTE, and the period counted in optimizer steps with the restart executed before the next update. Restarts at the terminal step are not performed. Drift is recorded as the running maximum of the blockwise spectral norm $\max_j \|S^j\|_2$, evaluated immediately before each restart.

J.2 Commonsense Reasoning

Every compared method is tuned on one shared learning-rate grid spanning 10^{-4} to 5×10^{-3}, searched as on GLUE, with each method keeping its own best. The backbone is Qwen3.5-9B (base), with 8.95B total parameters.[2] We evaluate on the eight-task suite of Table 4, comprising BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC-e, ARC-c and OBQA. The full configuration is given in Table 20. Every compared method is trained under it, with the same adapter placement.

We follow the Commonsense-15K protocol used by SVFT (Lingam et al., 2024), which takes a 15K instruction subset of the LLM-Adapters commonsense benchmark (Hu et al., 2023) and evaluates on the same eight tasks; PSOFT (Wu et al., 2026) subsequently uses the same setup: a single training run on that subset. Every method in Table 4 was run by us under this protocol, on the five seeds used elsewhere in the paper rather than the single run it asks for. HRA (Yuan et al., 2024) exceeds device memory at this scale and is reported as such. Training time is wall-clock in *mm:ss*, as everywhere else in the paper except the diffusion experiments.

BoolQ calibration. The periodic schedule's BoolQ deficit is a majority-class collapse. Answer extraction succeeds on every response (0 unparseable of 3270). At seed 42 the periodic arm answers "true" on 90.9% of items (2973 of 3270) against a base rate of 62.2% "true" labels (2033 of 3270); at seed 9 it answers "true" on 80.9% (2646). The tolerance arm answers "true" on 66.3% (2167), the closest of the three to the data's rate.

J.3 PEFT-Arena Stability–Plasticity Evaluation

We adopt the released protocol of Huang et al. (2026) without modification: the Qwen2.5-7B backbone, the mathematics domain, and the supervised-finetuning regime. Target performance is measured on Math-500, AMC23 and AIME24, and general performance on IFEval, NQ and BBH, each

[2] https://huggingface.co/Qwen/Qwen3.5-9B

reported together with its change against the un-adapted base model. We follow the same protocol for every row: same data, budget, optimizer and evaluation harness. We run RTOFT-B at block size 32 so that its trainable-parameter count matches the OFT block-32 entry exactly, which is the comparison the table is built around. Every remaining setting is theirs unchanged.

Timing comparisons are in Tables 2, 4 and 17.

J.4 Controllable Generation

The drift-triggered row of Table 5 uses $\tau = 0.10$ with $\|S\|_2$ evaluated at every optimizer step; at $b = 2$ the spectral norm of each block is the absolute value of its single parameter, so the monitor is a single fused reduction. All other settings are identical to the one-epoch arm. The trigger fires $1{,}872$ times over $27{,}820$ steps (6.7% of checks); the realized post-step maximum is 0.12, every forward pass below τ; peak memory unchanged from the period arm; wall-clock $21{:}35{:}36$ against its $21{:}26{:}06$, an overhead of 0.7%.

The trajectory comparison of Section 7 dumps the tangent state of matched no-restart and one-epoch arms at $20/40/60/80/100\%$ of training (CoLA and STS-B, main protocol, multiple seeds) and evaluates each dump twice offline, as parameterized and after a polar merge on a copy; the offline evaluation of the first dump reproduces the online evaluation exactly.

The dynamics-perturbation comparison of Section 7 freezes the matched pair, the one-epoch period and $\tau = 0.20$, which realize the same operating point on CoLA, and varies the learning-rate schedule (cosine, linear, constant), its scale ($0.5\times$, $2\times$), and the horizon (20 epochs), multiple seeds each under the ten-epoch sweep protocol with the monitor evaluated at every step. We follow the protocol of Qiu et al. (2023), Liu et al. (2024) and Yuan et al. (2024) on Stable Diffusion (Rombach et al., 2022): segmentation-to-image on ADE20K and landmark-to-face on CelebA-HQ, with the control signal injected in the same way for all compared methods. Control fidelity on segmentation-to-image is measured by segmenting the generated images and comparing the result against the conditioning map (mIoU, mAcc, aAcc). On landmark-to-face it is the landmark error between the control signal and the prediction. Baselines are DreamBooth (Ruiz et al., 2023), ControlNet (Zhang et al., 2023a), T2I-Adapter (Mou et al., 2024), LoRA (Hu et al., 2022), OFT and its constrained variant (Qiu et al., 2023), BOFT (Liu et al., 2024) and HRA (Yuan et al., 2024). Following Liu et al. (2024), every compared method is run on one shared grid: trainable-parameter budgets at roughly 1–2M, 10M, 20M and 26M, and learning rates from 10^{-4} to 10^{-3}. The entry reported for each method is its own best on that grid. LoRA's best is $r = 16$; HRA is reported at $r = 8, \lambda = 10^{-3}$, the configuration used by Yuan et al. (2024) in their controllable-generation experiments. All rows share hardware, step budget and batch size. Training cost in Table 17 is reported in *h:mm:ss*.