PALoRA: Projection-Adaptive LoRA for Preserving Reasoning in Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Efficiently updating Large Language Models (LLMs) with new or evolving factual knowledge can erode previously acquired reasoning abilities. Spectral parameter-efficient fine-tuning methods control this interference through a fixed-rank boundary in the singular value decomposition of pre-trained weights, treating singular-value rank as a proxy for skill relevance. We show, both theoretically and empirically, that this premise is incomplete: information essential for reasoning is distributed across the singular spectrum of multilayer perceptron weights rather than confined to either end of it. We therefore introduce PALoRA, a two-stage framework for relevance-based knowledge injection. PALoRA first trains a Singular Value Fine-Tuning (SVF) expert on a target reasoning skill and uses its learned singular scaling vector as a frozen probe of how strongly the skill relies on each singular component. It then injects factual knowledge with Low-Rank Adaptation (LoRA) while penalizing the projection of the update onto every singular direction in proportion to this reliance. We prove that this graded penalty bounds the first-order skill degradation of the update, with fixed-rank protection as its binary case. Across Llama 3.1 8B and Mistral 7B, and across mathematical, coding, and scientific reasoning benchmarks, PALoRA preserves on average 96% of the SVF expert's reasoning performance while maintaining competitive factual recall. It matches the strongest of five baselines without a fixed rank boundary, with the only positive backward transfer, retains more of the skill under sequential injection, and adds less than 0.006% parameter overhead.

1 Introduction

Large language models (LLMs) are trained on static corpora, but deployed systems must continually acquire corrected and newly emerging facts (Jang et al., 2022). Parameter-efficient fine-tuning (PEFT), especially LoRA (Hu et al., 2022), makes such updates computationally feasible. Yet growing evidence shows that factual knowledge injection can come at the cost of core reasoning abilities, degrading performance on mathematical, coding, and scientific tasks (Pletenev et al., 2025). This tension is particularly acute in Transformer MLP layers, which are known to support both factual recall and procedural reasoning (Geva et al., 2021; Dai et al., 2022; Cobbe et al., 2021; Austin et al., 2021; Clark et al., 2018; Sharma et al., 2024).

The rank cut-off premise. Recent spectral PEFT methods address this trade-off through the singular value decomposition of pre-trained weights, adapting its principal or minor components or protecting its top ones (Meng et al., 2024; Wang et al., 2025; Lingam et al., 2024; Xiong & Xie, 2026; Yang et al., 2024). Different as they are, these approaches share one premise: the components worth preserving are singled out by a cut-off in singular-value rank, and a method need only decide which side of the cut to adapt and which to protect. We argue that this view is incomplete (Hsu et al., 2022). A reasoning skill need not align with the directions that explain the most variance of the pre-trained model, and the components it relies on may be distributed across the singular spectrum rather than confined to either end of it; protecting a fixed top-k subspace, however strictly, then leaves the skill loading outside it exposed.

We test this premise directly. Measured by a Singular Value Fine-Tuning (SVF) expert (Sun et al., 2025), the singular components a reasoning skill relies on are spread across the spectrum and nearly

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.

disjoint from the leading ones, yet protecting either set preserves the skill equally well; adapting the minor components protects no more than unconstrained LoRA; and OPLoRA (Xiong & Xie, 2026), the strongest rank-based protection, is matched by a constraint that does not single out the leading directions. Singular-value rank alone thus does not determine which components to protect: protecting the top directions is one way to retain a skill, not a necessary one. This motivates a shift from rank-based to relevance-based protection, estimating the importance of every singular direction instead of selecting a protected region in advance.

PALoRA (Projection-Adaptive Low-Rank Adaptation) implements this shift as a two-stage framework that explicitly decouples stability and plasticity within shared MLP weights. In the first stage, we train the SVF expert on a target reasoning skill in a short policy-gradient run and freeze it; the amount by which it rescales a singular component measures how strongly the skill relies on it, whatever that component's singular-value rank. In the second stage, we inject factual knowledge via a LoRA adapter, penalizing the projection of its update onto every singular direction in proportion to that reliance, so that it acquires new knowledge in the directions the skill leaves free, wherever they lie.

Contributions. This work makes three contributions. First, we show that the rank cut-off shared by spectral PEFT methods does not locate the reasoning skills of MLP weights: the directions an SVF expert relies on are distributed across the spectrum, protecting them or the leading directions preserves the skill equally, and confining an update to the minor components does not protect the skill (Sections 4.2 and 4.3). Second, we give a theoretical account of this finding, showing that under a formal misalignment condition the SVF weights protect skill-relevant components outside the leading singular directions more than the leading ones, and that rank-fixed protection cannot exclude a first-order degradation, however strictly enforced (Section 3.3). Third, we introduce PALoRA, which grades the protection of every singular direction by its skill relevance, prove that this graded penalty bounds the skill degradation of the update, and show against five baselines sharing the same skill expert, optimization schedule, and adapter capacity that it matches the strongest rank-based protection with the only positive backward transfer, and exceeds it under sequential injection (Sections 3 and 4).

2 Related Work

Knowledge storage in Transformers. Transformer models (Vaswani et al., 2017) combine self-attention mechanisms with position-wise feed-forward, or MLP, blocks. Prior work has shown that these MLP layers play a central role in storing and retrieving parametric knowledge (Meng et al., 2022). Geva et al. (2021) characterize feed-forward layers as key–value memories: the first linear projection acts as a pattern detector, while the second maps activated patterns to corresponding output representations. Dai et al. (2022) provide complementary evidence by identifying individual "knowledge neurons" in MLP layers that mediate factual associations. This motivates our focus on MLP layers as the site where knowledge injection and skill preservation must be jointly controlled.

PEFT and the plasticity–stability trade-off. PEFT methods (Houlsby et al., 2019) differ in how they balance adaptation capacity with preservation of pre-trained capabilities. Additive methods such as LoRA (Hu et al., 2022) achieve strong plasticity by learning low-rank updates (Aghajanyan et al., 2021) on top of frozen model weights, enabling efficient acquisition of new knowledge and task-specific behaviors. In contrast, Singular Value Fine-Tuning (SVF) (Sun et al., 2025) modifies only the singular values of pre-trained weights while preserving their original singular directions, yielding a more stability-oriented adaptation mechanism.

Continual learning and interference mitigation. Catastrophic forgetting is a core challenge in continual learning (McCloskey & Cohen, 1989; Mermillod et al., 2013; Kotha et al., 2024; Shuttleworth et al., 2025). Regularization-based methods such as EWC (Kirkpatrick et al., 2017) penalize updates to parameters important for previous tasks, while orthogonality-based approaches reduce interference geometrically. OGD (Farajtabar et al., 2020) projects gradients away from previous-task directions. In PEFT, O-LoRA (Wang et al., 2023) enforces orthogonal adapter subspaces across tasks. PALoRA follows the same principle, but applies it at the level of the singular directions of the pre-trained MLP weights, weighted by the reliance an SVF expert reveals rather than restricted to a set selected in advance.

Spectral PEFT. A recent line of spectral PEFT methods uses the singular value decomposition of pre-trained weights to control adaptation subspaces. PiSSA (Meng et al., 2024) and MiLoRA (Wang et al., 2025) adapt principal and minor singular components, respectively, while SVFT (Lingam et al., 2024) learns sparse combinations of singular vectors. OPLoRA (Xiong & Xie, 2026) constrains LoRA updates to the orthogonal complement of the top-k singular vectors, and CorDA (Yang et al., 2024) preserves context-oriented principal directions using covariance statistics. Despite their differences, these methods separate the spectrum at a fixed rank, of the singular values or of a context covariance, and treat each side alike. Our work shows that reasoning skills are distributed across the singular spectrum, so that such a cut does not isolate them.

3 PALoRA: Method and Theoretical Motivation

PALoRA addresses the plasticity–stability trade-off through a two-stage adaptation framework. In the first stage, a frozen SVF expert measures how strongly a target reasoning skill relies on each singular direction. In the second stage, new factual knowledge is injected through a LoRA adapter constrained to remain orthogonal to each direction in proportion to that reliance, so that the protection follows the shape of the skill across the spectrum rather than a cut-off in singular-value rank. The two stages are illustrated in Figure 1 and summarized in Algorithm 1, and a consolidated notation reference is provided in Appendix A.

3.1 Architectural Co-location

Both the plasticity module (LoRA) and the stability module (SVF) operate directly on the MLP weight matrices of the Transformer (Vaswani et al., 2017). For a given MLP layer l with pre-trained weight matrix $\mathbf{W}_{\text{mlp}}^{(l)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the adapted weight is defined as

$$\mathbf{W}_{\text{mlp}}^{\prime(l)} = \text{SVF}\big(\mathbf{W}_{\text{mlp}}^{(l)}\big) + \text{LoRA}\big(\mathbf{W}_{\text{mlp}}^{(l)}\big), \tag{1}$$

where, writing the singular value decomposition as $\mathbf{W} = \mathbf{U}\,\text{diag}(\boldsymbol{\sigma})\mathbf{V}^\top$, the term $\text{SVF}(\mathbf{W}) = \mathbf{U}\,\text{diag}(\mathbf{z} \odot \boldsymbol{\sigma})\mathbf{V}^\top$ applies a learned scaling vector $\mathbf{z} \in \mathbb{R}^{\min(d_{\text{out}}, d_{\text{in}})}$, one entry per singular value, to the singular values, preserving the original singular directions and modulating only their magnitudes (Sun et al., 2025), and $\text{LoRA}(\mathbf{W})$ adds a low-rank update $\Delta\mathbf{W} = \mathbf{BA}$ with $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ and $r \ll \min(d_{\text{out}}, d_{\text{in}})$, introducing new trainable directions that enable rapid memorisation of novel facts (Hu et al., 2022).

This co-location is deliberate: the MLP layers are the primary site of both factual storage and procedural reasoning, so the two modules are placed on the very matrices where they compete, and the decoupling is obtained not by separating them architecturally but by governing their interaction through the training strategy described next. Equation 1 is to be read literally: the Phase 1 scaling is retained in the deployed model, so the SVF stage both selects the protected directions and forms part of the final weights. We therefore report retention against two reference points, the base model and the SVF expert, deploy the same expert under every method we compare against so that only the injection rule differs, and separate the two contributions experimentally in Section 4.4.

3.2 Two-Phase Training Framework

Phase 1: Skill Expert Creation

We train an SVF module following the standard procedure of Transformer-Squared (Sun et al., 2025), optimizing it with the REINFORCE policy-gradient algorithm (Williams, 1992) on a high-quality skill dataset (e.g., GSM8K for mathematical reasoning). For each MLP layer l, SVF learns a scaling vector $\mathbf{z}_l \in \mathbb{R}^{r_l}$ that modulates the singular values of $\mathbf{W}_{\text{mlp}}^{(l)} = \mathbf{U}_l\,\text{diag}(\boldsymbol{\sigma}_l)\mathbf{V}_l^\top$; training is short, 10 iterations that each take a single policy-gradient update over a batch of 256 sampled prompts, after which $\mathbf{z}_l^*$ is frozen and reused by every injection that targets the same skill. The deviation $|\mathbf{z}_l^* - \mathbf{1}|$ records how far the expert had to move each singular component in order to acquire the skill. Since the components it moves are spread across the whole spectrum rather than gathered at its top (Section 4.2), no cut-off in singular-value rank isolates them, and PALoRA therefore grades the protection of every component by the importance the expert assigns to it, $w_{l,i} = |z_{l,i}^* - 1| / \max_j |z_{l,j}^* - 1| \in [0, 1]$,

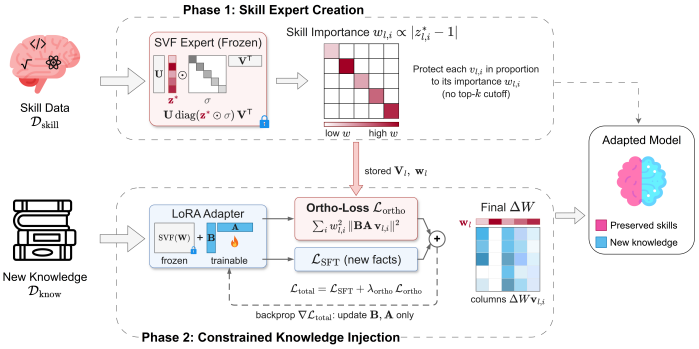


Figure 1: PALoRA. Phase 1 trains and freezes an SVF expert, whose learned scaling vector measures the skill relevance of every singular component. Phase 2 trains a LoRA adapter for new facts while penalizing projections of the update onto each component in proportion to that relevance.

Algorithm 1 PALoRA training

Input: $\mathcal{M}(\theta_{\text{base}})$, $\mathcal{D}_{\text{skill}}$, $\mathcal{D}_{\text{know}}$, λ_{ortho}
Phase 1: train SVF on $\mathcal{D}_{\text{skill}}$ for 10 iterations to obtain frozen $\{\mathbf{z}_l^*\}$
for all MLP layers l **do**
    $\mathbf{w}_l \leftarrow |\mathbf{z}_l^* - \mathbf{1}| / \max_i |z_{l,i}^* - 1|$; store $\mathbf{V}_l$ and $\mathbf{w}_l$
end for
Phase 2: deploy the expert, $\mathbf{W}_l \leftarrow \text{SVF}(\mathbf{W}_l)$, and initialize LoRA $\theta_{\text{LoRA}} = \{\mathbf{B}_l, \mathbf{A}_l\}_{l=1}^{L}$ with $\mathbf{B}_l = \mathbf{0}$
for training steps $1, \ldots, T$ **do**
    Sample minibatch from $\mathcal{D}_{\text{know}}$
    $\mathcal{L}_{\text{ortho},l} \leftarrow \|\mathbf{B}_l \mathbf{A}_l \mathbf{V}_l \,\text{diag}(\mathbf{w}_l)\|_F^2$ for all l (Eq. 3)
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{SFT}} + \lambda_{\text{ortho}} \sum_l \mathcal{L}_{\text{ortho},l}$ (Eq. 2)
    Update θ_{LoRA} using $\nabla \mathcal{L}_{\text{total}}$
end for
Return: adapted model $\text{SVF}(\mathbf{W}) + \mathbf{BA}$

instead of selecting a subset to be treated as uniformly critical. As the deviations fall off sharply (Figure 2b), the protection is far from flat: it concentrates on the components the skill rescales most and leaves the remainder nearly free. The *critical index set* $\mathcal{I}_{\text{crit},l}$, the top-$k$ entries of $|\mathbf{z}_l^* - \mathbf{1}|$, is their binary case $w_{l,i} \in \{0, 1\}$, analysed in Section 3.3 and compared with the grading in Section 4.2.

Phase 2: Constrained Knowledge Injection

With the base model and SVF expert frozen, we train a LoRA adapter (Hu et al., 2022) on a dataset of unknown facts using

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SFT}} + \lambda_{\text{ortho}} \sum_{l=1}^{L} \mathcal{L}_{\text{ortho},l}, \tag{2}$$

where $\mathcal{L}_{\text{SFT}}$ is the standard cross-entropy loss for knowledge injection, and $\mathcal{L}_{\text{ortho},l}$ is a structural orthogonality loss that penalizes the projection of the update onto each input singular direction in proportion to its protection weight, steering the adapter to inject new knowledge away from the skill-relevant pathways:

$$\mathcal{L}_{\text{ortho},l} = \sum_{i=1}^{r_l} w_{l,i}^2 \, \|\Delta \mathbf{W}_l \mathbf{v}_{l,i}\|_2^2 = \|\mathbf{B}_l \mathbf{A}_l \mathbf{V}_l \,\text{diag}(\mathbf{w}_l)\|_F^2 \,, \tag{3}$$

with $\Delta \mathbf{W}_l = \mathbf{B}_l \mathbf{A}_l$. Since $w_{l,i}$ enters squared, the directions the expert rescaled most dominate the penalty while those it left untouched stay nearly free, and the factored form, evaluated on the LoRA factors, adds little cost per step. Its value bounds the skill loss on the input side, where interference is governed (Theorem 2, Lemma 3).

3.3 THEORY: WHAT THE SVF PROBE MEASURES AND WHAT THE PENALTY GUARANTEES

If a skill relies on singular directions scattered across the spectrum, protecting only the leading ones, however exactly, leaves the skill loading outside them exposed, whereas a probe of which directions the skill uses, whatever their rank, can grade the protection of each. We formalize both along the two phases (proofs in Appendix B): for Phase 1, what the SVF deviation $|z_i^* - 1|$ estimates and when its weights protect skill-relevant components outside the leading singular directions more than the leading ones (Theorem 1); for Phase 2, a bound on the skill degradation of an update penalized by Eq. 3 (Theorem 2), whose binary case, rank-fixed protection, cannot exclude a first-order degradation (Corollary 1).

Setting and assumptions. Let $\mathbf{W} = \sum_{i \le r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \in \mathbb{R}^{m \times n}$, $\sigma_1 \ge \cdots \ge \sigma_r > 0$, and let a skill T correspond to an s-dimensional input subspace $\mathcal{U}_T$ with projector $\mathbf{P}_T$. We assume (Appendix B.1) that pre-training aligns the leading singular directions with high-variance corpus directions (Assumption 1); that the skill gradient is structured, $\nabla_{\mathbf{W}} L_T = \mathbf{G}_T + \mathbf{N}$ with $\mathrm{row}(\mathbf{G}_T) \subset \mathcal{U}_T$ and $\|\mathbf{N}\|_F \le \epsilon_{\text{noise}}$ (Assumption 3); and that the factors of $\mathbf{G}_T$ are isotropic, with curvature along the SVF coordinates bounded above and below (Assumption 5). Since a generic s-dimensional subspace already gives $\|\mathbf{P}_T \mathbf{v}_i\|_2^2 \approx s/n$, misalignment $\|\mathbf{P}_T \mathbf{v}_i\|_2^2 \le \delta$ for $i \le k$ (Assumption 2) is measured by the ratio $\mu := (s/n)/\delta$, with $\mu = 1$ at chance. Finally, the diagonal loadings $\gamma_i = \mathbf{u}_i^\top \mathbf{G}_T \mathbf{v}_i$ must not vanish (Assumption 4): a gradient such as $\mathbf{G}_T = \mathbf{u}_2 \mathbf{v}_1^\top$ rotates singular vectors without moving any singular value to first order, and is invisible to any singular-value probe. These describe Phase 1, from $\mathbf{W}_0$; Phase 2 starts from $\mathbf{W}^\star = \mathrm{SVF}(\mathbf{W}_0)$ (same singular vectors), where Assumption 6 gives $\nabla_{\mathbf{W}} L_T(\mathbf{W}^\star) = \mathbf{G}_T^\star + \mathbf{N}^\star$ the same structure, $\|\mathbf{N}^\star\|_F \le \epsilon^\star$, with L_T ℓ-smooth.

Lemma 1 (Spectral loading)**.** *Under Assumptions 3 and 5, $\mathbb{E}[\gamma_i^2] = \|\mathbf{G}_T\|_F^2 \, \|\mathbf{P}_T \mathbf{v}_i\|_2^2/(sm)$: the loading of a component scales with the skill overlap of $\mathbf{v}_i$, and with nothing else.*

Lemma 2 (Two regimes of the SVF deviation)**.** *Let $c_i > 0$ be the Gauss–Newton curvature of L_T along $\mathbf{u}_i \mathbf{v}_i^\top$ and $\nu_i = \mathbf{u}_i^\top \mathbf{N} \mathbf{v}_i$. Diagonal gradient flow from $\mathbf{z}(0) = \mathbf{1}$ gives $z_i(t) - 1 = -(\gamma_i + \nu_i)\big(1 - e^{-\sigma_i^2 c_i t}\big)/(\sigma_i c_i)$, so that $|z_i - 1| \approx t\,\sigma_i |\gamma_i + \nu_i|$ in the transient regime $\sigma_i^2 c_i t \ll 1$ and $|z_i - 1| \to |\gamma_i + \nu_i|/(\sigma_i c_i)$ in the saturated regime $\sigma_i^2 c_i t \gg 1$.*

The two regimes have opposite dependence on σ_i; leading components saturate first (crossover at $\sigma_i \approx (c_i t)^{-1/2}$). With Lemma 1, the saturated score is $|z_i^* - 1| \asymp \|\mathbf{P}_T \mathbf{v}_i\|_2/(\sigma_i c_i)$ up to a common factor: skill overlap amplified by the inverse singular value.

Theorem 1 (Detection of off-principal skill components)**.** *Let $i^* > k$ have chance-level overlap, $\kappa = c_{\max}/c_{\min}$, and $R := |z_{i^*}^* - 1| / \max_{i \le k} |z_i^* - 1| = w_{i^*} / \max_{i \le k} w_i$. Under Assumptions 1–5, at typical loading magnitudes and up to noise, $R \gtrsim \kappa^{-1}(\sigma_k/\sigma_{i^*})\sqrt{\mu}$ in the saturated regime and $R \gtrsim (\sigma_{i^*}/\sigma_1)\sqrt{\mu}$ in the transient one. When the bound exceeds one, Eq. 3 weights i^* above every $i \le k$, and a threshold ε places i^* in $\mathcal{S}_T = \{i \mid |z_i^* - 1| > \varepsilon\}$ but no $i \le k$, so $\mathcal{S}_T \not\subset \{1, \ldots, k\}$; this holds (i) if $\mu > \kappa^2(\sigma_{i^*}/\sigma_k)^2$ when saturated, and (ii) if $\mu > (\sigma_1/\sigma_{i^*})^2$ when transient.*

$R > 1$ is the *separation condition*: a transient probe favours the spectrum's top, while a saturated one, amplifying skill overlap by the inverse singular value, inverts this and, for a decaying spectrum and $\kappa = O(1)$, separates at chance-level misalignment.

Proposition 1 (Stability of the protection weights)**.** *Let $\theta_i = |z_i^* - 1|$, $M = \max_i \theta_i$, let the probe return $\hat{z}_i = z_i^* + \xi_i$ with ξ_i sub-Gaussian of parameter τ, $\hat{w}_i = |\hat{z}_i - 1| / \max_j |\hat{z}_j - 1|$, and $\tau_\eta = \tau\sqrt{2\log(2r/\eta)} < M$. Then with probability at least $1 - \eta$, $\max_i |\hat{w}_i - w_i| \le 2\tau_\eta/(M - \tau_\eta)$ and $|\mathcal{L}_{ortho}(\Delta\mathbf{W}; \hat{\mathbf{w}})^{1/2} - \mathcal{L}_{ortho}(\Delta\mathbf{W}; \mathbf{w})^{1/2}| \le \|\Delta\mathbf{W}\|_F \max_i |\hat{w}_i - w_i|$ for every $\Delta\mathbf{W}$.*

Unlike subset rules (Lemma 4), graded weights need no margin between relevant and irrelevant components, and the number of Phase 1 rollouts required to estimate them grows as $\log r$.

Theorem 2 (Graded protection bound)**.** *Complete $\{\mathbf{v}_i\}$ to an orthonormal basis of $\mathbb{R}^n$ with $w_i := 0$ for $i > r$, and let $g_i = \|\mathbf{G}_T^\star \mathbf{v}_i\|_2$ be the skill column loading. Under Assumption 6, for every update $\Delta\mathbf{W}$, with $\varrho = \|\Delta\mathbf{W}\|_F$, and every $\zeta > 0$,*

$$L_T(\mathbf{W}^\star + \Delta\mathbf{W}) - L_T(\mathbf{W}^\star) \le \Phi_\zeta(\mathbf{w})\big(\mathcal{L}_{ortho}(\Delta\mathbf{W}) + \zeta\varrho^2\big)^{1/2} + \varrho\,\epsilon^\star + \tfrac{\ell}{2}\varrho^2, \tag{4}$$

where $\Phi_\zeta(\mathbf{w})^2 = \sum_i g_i^2/(w_i^2 + \zeta)$. If $g_i = 0$ whenever $w_i = 0$, the first term tends to $\Phi_0(\mathbf{w})\mathcal{L}_{ortho}(\Delta\mathbf{W})^{1/2}$ as $\zeta \to 0$, with $\Phi_0(\mathbf{w})^2 = \sum_{i: w_i > 0} g_i^2/w_i^2$.

Table 1: The injection rules compared. All methods start from the same deployed SVF expert with weights $\mathbf{W}$, and each singles out spectral components of $\mathbf{W}$ either to place the update in them or to keep it out. They differ in how those components are identified and where the boundary falls: every baseline draws one at a fixed rank, whereas PALoRA draws none and weights each component.

Method	Components singled out	Role	Boundary	Enforced
LoRA (Hu et al., 2022)	none	—	—	—
PiSSA (Meng et al., 2024)	leading components	adapted	fixed rank	init.
MiLoRA (Wang et al., 2025)	minor components	adapted	fixed rank	init.
OPLoRA (Xiong & Xie, 2026)	leading components	protected	fixed rank	training
CorDA (Yang et al., 2024)	skill-activation covariance	protected	fixed rank	init.
PALoRA (ours)	SVF deviation $\lvert\mathbf{z}^* - \mathbf{1}\rvert$	protected	none, graded	training

Skill interference is thus controlled by the very quantity PALoRA minimizes. The constant $\Phi_0(\mathbf{w})$ is small when the weights are large wherever the column loading g_i is large, and since g_i scales with the overlap $\|\mathbf{P}_T\mathbf{v}_i\|_2$ as w_i does, g_i/w_i depends only on the regime of component i (Lemma 3): every component that carries skill loading is penalized, whatever its position in the spectrum.

Corollary 1 (Rank-fixed protection as the binary case)**.** *For $w_i = \mathbb{I}[i \in \mathcal{S}]$ and hard protection $\Delta\mathbf{W}\mathbf{v}_i = \mathbf{0}$, $i \in \mathcal{S}$, the worst admissible update of norm ϱ has $\Delta L_T = \varrho\|(g_i)_{i\notin\mathcal{S}}\|_2$ up to $\pm(\varrho\,\epsilon^\star + \ell\varrho^2/2)$. For $\mathcal{S} = \{1, \dots, k\}$, i^* of Theorem 1 has $\mathbb{E}[g_{i^*}^2] \asymp \|\mathbf{G}_T^\star\|_F^2/n > 0$, so rank-fixed protection such as OPLoRA cannot exclude a first-order degradation, however strictly enforced.*

From $\mathbf{B}_l = \mathbf{0}$, any run decreasing $\mathcal{L}_{\text{total}}$ has $\sum_l \mathcal{L}_{\text{ortho},l} \le (\mathcal{L}_{\text{SFT}}(0) - \mathcal{L}_{\text{SFT}})/\lambda_{\text{ortho}}$, so the graded term decays as $\lambda_{\text{ortho}}^{-1/2}$, whereas the remaining terms of Eq. 4 and the binary leak do not (Corollary 2). Appendix B.8 adds the regime dependence of $\mathbf{w}$ (Corollary 3) and multi-skill bounds (Corollary 4).

4 Experiments

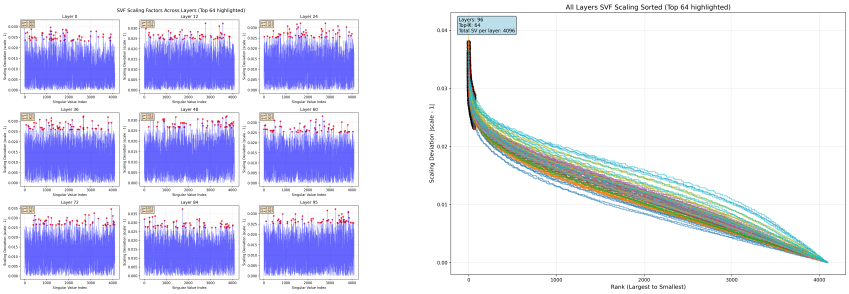
We test the rank premise of spectral PEFT through four questions: whether the skill-critical directions are the leading ones (Q1), whether retention across domains, architectures, and baselines requires protecting them (Q2), the contribution of each stage (Q3), and sequential injection (Q4).

4.1 Setup, Baselines, and Matched Budget

We use Llama-3.1-8B-Instruct for the main experiments and ablations and Mistral-7B-Instruct-v0.3 for cross-architecture validation, preserving GSM8K (final-answer accuracy), MBPP (pass@1), and AI2-ARC Challenge (accuracy), each with its own SVF expert. Knowledge comes from the multi-domain `Unknown` split (Gekhman et al., 2024) of Pletenev et al. (2025), whose facts are by construction not recalled before fine-tuning, so all measured recall is acquired during injection; we inject 100, 250, 500, and 1000 of them and report recall as the fraction answered correctly afterwards. The adapter defaults are $r = 16$ and $\alpha = 32$, and PALoRA has no protected-subspace size to set, since Eq. 3 grades the full spectrum; λ_{ortho} controls how much plasticity the injection is allowed while stability is maintained (Section 4.4), and the full configuration, hardware, and evaluation protocol are given in Appendix F. All results we report are means over three seeds.

Every method is trained on, and deployed with, the same Phase 1 SVF expert, so the skill information present before injection is identical across the comparison and each spectral baseline decomposes the very weights PALoRA protects. What remains different is the injection rule, and we select baselines spanning its axes (Table 1). PiSSA, MiLoRA, and OPLoRA identify their components by a rule on the spectrum, occupying the principal ones, occupying the minor ones, or protecting the top-k; CorDA identifies them from skill activations as PALoRA does, and OPLoRA enforces its choice during training as PALoRA does. Where all four agree, and where PALoRA departs, is the boundary: each separates the spectrum at a fixed rank and treats the protected side alike, so together they test whether such a cut-off, or its side, decides retention. All methods share an identical training loop, objective, optimization schedule (except PiSSA), target modules, and adapter shape (Appendix G).

Since the expert is shared, its Phase 1 cost is common to the comparison and paid once per skill, amortized over every later injection that reuses it. PALoRA's own overhead is the constraint: its



(a) $|z_i^* - 1|$ across the spectrum. (b) Top-64 magnitudes, sorted by size.

Figure 2: Visual evidence for distributed skill information. The SVF scaling factors show a sharp elbow and a long tail, while the top SVF-selected components are scattered across singular-value rank rather than aligned with the largest singular values.

Table 2: Validation of SVF-guided subspace selection on Llama-3.1-8B with the GSM8K expert and 1000 unknown facts. Every row deploys the SVF expert.

	Performance		Overlap with SVF-Guided	
Method	**Recall (%)**	**GSM8K (%)**	**Micro Overlap**	**Micro Jaccard**
SVF-Guided	**62.6**	81.12	—	—
Top-k Raw	61.6	81.12	0.0229	0.0116
Random-k	62.2	79.61	0.0145	0.0073
PALoRA (ours)	61.1	**83.04**	—	—

injection stage runs within 1.3–$1.4\times$ the baselines, the deployed model carries less than 0.006% additional parameters, and the protected singular vectors are needed only in Phase 2 (Appendix I.1).

4.2 Q1: Are the Skill-Critical Directions the Leading Singular Directions?

We first test the premise shared by the spectral baselines, and with it the conclusion of Theorem 1. If skill-relevant information were concentrated in the largest singular directions, raw top-k protection would already suffice. Figure 2 visualizes the SVF scaling pattern for the GSM8K expert on Llama-3.1-8B, with every expert and model family in Appendix E, and Table 2 compares SVF-guided selection with raw top-k and random-k protection under an otherwise identical injection run. The selected components spread across the spectrum and barely overlap the raw top-k subspace, at a micro Jaccard index of 0.012, yet the two sets preserve the skill equally well, 81.12 on GSM8K, while random-k reaches 79.61 at comparable recall: two nearly disjoint sets of 64 directions carry the same share of the skill, as a distributed skill predicts, and top-k protection succeeds without the skill being concentrated where it protects. If no set of 64 directions is privileged, the protection should follow the expert's importance profile over the whole spectrum rather than any cut of it, which is what PALoRA does. Since that importance falls off sharply (Figure 2b), the protection still concentrates where the skill relies on it most, raising GSM8K accuracy to 83.04 at comparable recall.

4.3 Q2: Does Retention Require Protecting the Leading Directions?

Table 3 reports both model families under an identical protocol. On GSM8K, MBPP, and AI2-ARC the base models reach 78.54, 70.7, 90.32 (Llama-3.1-8B-Instruct) and 44.20, 49.5, 81.65 (Mistral-7B-Instruct-v0.3), with the Phase 1 experts at 83.85, 72.7, 90.11 and 48.29, 51.5, 85.06.

Along the rank axis of Table 3, keeping the update off the leading components does not protect the skill: PiSSA, adapting them, trails PALoRA by 6.7 GSM8K points, yet MiLoRA, adapting the minor ones, stays within 0.1 points of LoRA in backward transfer. CorDA uses skill activations only to initialize, and falls below the base model everywhere. OPLoRA protects the $k = 128$ leading directions during training, leaving the skill loading outside them exposed however strictly enforced (Corollary 1). PALoRA does not single out these directions, nearly disjoint from those it weights most

Table 3: Knowledge recall (Rec.) and skill accuracy (Acc.) in percent, with #F injected facts and the same Phase 1 SVF expert for every method. Bold marks the best skill accuracy in each row. ACC and BWT average all 24 configurations: ACC is the mean of skill and knowledge accuracy, and BWT the mean change in skill accuracy against the base model. Standard deviations are given in Appendix H.

		LoRA		PiSSA		MiLoRA		OPLoRA		CorDA		PALoRA	
Skill	**#F**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**
Llama-3.1-8B-Instruct													
GSM8K	100	68.0	78.24	56.0	69.17	66.3	80.11	66.7	80.19	72.0	76.40	60.0	**82.31**
	250	66.0	80.21	52.3	78.17	65.1	81.45	64.3	82.56	67.5	75.06	58.8	**82.87**
	500	66.4	81.43	50.9	77.43	63.9	80.34	63.1	82.49	67.5	66.16	61.2	**82.76**
	1000	62.0	78.80	49.4	78.42	63.0	77.66	62.1	81.65	61.8	58.33	61.1	**83.04**
MBPP	100	62.3	66.67	51.3	**68.35**	63.7	63.97	64.7	64.98	74.0	43.10	66.0	65.99
	250	63.5	65.32	53.9	65.99	64.4	65.99	65.9	**66.67**	60.0	39.06	64.7	**66.67**
	500	64.7	62.96	51.7	**68.01**	61.5	62.96	63.7	66.33	53.1	23.23	62.7	67.34
	1000	62.5	64.98	51.7	64.31	62.6	60.94	60.8	65.66	50.1	9.09	58.1	**65.99**
AI2-ARC	100	68.3	88.52	50.7	88.69	67.7	88.99	69.3	87.95	69.3	82.72	63.7	**89.27**
	250	66.9	88.43	51.6	88.33	65.5	88.72	66.8	**89.13**	66.3	78.54	65.9	88.38
	500	64.8	87.58	49.1	88.29	64.2	87.60	63.7	**89.06**	49.9	53.79	63.9	88.48
	1000	62.8	86.98	49.7	87.71	60.3	87.12	64.4	**88.90**	46.1	40.31	60.1	88.31
Mistral-7B-Instruct-v0.3													
GSM8K	100	92.0	42.48	91.0	39.50	93.3	43.95	91.3	43.19	93.7	26.00	88.7	**45.62**
	250	87.9	42.15	85.7	38.74	89.1	42.86	87.5	**44.02**	83.3	24.03	83.2	43.77
	500	88.3	38.59	84.1	39.50	80.5	43.09	78.0	**44.93**	80.1	12.23	78.0	44.25
	1000	81.4	40.36	79.3	33.89	80.4	37.76	76.3	43.77	83.0	13.37	71.2	**44.02**
MBPP	100	91.7	**51.18**	89.7	45.45	91.7	48.82	89.0	49.16	89.0	34.68	91.7	49.83
	250	86.3	49.16	85.2	46.80	89.5	48.15	85.2	49.16	86.7	17.85	82.4	**51.85**
	500	82.4	50.51	82.9	45.12	81.4	51.85	75.6	**52.53**	89.4	7.74	79.6	49.16
	1000	82.9	49.49	79.8	44.78	83.2	**51.52**	73.0	**51.52**	84.8	1.01	80.0	50.84
AI2-ARC	100	90.0	83.63	87.0	84.02	90.3	**84.32**	87.7	83.77	84.3	75.88	88.3	84.22
	250	84.3	83.77	81.9	83.49	82.3	83.81	81.3	**84.05**	80.0	57.80	83.5	83.92
	500	83.3	83.26	79.0	84.06	78.1	84.16	75.3	84.16	84.3	46.30	76.9	**84.89**
	1000	79.5	83.22	77.0	83.33	75.2	84.06	71.8	**84.51**	87.3	26.21	73.9	83.96
ACC ↑		**71.6**		66.9		71.1		71.0		57.3		70.7	
BWT ↑		-1.3		-2.8		-1.2		0.0		-27.9		$+\mathbf{0.3}$	

Table 4: Ablations on Llama-3.1-8B, GSM8K expert, 1000 unknown facts. **(a)** Contribution of each stage. **(b)** Sensitivity to the orthogonality weight.

(a)	**Configuration**	λ_{ortho}	**Recall**	**GSM8K**	**vs. base**
(R0)	Base model, no injection	—	—	78.54	—
(R1)	SVF expert, no injection	—	—	83.85	$+5.31$
(A)	LoRA	0	61.7	72.33	-6.21
(B)	SVF + unconstrained LoRA	0	62.0	78.80	$+0.26$
(C)	Constraint only	0.0003	59.3	75.13	-3.41
(D)	PALoRA	0.0003	61.1	**83.04**	$+\mathbf{4.50}$

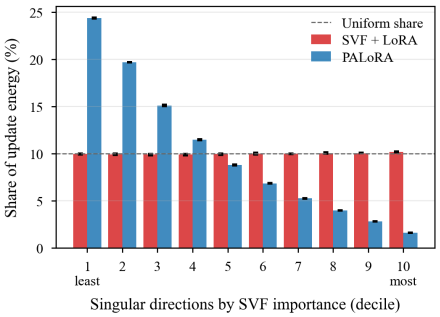
(b) λ_{ortho}	**Recall**	**GSM8K**
0	62.0	78.8
0.0003	61.1	83.0
0.001	59.1	82.1
0.003	55.7	82.0

(Table 2), yet matches OPLoRA: 70.7 against 71.0 ACC and 71.8% against 72.8% recall, averaged over all 24 configurations. It exceeds OPLoRA in skill accuracy in 14 of them and has the only positive backward transfer (Lopez-Paz & Ranzato, 2017), $+0.3$ against 0.0, so retention does not require the leading directions, as Q1 predicts.

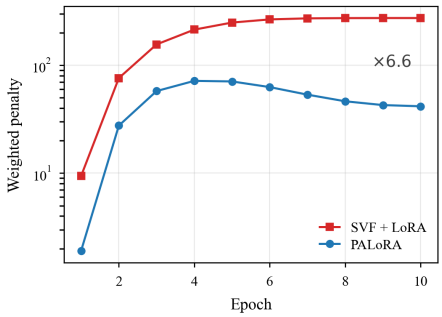
4.4 Q3: Which Stage Produces the Retention?

Since the SVF scaling is retained through Eq. 1, the retention could come from the scaling alone. A 2×2 design over {SVF deployed} $\times$ {constraint applied} separates the two (Table 4(a)).

Row (B) is decisive: injecting into the SVF expert without the constraint reaches 78.80, 5.05 points below the expert it started from, so the retention cannot be attributed to the SVF scaling alone. Relative to row (A), the SVF stage is worth $+6.47$ GSM8K points and the constraint a further $+4.24$, so that (D) ends within 0.81 points of the expert. The two stages act as a single mechanism: the constraint protects exactly those directions the expert identifies and occupies. Figure 3 shows where



(a) Update energy per SVF importance decile.



(b) Weighted penalty during training.

Figure 3: Where the update goes, for Llama-3.1-8B with the GSM8K expert and 1000 unknown facts, over three seeds; the standard deviation is smaller than the markers. **(a)** Share of the final update energy $\|\Delta \mathbf{W} \mathbf{v}_i\|_2^2$ per decile of the SVF importance $|z_i^* - 1|$, formed per layer; decile 10 holds the directions the expert rescaled most. **(b)** Weighted penalty of Eq. 3, monitored without optimization for unconstrained LoRA.

Table 5: Sequential injection on Llama-3.1-8B, GSM8K expert, 4 rounds × 250 unknown facts. Recall is averaged over the rounds seen so far, and standard deviations are given in Appendix C.

	GSM8K				Avg. recall			
Round	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
(A) LoRA	74.7	75.6	71.1	70.5	62.9	58.9	55.5	51.1
(B) SVF + LoRA	**82.3**	78.0	76.1	74.7	66.4	60.3	53.0	45.8
(O) SVF + OPLoRA	79.9	76.6	78.0	76.5	67.1	61.8	56.2	51.9
(D) PALoRA	81.2	**81.1**	**81.5**	**79.7**	63.6	57.4	53.3	47.4

the update goes. Unconstrained LoRA spreads its energy evenly over the SVF importance deciles, about 10% in each, whereas PALoRA places only 1.6% of its update in the top decile and 24.4% in the bottom one, so the protection is graded rather than a cut (Figure 3a). During training, the weighted penalty of the unconstrained update, which bounds its skill degradation (Theorem 2), grows and saturates, whereas PALoRA pulls it back after the first epochs and ends $6.6\times$ lower (Figure 3b); the SFT loss on the facts ends at 0.357 against 0.356, so this redistribution does not impede acquisition.

Table 4(b) shows the trade-off this mechanism repositions: GSM8K rises from 78.8 at $\lambda_{\text{ortho}} = 0$ to 83.0 at 0.0003 at nearly unchanged recall, and a stronger weight mostly costs recall (Corollary 2).

4.5 Q4: Does the Advantage Persist Under Sequential Injection?

We finally examine the continual-adaptation setting of our motivation, splitting the 1000 facts into four rounds of 250 in one adapter (Table 5). Measured from where each starts, unconstrained LoRA loses a similar amount without and with the expert (8.0 and 9.2 points), so the expert only raises where degradation begins. SVF + OPLoRA slows the decay but still ends 2.1 points below the base model. PALoRA loses 1.5 points, stays above the base model at every round, and ends 5.0 points above SVF + LoRA and 3.2 above SVF + OPLoRA, with average recall 1.6 above the former and 4.5 below the latter; as the updates accumulate, graded protection retains more than rank-fixed protection.

5 Conclusion

We revisited the premise of spectral PEFT that singular-value rank identifies the components to protect, showing theoretically and empirically that reasoning-relevant information in MLP weights is distributed across the singular spectrum: protecting the leading directions is not necessary to retain a skill and cannot exclude a first-order degradation. PALoRA grades the penalty on every singular direction by the skill relevance a frozen SVF expert measures, bounding first-order skill degradation. Across models, domains, and sequential injection, it matches the strongest rank-based protection and retains more as updates accumulate, so skill preservation is better framed as importance estimation than subspace selection. Its main limitation is that injection capacity depends on protection strength, so the trade-off is repositioned rather than removed. Future work should extend PALoRA to several skills (Appendix D), larger models and architectures, and longer continual-adaptation horizons.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and editing of the manuscript (grammar, phrasing, and copy-editing of text drafted by the authors) and as a coding assistant for parts of the experimental and plotting code. We did not use generative AI tools for research ideation, for deriving the theoretical results, or for producing or interpreting any experimental result; retrieval and discovery of related work were carried out by the authors. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was read, tested, and verified against the protocols described in Appendix F. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work studies parameter-efficient adaptation of publicly released large language models and does not involve human subjects, personally identifying information, or the collection of new data. All models (Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3) and datasets (GSM8K, MBPP, AI2-ARC, and the factual corpus of Pletenev et al. (2025)) are publicly available and were used in accordance with their licenses and intended research use. PALoRA is a mechanism for injecting new facts into a model while preserving reasoning ability; as with any knowledge-editing method, it could in principle be used to insert incorrect or misleading facts, and our evaluation covers only non-conflicting factual updates. We therefore recommend that deployments pair knowledge injection with verification of the injected facts. Our experiments inherit any biases present in the pre-trained base models and in the public datasets used, and we make no claim that PALoRA mitigates such biases.

REPRODUCIBILITY STATEMENT

We have aimed to make every result in this paper reproducible. The full experimental setup, including base models, skill and knowledge datasets, dataset splits, default hyperparameters, and the hardware environment, is specified in Section 4 and detailed in Appendix F, which also documents the training protocol for both phases and the evaluation protocol used for skill and recall measurements. Appendix G describes how each baseline was adapted to the same data interface, optimization schedule, and adapter capacity, so that the comparisons in Section 4.3 are matched by construction. Appendix I and Appendix I.1 report the parameter, memory, and compute budget of every method, including measured runtimes. All assumptions underlying our theoretical claims are stated formally in Appendix B.1, complete proofs of the theorems, lemmas, proposition, and corollaries are given in Appendix B, and a consolidated notation reference is provided in Appendix A. Our implementation of PALoRA, together with the training and evaluation scripts needed to reproduce the reported numbers, is provided as anonymized supplementary material and will be released publicly upon acceptance.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.444. URL https://aclanthology.org/2024.emnlp-main.444/.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uPv9Y3gmAI5.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vfsRB5MImo9.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VrHiF2hsrm.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 41425–41446. Curran Associates, Inc., 2024. doi: 10.52202/079017-1311. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/48c368f105e8145b945227b73255635a-Paper-Conference.pdf.

David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower (ed.), *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6ZBHIEtdP4.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Martial Mermillod, Aurélia Bugaiska, and Patrick BONIN. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, Volume 4 - 2013, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00504. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2013.00504.

Sergey Pletenev, Maria Marina, Daniil Moskovskiy, Vasily Konovalov, Pavel Braslavski, Alexander Panchenko, and Mikhail Salnikov. How much knowledge can you pack into a LoRA adapter without harming LLM? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4309–4322, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.243. URL https://aclanthology.org/2025.findings-naacl.243/.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ozX92bu8VA.

Reece S Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=xp7B8rkh7L.

Gilbert W Stewart and Ji-guang Sun. Matrix perturbation theory. *(No Title)*, 1990.

Qi Sun, Edoardo Cetin, and Yujin Tang. Transformer-squared: Self-adaptive LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dh4t9qmcvK.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

Papers), pp. 4823–4836, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.248. URL https://aclanthology.org/2025.naacl-long.248/.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715/.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Yifeng Xiong and Xiaohui Xie. Oplora: Orthogonal projection lora prevents catastrophic forgetting during parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34088–34096, 2026.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Gi00NVru6n.

APPENDIX

This appendix is organized in three parts. The first part supports the theoretical claims of Section 3.3: Appendix A consolidates the notation used throughout the paper, and Appendix B states the formal assumptions, the full proofs of the main theorems, the detailed corollaries, and a discussion of how the results should be read geometrically and where they cease to hold. The second part reports additional empirical results: the sequential-injection experiment (Appendix C), a measurement of how the protected subspace grows when several skills must be preserved at once (Appendix D), and SVF adapter visualizations for every expert (Appendix E). The third part documents reproducibility: the datasets, training configuration, and evaluation protocol (Appendix F), the adaptation and fairness protocol applied to the baselines (Appendix G), and the parameter, memory, and compute analysis together with the measured data and compute budget of every method (Appendix I).

A GLOBAL NOTATION REFERENCE

This section consolidates the notation used throughout the paper. We first define the symbols that carry the theoretical analysis of Section 3.3 and Appendix B, and then tabulate every symbol, grouped by its role in the overall training pipeline (Table 6), the PALoRA parameterization (Table 7), and the theoretical analysis (Table 8).

A.1 DEFINITIONS FOR THE THEORETICAL ANALYSIS

Let $\mathbf{W} \in \mathbb{R}^{m\times n}$ be an MLP weight matrix of a Transformer layer. Its singular value decomposition is

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top,$$

where $r = \operatorname{rank}(\mathbf{W})$, $\mathbf{u}_i \in \mathbb{R}^m$ and $\mathbf{v}_i \in \mathbb{R}^n$ are the left and right singular vectors, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

Let $\mathbf{z} \in \mathbb{R}^r$ denote the SVF scaling vector. The scaled weight matrix is

$$\mathbf{W}_\mathbf{z} = \mathbf{U}\operatorname{diag}(\mathbf{z}\odot\boldsymbol{\sigma})\mathbf{V}^\top = \sum_{i=1}^{r} z_i \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Let $\mathbf{z}^*$ denote the optimal SVF scaling vector obtained after skill fine-tuning on task T. Let $L_T(\mathbf{W})$ denote the task-specific loss associated with skill T.

Let $\mathcal{U}_T \subset \mathbb{R}^n$ be the task-specific input subspace associated with skill T, with orthogonal projector

$$\mathbf{P}_T = \mathbf{V}_T\mathbf{V}_T^\top, \qquad \mathbf{V}_T \in \mathbb{R}^{n\times s}, \qquad s \ll n.$$

We define the set of skill-relevant singular components by

$$\mathcal{S}_T = \{i \in \{1,\dots,r\} \mid |z_i^* - 1| > \varepsilon\},$$

for a threshold $\varepsilon > 0$.

The deployed expert of Phase 2 is $\mathbf{W}^\star = \mathbf{W}_{\mathbf{z}^*}$. It has the same singular vectors as $\mathbf{W}$ (only the ordering of its singular values may change), so the indices i of $\mathbf{u}_i, \mathbf{v}_i$ refer to the same directions before and after Phase 1. We complete $\{\mathbf{v}_i\}_{i=1}^r$ to an orthonormal basis $\{\mathbf{v}_i\}_{i=1}^n$ of $\mathbb{R}^n$ and $\{\mathbf{u}_i\}_{i=1}^r$ to an orthonormal basis $\{\mathbf{u}_i\}_{i=1}^m$ of $\mathbb{R}^m$; the added vectors span the null spaces of $\mathbf{W}^\top$ and $\mathbf{W}$, are unaffected by SVF, and receive protection weight $w_i := 0$ for $i > r$. The protection weights and the graded penalty of a generic layer are

$$w_i = \frac{|z_i^* - 1|}{\max_j |z_j^* - 1|} \in [0,1], \qquad \mathcal{L}_{\text{ortho}}(\Delta\mathbf{W};\mathbf{w}) = \sum_{i=1}^{n} w_i^2 \|\Delta\mathbf{W}\mathbf{v}_i\|_2^2 = \|\Delta\mathbf{W}\mathbf{V}\operatorname{diag}(\mathbf{w})\|_F^2,$$

and the skill column loading at the deployed expert is $g_i = \|\mathbf{G}_T^\star \mathbf{v}_i\|_2$, where $\mathbf{G}_T^\star$ is the structured part of $\nabla_\mathbf{W} L_T(\mathbf{W}^\star)$ (Assumption 6).

We use the Frobenius inner product

$$\langle \mathbf{A}, \mathbf{B}\rangle_F = \operatorname{tr}(\mathbf{A}^\top\mathbf{B}),$$

and the Frobenius norm $\|\mathbf{A}\|_F$.

A.2 SYMBOL TABLES

A few symbols recur across the tables, since the method and the analysis refer to the same object with layer-indexed and generic notation respectively; each table lists the form used in its own context.

Table 6: General notation, data, and optimization variables used throughout the paper.

Symbol	Description
$\mathcal{M}(\theta_{\text{base}})$	Pre-trained base language model with frozen parameters θ_{base}.
θ_{base}	Parameters of the frozen base model.
θ_{LoRA}	Trainable LoRA parameters, collected as $\{\mathbf{B}_l, \mathbf{A}_l\}_{l=1}^{L}$.
$\mathcal{D}_{\text{skill}}$	Skill dataset used in Phase 1 to train the SVF expert.
$\mathcal{D}_{\text{know}}$	Knowledge dataset used in Phase 2 for factual knowledge injection.
$\mathcal{D}_{\text{prior}}$	Pre-training data distribution used in the formal assumptions.
L	Number of adapted MLP layers.
l	Layer index, ranging over $1, \ldots, L$.
T	Number of Phase 2 training steps in Algorithm 1; in the theory section, T also denotes a task or skill, with the meaning determined by context.
T_m	The m-th skill in the multi-skill analysis; here m is a skill index and is unrelated to the matrix dimension m of Table 8.
k	Number of singular directions in a top-k protected set, used by the analysis, by the top-k and random-k comparisons, and by the rank-fixed baselines; PALoRA itself sets no k, since its protection is graded over the full spectrum.
$w_{l,i}$	Protection weight of the i-th singular direction of layer l, $\lvert z^*_{l,i} - 1\rvert / \max_j \lvert z^*_{l,j} - 1\rvert \in [0,1]$; top-$k$ selection is the special case $w_{l,i} \in \{0,1\}$.
r	LoRA rank in the adaptation module; when used in the theory section as $r = \text{rank}(\mathbf{W})$, it denotes the rank of the weight matrix. The layer-indexed r_l of Table 7 is a distinct quantity.
α	LoRA scaling hyperparameter in the experiments; in Assumption 1, $\alpha > 0$ also denotes the spectral decay exponent of the covariance eigenvalues.
λ_{ortho}	Weight of the orthogonality regularization term in Phase 2.
$\mathcal{L}_{\text{SFT}}$	Supervised fine-tuning loss used for factual knowledge injection; $\mathcal{L}_{\text{SFT}}(0)$ is its value at the LoRA initialization $\mathbf{B}_l = \mathbf{0}$.
$\mathcal{L}_{\text{ortho},l}$	Orthogonality penalty applied at layer l.
$\mathcal{L}_{\text{total}}$	Total Phase 2 objective, $\mathcal{L}_{\text{SFT}} + \lambda_{\text{ortho}} \sum_{l=1}^{L} \mathcal{L}_{\text{ortho},l}$.

Table 7: Notation for the PALoRA parameterization, SVD quantities, and protected subspaces.

Symbol	Description
$\mathbf{W}_{\text{mlp}}^{(l)}$	Pre-trained MLP weight matrix at layer l, with shape $\mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$.
$\mathbf{W}_{\text{mlp}}^{\prime(l)}$	Adapted MLP weight matrix after combining the SVF and LoRA components.
$\text{SVF}(\mathbf{W})$	Singular Value Fine-Tuning transformation applied to a pre-trained weight matrix $\mathbf{W}$.
$\text{LoRA}(\mathbf{W})$	Low-rank adaptation term applied to a pre-trained weight matrix $\mathbf{W}$.
$\Delta\mathbf{W}_l$	LoRA update at layer l, defined as $\mathbf{B}_l\mathbf{A}_l$.
$\mathbf{B}_l$	Left LoRA factor for layer l, with shape $\mathbb{R}^{d_{\text{out}} \times r}$.
$\mathbf{A}_l$	Right LoRA factor for layer l, with shape $\mathbb{R}^{r \times d_{\text{in}}}$.
$d_{\text{out}}, d_{\text{in}}$	Output and input dimensions of an adapted MLP weight matrix; written $d_{\text{out},l}, d_{\text{in},l}$ when the layer must be made explicit.
r_l	Number of singular values of the MLP weight matrix at layer l, $r_l = \min(d_{\text{out},l}, d_{\text{in},l})$; this is the dimension of the SVF scaling vector $\mathbf{z}_l$, and is distinct from the LoRA rank r.
$\mathbf{W}_{\text{mlp}}^{(l)} = \mathbf{U}_l \operatorname{diag}(\boldsymbol{\sigma}_l)\mathbf{V}_l^\top$	Singular value decomposition of the pre-trained MLP weight matrix at layer l.
$\mathbf{U}_l, \mathbf{V}_l$	Left and right singular-vector matrices for layer l.
$\boldsymbol{\sigma}_l$	Vector of singular values of $\mathbf{W}_{\text{mlp}}^{(l)}$.
$\sigma_{l,i}$	i-th singular value of the pre-trained MLP weight matrix at layer l.
$\mathbf{u}_{l,i}$	i-th left singular vector of layer l, spanning the protected output subspace.
$\mathbf{v}_{l,i}$	i-th right singular vector of layer l, spanning the protected input subspace.
$\mathbf{z}_l$	Trainable SVF scaling vector for layer l.
$\mathbf{z}_l^*$	Frozen SVF scaling vector obtained after Phase 1 training.
$\mathbf{1}$	All-ones vector used as the identity reference for SVF scaling.
$\odot$	Elementwise (Hadamard) product, used in the SVF scaling $\mathbf{z} \odot \boldsymbol{\sigma}$.
$\mathcal{I}_{\text{crit},l}$	Set of top-k singular-component indices in layer l selected by the magnitude of $\lvert z_{l,i}^* - 1\rvert$; written $\mathcal{I}_{\text{crit}}$ when the layer is clear from context; it is the binary special case of $\mathbf{w}_l$.
TopK_i	Operator returning the k indices i with the largest argument values, used to define $\mathcal{I}_{\text{crit},l}$.
$P_{\text{SVF}}, P_{\text{LoRA}}$	Trainable parameter counts of the Phase 1 SVF stage and the Phase 2 LoRA stage (Appendix I).

Table 8: Notation used in the theoretical analysis and proofs.

Symbol	Description
$\mathbf{W} \in \mathbb{R}^{m \times n}$	Generic pre-trained weight matrix used in the theoretical analysis.
$\mathbf{W}_0$	Initial pre-trained weight matrix before task adaptation.
$\mathbf{W}_\mathbf{z}$	SVF-parameterized weight matrix obtained by scaling the singular values of $\mathbf{W}$.
$\mathbf{W}^\star$	Deployed SVF expert $\mathbf{W}_{\mathbf{z}^*}$, the point from which Phase 2 starts.
m, n	Output and input dimensions of the generic theoretical weight matrix $\mathbf{W}$.
$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}$	SVD factors of $\mathbf{W}$, where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$.
σ_i	i-th singular value of $\mathbf{W}$.
$\mathbf{u}_i, \mathbf{v}_i$	i-th left and right singular vectors of $\mathbf{W}$; for $i > r$, the orthonormal completion to a basis of $\mathbb{R}^m$ and $\mathbb{R}^n$.
i^*	Off-principal index, $i^* > k$, with chance-level skill overlap, used in Theorem 1.
$\mathbf{x}$	Input vector to the layer, drawn from $\mathcal{D}_{\text{prior}}$ or from the task distribution.
$L_T(\mathbf{W})$	Task-specific loss for skill T, minimized during skill adaptation.
$\mathcal{U}_T$	Low-dimensional task-specific input subspace associated with skill T.
$\mathbf{P}_T$	Orthogonal projector onto $\mathcal{U}_T$, written as $\mathbf{V}_T \mathbf{V}_T^\top$.
$\mathbf{V}_T$	Basis matrix whose columns span the task-specific subspace $\mathcal{U}_T$.
f_T	Task-specific compositional function associated with skill T, $f_T : \mathbb{R}^n \to \mathbb{R}^d$.
d	Output dimension of the skill function f_T.
$\mathbf{C}_{\text{prior}}$	Input covariance matrix induced by the pre-training distribution.
λ_i	i-th eigenvalue of the covariance matrix $\mathbf{C}_{\text{prior}}$.
$\mathbf{G}_T$	Low-rank structured component of the task gradient at $\mathbf{W}_0$.
$\mathbf{N}, \epsilon_{\text{noise}}$	Zero-mean noise component in the gradient decomposition, and the upper bound on its Frobenius norm.
$\mathbf{G}_T^\star, \mathbf{N}^\star, \epsilon^\star$	Structured part, noise part, and noise bound of the task gradient at the deployed expert $\mathbf{W}^\star$ (Assumption 6).
$\beta_j, \mathbf{g}_j, \mathbf{h}_j$	Singular values and orthonormal singular vectors in $\mathbf{G}_T = \sum_{j=1}^{s} \beta_j \mathbf{g}_j \mathbf{h}_j^\top$, with $\mathbf{h}_j \in \mathcal{U}_T$.
s	Rank bound on the structured task-gradient term $\mathbf{G}_T$, and dimension of the task subspace $\mathcal{U}_T$.
γ_i	Spectral loading coefficient $\mathbf{u}_i^\top \mathbf{G}_T \mathbf{v}_i$, measuring alignment of the task gradient with singular component i.
g_i	Column loading $\|\mathbf{G}_T^\star \mathbf{v}_i\|_2$ of the skill gradient at the deployed expert along input direction $\mathbf{v}_i$ (Lemma 3).
ν_i	Noise loading $\mathbf{u}_i^\top \mathbf{N} \mathbf{v}_i$, with $\|\nu_i\| \le \epsilon_{\text{noise}}$.
ρ, ρ_0	Diagonal energy fraction $\sum_i \gamma_i^2 / \|\mathbf{G}_T\|_F^2$ and its lower-bound constant (Assumption 4).
δ	Bound on the overlap $\|\mathbf{P}_T \mathbf{v}_i\|_2^2$ of the top-k right singular directions with the task subspace.
μ	Misalignment ratio $(s/n)/\delta$; $\mu = 1$ is chance level.
c_i; $c_{\min}, c_{\max}$; κ	Gauss–Newton curvature in the i-th SVF coordinate; its bounds; the curvature spread $c_{\max}/c_{\min}$ (Assumption 5).
t	Gradient-flow time of Phase 1 in Lemma 2.
R	Detection ratio of Theorem 1 (equal to $w_{i^*} / \max_{i \le k} w_i$); $R > 1$ is the separation condition.
ε	Threshold used to define the set of skill-relevant singular components.
θ_i, M	Noiseless probe score $\|z_i^* - 1\|$ and its maximum $\max_i \theta_i$ (Proposition 1).
$\hat{s}_i, \xi_i, \tau, \eta$	Observed probe score $\|z_i^* + \xi_i - 1\|$, its estimation noise, the sub-Gaussian parameter of that noise, and the failure probability of Proposition 1.
$\hat{w}_i, \tau_\eta$	Protection weight computed from the observed scores, and the uniform noise level $\tau\sqrt{2\log(2r/\eta)}$.
Δ_{gap}	Margin between the smallest skill-relevant and the largest irrelevant probe score.
B, V	Rollouts per REINFORCE update and variance proxy of the policy-gradient estimator (Assumption 3).
ζ	Free regularization level of the graded bound, $\zeta > 0$ (Theorem 2).

Continued on next page

Table 8 continued

Symbol	Description
$\Phi_\zeta(\mathbf{w}), \Phi_0(\mathbf{w})$	Skill-loading constant $\left(\sum_i g_i^2/(w_i^2+\zeta)\right)^{1/2}$ of Theorem 2 and its limit under the support condition.
$\mathcal{S}, \mathcal{A}_\mathcal{S}$	Binary protected index set of a rank-fixed rule and the set of updates it admits, $\{\Delta\mathbf{W} \mid \Delta\mathbf{W}\mathbf{v}_i = \mathbf{0}\ \forall i \in \mathcal{S}\}$ (Corollary 1); PALoRA passes the graded weights $\mathbf{w}$ instead, of which $\mathcal{S}$ is the binary case.
$\mathbf{\Pi}_\mathcal{S}, \mathbf{\Pi}_\mathcal{S}^U$	Projectors $\sum_{i\in\mathcal{S}} \mathbf{v}_i\mathbf{v}_i^\top$ and $\sum_{i\in\mathcal{S}} \mathbf{u}_i\mathbf{u}_i^\top$ onto the protected input and output directions.
ϱ	Frobenius norm of the knowledge-injection update $\Delta\mathbf{W}$ in Theorem 2 and Corollary 1.
ℓ	Smoothness constant of L_T in Assumption 6.
$\mathbf{z}, \mathbf{z}^*$	Generic (non-layer-indexed) SVF scaling vector and its value after skill fine-tuning, with entries z_i and z_i^*.
$z_i(t)$	SVF scaling trajectory under gradient flow from $z_i(0) = 1$ (Lemma 2).
$\mathcal{S}_T$	Set of skill-relevant singular indices identified by the SVF scaling deviation $\lvert z_i^* - 1\rvert$.
M, p	Number of skills in the multi-skill corollary, and the mean pairwise overlap fraction of their skill-critical index sets.
$\mathbf{w}^{(m)}, \mathbf{w}^\cup$	Protection weights of skill T_m and their elementwise maximum (Corollary 4).
$\langle\mathbf{A}, \mathbf{B}\rangle_F$	Frobenius inner product between matrices $\mathbf{A}$ and $\mathbf{B}$.
$\lVert\mathbf{A}\rVert_F$	Frobenius norm of matrix $\mathbf{A}$.

B Full Proofs of the Theoretical Results

B.1 Formal Assumptions

Assumption 1 (Pre-training data diversity)**.** *The pre-training corpus is drawn from a distribution $\mathcal{D}_{prior}$ with broad spectral support. The input covariance matrix*

$$\mathbf{C}_{prior} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{prior}}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{n\times n}$$

has full rank, and its eigenvalues satisfy

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0, \qquad \lambda_i \asymp i^{-\alpha}, \qquad \alpha > 0.$$

Consequently, the pre-trained weight matrix $\mathbf{W}_0$ concentrates its principal singular vectors along directions of maximum input variance under $\mathcal{D}_{prior}$.

Assumption 2 (Skill compositionality and misalignment)**.** *A skill T corresponds to a compositional function $f_T : \mathbb{R}^n \to \mathbb{R}^d$ that depends only on a low-dimensional, task-specific subspace $\mathcal{U}_T \subset \mathbb{R}^n$. Formally, there exists a projector $\mathbf{P}_T = \mathbf{V}_T\mathbf{V}_T^\top$, $\mathbf{V}_T \in \mathbb{R}^{n\times s}$, such that*

$$f_T(\mathbf{x}) = f_T(\mathbf{P}_T\mathbf{x}) \quad \textit{almost surely under the task distribution.}$$

The overlap of the top-k right singular directions of $\mathbf{W}_0$ with the task subspace is bounded, and is measured relative to chance: for every $i \leq k$,

$$\|\mathbf{P}_T\mathbf{v}_i\|_2^2 \leq \delta, \qquad \mu := \frac{s/n}{\delta}, \tag{5}$$

while off-principal directions have chance-level overlap, $\|\mathbf{P}_T\mathbf{v}_i\|_2^2 \asymp s/n$ for the indices $i > k$ considered below. A uniformly random s-dimensional subspace gives $\mathbb{E}\|\mathbf{P}_T\mathbf{v}_i\|_2^2 = s/n$, so $\mu = 1$ is chance-level misalignment and $\mu > 1$ indicates that the skill subspace avoids the leading singular directions more than a random subspace would. The ratio μ makes the size of δ quantitative, and Theorem 1 states for which values of μ its conclusion holds.

Assumption 3 (Gradient structure and estimator noise)**.** *Skill adaptation is driven by a stochastic estimator of $\nabla_{\mathbf{W}} L_T$, starting from the pre-trained weight matrix $\mathbf{W}_0$. In Phase 1 this estimator is the REINFORCE policy gradient over B rollouts with bounded reward, so that it is unbiased and its variance in each SVF coordinate is at most V/B for a variance proxy V. The gradient at $\mathbf{W}_0$ decomposes as*

$$\nabla_{\mathbf{W}} L_T(\mathbf{W}_0) = \mathbf{G}_T + \mathbf{N}, \tag{6}$$

where $\mathbf{G}_T$ is a low-rank structured term with $\mathrm{rank}(\mathbf{G}_T) \leq s$ and row space in $\mathcal{U}_T$. We write its singular value decomposition as

$$\mathbf{G}_T = \sum_{j=1}^{s} \beta_j\, \mathbf{g}_j\mathbf{h}_j^\top, \qquad \mathbf{h}_j \in \mathcal{U}_T\ \forall j, \qquad \textstyle\sum_j \beta_j^2 = \|\mathbf{G}_T\|_F^2,$$

with orthonormal $\{\mathbf{g}_j\} \subset \mathbb{R}^m$ and orthonormal $\{\mathbf{h}_j\} \subset \mathcal{U}_T$. The residual term $\mathbf{N}$ is a zero-mean noise matrix with small Frobenius norm

$$\|\mathbf{N}\|_F \leq \epsilon_{noise}.$$

Assumption 4 (Non-degenerate diagonal loading)**.** *With $\gamma_i = \mathbf{u}_i^\top \mathbf{G}_T \mathbf{v}_i$, the diagonal of $\mathbf{G}_T$ in the singular basis of $\mathbf{W}_0$ carries at least a constant fraction of the energy that an isotropic gradient would place on it:*

$$\rho := \frac{\sum_{i=1}^{r} \gamma_i^2}{\|\mathbf{G}_T\|_F^2} \;\geq\; \frac{\rho_0}{m}, \qquad \rho_0 > 0. \tag{7}$$

Under Assumption 5, Lemma 1 gives $\mathbb{E}[\rho] = \mathrm{tr}(\mathbf{P}_T\mathbf{V}\mathbf{V}^\top)/(sm)$, which equals $1/m$ when $r = n$, so the assumption holds in expectation. Its role is to exclude gradients whose energy lies off the diagonal. It cannot be dispensed with: for $\mathbf{G}_T = \mathbf{u}_2\mathbf{v}_1^\top$ one has $\gamma_i = 0$ for every i, so every singular value is stationary to first order and no singular-value probe can detect the skill, although $\|\mathbf{G}_T\|_F = 1$.

Assumption 5 (Isotropic alignment and curvature regularity)**.** *The frames $\{\mathbf{g}_j\}$ and $\{\mathbf{h}_j\}$ of Assumption 3 are distributed uniformly (Haar) among orthonormal s-frames of $\mathbb{R}^m$ and of $\mathcal{U}_T$ respectively, independently of each other and of $(\mathbf{U}, \mathbf{V})$. The Gauss–Newton curvatures c_i of Lemma 2 satisfy $0 < c_{\min} \leq c_i \leq c_{\max}$, and we write $\kappa = c_{\max}/c_{\min}$.*

Assumption 6 (Skill gradient at the deployed expert)**.** *L_T is ℓ-smooth in $\mathbf{W}$. At the deployed expert $\mathbf{W}^\star = \mathbf{W}_{\mathbf{z}^*}$ the skill gradient decomposes as*

$$\nabla_{\mathbf{W}} L_T(\mathbf{W}^\star) = \mathbf{G}_T^\star + \mathbf{N}^\star, \qquad \mathrm{row}(\mathbf{G}_T^\star) \subset \mathcal{U}_T, \quad \mathrm{rank}(\mathbf{G}_T^\star) \le s, \quad \|\mathbf{N}^\star\|_F \le \epsilon^\star, \tag{8}$$

and the singular frames of $\mathbf{G}_T^\star = \sum_j \beta_j^\star \mathbf{g}_j^\star \mathbf{h}_j^{\star\top}$ satisfy the isotropy of Assumption 5.

Assumption 6 is needed because Phase 2 does not start from $\mathbf{W}_0$: the injected update is added to $\mathbf{W}^\star$, so its first-order effect on the skill is governed by the gradient at $\mathbf{W}^\star$, not at $\mathbf{W}_0$. It keeps the same skill subspace $\mathcal{U}_T$, since SVF changes only singular values and the skill still reads its inputs through $\mathbf{P}_T$. It does not require the diagonal loadings at $\mathbf{W}^\star$ to be non-zero: at a converged expert the SVF gradient $\partial L_T / \partial z_i = \sigma_i \mathbf{u}_i^\top \nabla_{\mathbf{W}} L_T(\mathbf{W}^\star)\mathbf{v}_i$ vanishes, so the diagonal of $\mathbf{G}_T^\star + \mathbf{N}^\star$ in the singular basis is (close to) zero, and the interference of an update with the skill must be measured through off-diagonal energy. This is why Phase 2 is analysed through the column loadings g_i rather than through γ_i.

B.2 Proof of Lemma 1 (Spectral Loading)

Proof. Expanding $\mathbf{G}_T$ in its singular value decomposition,

$$\gamma_i = \mathbf{u}_i^\top \mathbf{G}_T \mathbf{v}_i = \sum_{j=1}^{s} \beta_j \, (\mathbf{u}_i^\top \mathbf{g}_j)(\mathbf{h}_j^\top \mathbf{v}_i). \tag{9}$$

Under Assumption 5, the two frames are independent, and the joint law of $\{\mathbf{g}_j\}$ is invariant under flipping the sign of any single $\mathbf{g}_j$, so $\mathbb{E}[(\mathbf{u}_i^\top \mathbf{g}_j)(\mathbf{u}_i^\top \mathbf{g}_l)] = 0$ for $j \neq l$ and the cross terms vanish in expectation. Hence

$$\mathbb{E}[\gamma_i^2] = \sum_j \beta_j^2 \, \mathbb{E}\big[(\mathbf{u}_i^\top \mathbf{g}_j)^2\big] \, \mathbb{E}\big[(\mathbf{h}_j^\top \mathbf{v}_i)^2\big].$$

Each $\mathbf{g}_j$ is uniform on the unit sphere of $\mathbb{R}^m$, so $\mathbb{E}[(\mathbf{u}_i^\top \mathbf{g}_j)^2] = 1/m$. Since $\mathbf{h}_j \in \mathcal{U}_T$, we have $\mathbf{h}_j^\top \mathbf{v}_i = \mathbf{h}_j^\top \mathbf{P}_T \mathbf{v}_i$, and since $\mathbf{h}_j$ is uniform on the unit sphere of the s-dimensional $\mathcal{U}_T$, $\mathbb{E}[\mathbf{h}_j \mathbf{h}_j^\top] = \mathbf{P}_T / s$ and $\mathbb{E}[(\mathbf{h}_j^\top \mathbf{P}_T \mathbf{v}_i)^2] = \|\mathbf{P}_T \mathbf{v}_i\|_2^2 / s$. Summing with $\sum_j \beta_j^2 = \|\mathbf{G}_T\|_F^2$ gives $\mathbb{E}[\gamma_i^2] = \|\mathbf{G}_T\|_F^2 \|\mathbf{P}_T \mathbf{v}_i\|_2^2 / (sm)$. Summing over i gives the expectation of ρ stated after Assumption 4. □

Throughout, we write $|\gamma_i| \asymp \|\mathbf{G}_T\|_F \|\mathbf{P}_T \mathbf{v}_i\|_2 / \sqrt{sm}$ for the typical magnitude of the loading, i.e., its root mean square. This is the sense in which Theorem 1 holds with loadings at their typical magnitude; fluctuations of individual loadings around it enter the ratio R as constant factors, and the maximum over the k leading indices adds at most a factor logarithmic in k.

Remark (why the diagonal is the right object). For a simple singular value, first-order perturbation theory gives $\mathrm{d}\sigma_i = \mathbf{u}_i^\top (\mathrm{d}\mathbf{W}) \mathbf{v}_i$ (Stewart & Sun, 1990; Horn & Johnson, 2012; Golub & Van Loan, 2013), so only the diagonal entries γ_i of $\mathbf{U}^\top \mathbf{G}_T \mathbf{V}$ govern first-order singular-value motion, and hence the SVF probe; the off-diagonal entries rotate singular vectors instead. Since the diagonal of a matrix can vanish identically while its Frobenius norm does not, no lower bound on $\max_i |\gamma_i|$ follows from $\|\mathbf{G}_T\|_F > 0$ alone. This is what Assumption 4 supplies. The diagonal is the right object for the probe of Phase 1; for the interference of Phase 2 it is the column loading of Lemma 3.

B.3 Column Loading at the Deployed Expert

Lemma 3 (Column loading)**.** *Under Assumption 6, for every unit vector $\mathbf{v} \in \mathbb{R}^n$ and every unit vector $\mathbf{u} \in \mathbb{R}^m$,*

$$\mathbb{E}\big[\|\mathbf{G}_T^\star \mathbf{v}\|_2^2\big] = \frac{\|\mathbf{G}_T^\star\|_F^2 \, \|\mathbf{P}_T \mathbf{v}\|_2^2}{s}, \qquad \mathbb{E}\big[(\mathbf{u}^\top \mathbf{G}_T^\star \mathbf{v})^2\big] = \frac{\|\mathbf{G}_T^\star\|_F^2 \, \|\mathbf{P}_T \mathbf{v}\|_2^2}{sm}.$$

In particular, for every basis vector $\mathbf{v}_i$ (including the completion $i > r$), $\mathbb{E}[g_i^2] = \|\mathbf{G}_T^\star\|_F^2 \|\mathbf{P}_T \mathbf{v}_i\|_2^2 / s$, and the diagonal entry $\mathbf{u}_i^\top \mathbf{G}_T^\star \mathbf{v}_i$ carries only a fraction $1/m$ of it in expectation.

Proof. $\mathbf{G}_T^\star \mathbf{v} = \sum_j \beta_j^\star \mathbf{g}_j^\star (\mathbf{h}_j^{\star\top} \mathbf{v})$. Since $\{\mathbf{g}_j^\star\}$ is orthonormal, $\|\mathbf{G}_T^\star \mathbf{v}\|_2^2 = \sum_j \beta_j^{\star 2} (\mathbf{h}_j^{\star\top} \mathbf{v})^2$ exactly, without cross terms. As $\mathbf{h}_j^\star \in \mathcal{U}_T$ is uniform on its unit sphere, $\mathbb{E}[(\mathbf{h}_j^{\star\top} \mathbf{v})^2] = \mathbb{E}[(\mathbf{h}_j^{\star\top} \mathbf{P}_T \mathbf{v})^2] = \|\mathbf{P}_T \mathbf{v}\|_2^2 / s$, and summing with $\sum_j \beta_j^{\star 2} = \|\mathbf{G}_T^\star\|_F^2$ gives the first identity. The second follows exactly as in the proof of Lemma 1, with $\mathbf{u}$ in place of $\mathbf{u}_i$. □

Lemma 3 reconciles two facts. At a converged expert the diagonal entries $\mathbf{u}_i^\top \nabla L_T(\mathbf{W}^\star) \mathbf{v}_i$ vanish (Assumption 6), yet the column loading g_i loses only this one entry out of m, so g_i still scales with the skill overlap $\|\mathbf{P}_T \mathbf{v}_i\|_2$, exactly as the Phase 1 loading γ_i does in Lemma 1. The probe therefore measures, through γ_i, the same overlap that governs interference at the deployed expert through g_i, and it does so for the input directions $\mathbf{v}_i$, which is why Eq. 3 applies the weights on the input side.

B.4 Proof of Lemma 2 (Two Regimes)

Proof. In the SVF parameterization $\mathbf{W}_\mathbf{z} = \sum_i z_i \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ we have $\partial \mathbf{W}_\mathbf{z} / \partial z_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Hence at $\mathbf{z} = \mathbf{1}$,

$$b_i := \frac{\partial L_T}{\partial z_i} = \left\langle \nabla_\mathbf{W} L_T, \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \right\rangle_F = \sigma_i (\gamma_i + \nu_i), \qquad \nu_i = \mathbf{u}_i^\top \mathbf{N} \mathbf{v}_i, \quad |\nu_i| \le \epsilon_{\text{noise}}, \tag{10}$$

and the Gauss–Newton curvature in the same coordinate is $H_{ii} = \sigma_i^2 c_i$, with $c_i = \langle \mathbf{u}_i \mathbf{v}_i^\top, \nabla_\mathbf{W}^2 L_T[\mathbf{u}_i \mathbf{v}_i^\top] \rangle_F$. Under the diagonal quadratic model of L_T around $\mathbf{z} = \mathbf{1}$, gradient flow reads $\dot{z}_i = -(b_i + H_{ii}(z_i - 1))$ with $z_i(0) = 1$, whose solution is

$$z_i(t) - 1 = -\frac{b_i}{H_{ii}} \left(1 - e^{-H_{ii} t}\right) = -\frac{\gamma_i + \nu_i}{\sigma_i c_i} \left(1 - e^{-\sigma_i^2 c_i t}\right). \tag{11}$$

Expanding for $H_{ii} t \ll 1$ gives $-b_i t = -t\, \sigma_i (\gamma_i + \nu_i)$, which is the one-step expression with learning rate t; letting $H_{ii} t \to \infty$ gives $-b_i / H_{ii} = -(\gamma_i + \nu_i)/(\sigma_i c_i)$. The crossover $H_{ii} t = 1$ occurs at $\sigma_i = (c_i t)^{-1/2}$. □

The dependence on σ_i therefore inverts between the two regimes. At fixed γ_i and c_i, the profile $|z_i(t) - 1|$ vanishes both as $\sigma_i \to 0$ and as $\sigma_i \to \infty$, with a maximum near the crossover. Because the crossover depends on σ_i, a single expert can hold its leading components in the saturated regime and its trailing components in the transient regime at the same time.

B.5 Proof of Theorem 1 (Detection)

Proof. Write $K = \|\mathbf{G}_T\|_F / \sqrt{sm}$, so that $|\gamma_i| \asymp K \|\mathbf{P}_T \mathbf{v}_i\|_2$ by Lemma 1, and neglect ν_i against γ_i. By Assumption 2, $\|\mathbf{P}_T \mathbf{v}_i\|_2 \le \sqrt{\delta}$ for $i \le k$ and $\|\mathbf{P}_T \mathbf{v}_{i^*}\|_2 \asymp \sqrt{s/n}$.

Saturated regime. By Lemma 2,

$$|z_i^* - 1| \;\asymp\; \frac{|\gamma_i|}{\sigma_i c_i} \;\asymp\; \frac{K}{c_i} \cdot \frac{\|\mathbf{P}_T \mathbf{v}_i\|_2}{\sigma_i}. \tag{12}$$

For $i \le k$, $\sigma_i \ge \sigma_k$ and $c_i \ge c_{\min}$, so $\max_{i \le k} |z_i^* - 1| \lesssim K\sqrt{\delta} / (c_{\min} \sigma_k)$, while $|z_{i^*}^* - 1| \gtrsim K\sqrt{s/n} / (c_{\max} \sigma_{i^*})$. Therefore

$$R = \frac{|z_{i^*}^* - 1|}{\max_{i \le k} |z_i^* - 1|} \;\gtrsim\; \frac{c_{\min}}{c_{\max}} \cdot \frac{\sqrt{s/n} \,/\, \sigma_{i^*}}{\sqrt{\delta} \,/\, \sigma_k} = \kappa^{-1} \frac{\sigma_k}{\sigma_{i^*}} \sqrt{\mu}. \tag{13}$$

Transient regime. By Lemma 2, $|z_i - 1| \approx t\, \sigma_i |\gamma_i|$, and the curvature does not enter. For $i \le k$, $\sigma_i \le \sigma_1$, so $\max_{i \le k} |z_i - 1| \lesssim tK\sigma_1 \sqrt{\delta}$, while $|z_{i^*} - 1| \gtrsim tK\sigma_{i^*} \sqrt{s/n}$, and therefore $R \gtrsim (\sigma_{i^*} / \sigma_1) \sqrt{\mu}$.

In either regime, if the lower bound on R exceeds one, then any threshold $\varepsilon \in [\max_{i \le k} |z_i^* - 1|, |z_{i^*}^* - 1|)$ places i^* in $\mathcal{S}_T$ while excluding every $i \le k$, hence $\mathcal{S}_T \cap \{k+1, \dots, r\} \ne \emptyset$ and $\mathcal{S}_T \not\subset \{1, \dots, k\}$. The bound exceeds one if and only if $\mu > \kappa^2 (\sigma_{i^*} / \sigma_k)^2$ in the saturated regime and $\mu > (\sigma_1 / \sigma_{i^*})^2$ in the transient regime. Finally, since $w_i = |z_i^* - 1| / \max_j |z_j^* - 1|$ divides every score by the same constant, $R = w_{i^*} / \max_{i \le k} w_i$, so $R > 1$ states equivalently that the penalty of Eq. 3 charges direction i^* more than any of the k leading directions. □

In the saturated regime the requirement becomes weaker the faster the spectrum decays between k and i^*: when $\sigma_k/\sigma_{i^*} > \kappa$, it holds already at chance-level misalignment $\mu = 1$, and it tolerates $\mu < 1$, i.e., a skill subspace that overlaps the leading directions more than chance. In the transient regime the requirement grows with the spread of the spectrum. When the leading components are saturated and i^* is still transient, combining the two cases of Lemma 2 gives $R \gtrsim c_{\min}\, t\, \sigma_k \sigma_{i^*} \sqrt{\mu}$, i.e., the saturated bound reduced by the factor $\sigma_{i^*}^2 c_{\max} t < 1$ by which i^* falls short of saturation: an off-principal component is detected most easily once it has itself saturated.

B.6 Proof of Proposition 1 (Stability of the Protection Weights)

Rules that select a subset, including the top-k comparisons of Section 4.2, need more than Proposition 1: they must recover membership, which requires a margin.

Lemma 4 (Binary support recovery)**.** *In the setting of Proposition 1, let $\hat{s}_i = |\hat{z}_i - 1|$ and $\Delta_{gap} = \min_{i \in \mathcal{S}_T} \theta_i - \max_{i \notin \mathcal{S}_T} \theta_i$. If $\Delta_{gap} > 2\tau_\eta$, top-$|\mathcal{S}_T|$ selection on $\hat{s}_i$ recovers $\mathcal{S}_T$ with probability at least $1 - \eta$.*

We prove Proposition 1 as part (a) and Lemma 4 as part (b).

Proof. Write $\theta_i = |z_i^* - 1|$ and $\hat{s}_i = |z_i^* + \xi_i - 1|$. By the reverse triangle inequality, $|\hat{s}_i - \theta_i| \le |\xi_i|$. Since each ξ_i is sub-Gaussian with parameter τ, $\mathbb{P}(|\xi_i| > x) \le 2e^{-x^2/(2\tau^2)}$, and a union bound over the r components gives $\mathbb{P}(\max_i |\xi_i| > \tau_\eta) \le 2r\, e^{-\log(2r/\eta)} = \eta$. We work on the complementary event, on which $|\hat{s}_i - \theta_i| \le \tau_\eta$ for every i.

(a) Let $\hat{M} = \max_j \hat{s}_j$. Then $|\hat{M} - M| \le \max_j |\hat{s}_j - \theta_j| \le \tau_\eta$, so $\hat{M} \ge M - \tau_\eta > 0$. For every i,

$$\hat{w}_i - w_i = \frac{\hat{s}_i - \theta_i}{\hat{M}} + \theta_i \frac{M - \hat{M}}{M\hat{M}}, \qquad |\hat{w}_i - w_i| \le \frac{\tau_\eta}{\hat{M}} + \frac{\theta_i}{M} \cdot \frac{\tau_\eta}{\hat{M}} \le \frac{2\tau_\eta}{\hat{M}} \le \frac{2\tau_\eta}{M - \tau_\eta},$$

using $\theta_i \le M$. For the penalty, $\mathcal{L}_{\text{ortho}}(\Delta\mathbf{W}; \mathbf{w})^{1/2} = \|\Delta\mathbf{W}\mathbf{V}\,\mathrm{diag}(\mathbf{w})\|_F$ is a seminorm of $\mathbf{w}$ for fixed $\Delta\mathbf{W}$, so by the triangle inequality

$$\begin{aligned} \big|\, \|\Delta\mathbf{W}\mathbf{V}\,\mathrm{diag}(\hat{\mathbf{w}})\|_F - \|\Delta\mathbf{W}\mathbf{V}\,\mathrm{diag}(\mathbf{w})\|_F \,\big| &\le \|\Delta\mathbf{W}\mathbf{V}\,\mathrm{diag}(\hat{\mathbf{w}} - \mathbf{w})\|_F \\ &\le \max_i |\hat{w}_i - w_i|\, \|\Delta\mathbf{W}\mathbf{V}\|_F \\ &\le \max_i |\hat{w}_i - w_i|\, \|\Delta\mathbf{W}\|_F, \end{aligned}$$

where the last step uses that $\mathbf{V}$ has orthonormal columns.

(b) For any $i \in \mathcal{S}_T$ and $j \notin \mathcal{S}_T$,

$$\hat{s}_i - \hat{s}_j \;\ge\; \theta_i - \theta_j - 2 \max_l |\hat{s}_l - \theta_l| \;\ge\; \Delta_{\text{gap}} - 2\tau_\eta \;>\; 0 \tag{14}$$

by hypothesis, so every member of $\mathcal{S}_T$ is ranked above every non-member, and top-k selection with $k = |\mathcal{S}_T|$ returns $\mathcal{S}_T$ exactly. □

Rollout budget. By Assumption 3, the per-coordinate variance of the policy-gradient estimate is at most V/B, so at a fixed number of Phase 1 updates the noise parameter scales as $\tau \propto \sqrt{V/B}$. The graded weights are accurate to $\max_i |\hat{w}_i - w_i| \le \varepsilon_w$ once $\tau_\eta \le \varepsilon_w M/(2 + \varepsilon_w)$, i.e., once $B \gtrsim V \log(2r/\eta)/(\varepsilon_w^2 M^2)$, whereas binary support recovery requires $B \gtrsim V \log(2r/\eta)/\Delta_{\text{gap}}^2$, up to a factor set by the number of updates in both cases. Both budgets grow logarithmically in the number of screened components. They differ in the margin: $\Delta_{\text{gap}} \le M$, and when the sorted deviations decay without a clear gap below the elbow, as in Figure 2b, Δ_{gap} can be much smaller than M, so a graded rule is estimated reliably from fewer rollouts than any rule that must decide membership.

B.7 Proof of Theorem 2 (Graded Protection Bound)

Proof. By ℓ-smoothness (Assumption 6),

$$L_T(\mathbf{W}^\star + \Delta\mathbf{W}) - L_T(\mathbf{W}^\star) \le \langle \mathbf{G}_T^\star, \Delta\mathbf{W} \rangle_F + \langle \mathbf{N}^\star, \Delta\mathbf{W} \rangle_F + \tfrac{\ell}{2}\varrho^2,$$

and $\langle \mathbf{N}^\star, \Delta\mathbf{W}\rangle_F \le \|\mathbf{N}^\star\|_F \|\Delta\mathbf{W}\|_F \le \epsilon^\star \varrho$ by Cauchy–Schwarz. Since $\{\mathbf{v}_i\}_{i=1}^n$ is an orthonormal basis of $\mathbb{R}^n$, $\sum_i \mathbf{v}_i\mathbf{v}_i^\top = \mathbf{I}_n$, so with $e_i = \|\Delta\mathbf{W}\mathbf{v}_i\|_2$,

$$\langle \mathbf{G}_T^\star, \Delta\mathbf{W}\rangle_F = \mathrm{tr}\Big(\mathbf{G}_T^{\star\top}\Delta\mathbf{W}\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i^\top\Big) = \sum_{i=1}^n (\mathbf{G}_T^\star\mathbf{v}_i)^\top(\Delta\mathbf{W}\mathbf{v}_i) \le \sum_{i=1}^n g_i e_i.$$

For any $\zeta > 0$, weighted Cauchy–Schwarz gives

$$\sum_{i=1}^n g_i e_i = \sum_{i=1}^n \frac{g_i}{\sqrt{w_i^2+\zeta}} \cdot \sqrt{w_i^2+\zeta}\, e_i \le \Big(\sum_{i=1}^n \frac{g_i^2}{w_i^2+\zeta}\Big)^{1/2}\Big(\sum_{i=1}^n (w_i^2+\zeta)e_i^2\Big)^{1/2}.$$

The second factor is $\big(\mathcal{L}_{\text{ortho}}(\Delta\mathbf{W}) + \zeta\varrho^2\big)^{1/2}$, since $\sum_i w_i^2 e_i^2 = \mathcal{L}_{\text{ortho}}(\Delta\mathbf{W})$ ($w_i = 0$ for $i > r$) and $\sum_i e_i^2 = \|\Delta\mathbf{W}\mathbf{V}_{\text{full}}\|_F^2 = \varrho^2$ for the orthogonal completed basis $\mathbf{V}_{\text{full}}$. This proves Eq. 4. Under the support condition, the terms with $w_i = 0$ vanish for every ζ, and for the others $g_i^2/(w_i^2+\zeta) \uparrow g_i^2/w_i^2$ as $\zeta \downarrow 0$; since the first-order term is bounded by $\Phi_\zeta(\mathbf{w})(\mathcal{L}_{\text{ortho}} + \zeta\varrho^2)^{1/2} \le \Phi_0(\mathbf{w})(\mathcal{L}_{\text{ortho}} + \zeta\varrho^2)^{1/2}$ for every ζ, letting $\zeta \to 0$ gives $\Phi_0(\mathbf{w})\mathcal{L}_{\text{ortho}}(\Delta\mathbf{W})^{1/2}$. □

Several layers. Stacking the adapted layers into one parameter vector, the same argument applied layerwise and Cauchy–Schwarz over layers give $\Delta L_T \le \big(\sum_l \Phi_{\zeta,l}^2\big)^{1/2}\big(\sum_l \mathcal{L}_{\text{ortho},l} + \zeta\varrho^2\big)^{1/2} + \varrho\,\epsilon^\star + \frac{\ell}{2}\varrho^2$ with $\varrho^2 = \sum_l \|\Delta\mathbf{W}_l\|_F^2$, which is the form used with the summed penalty of Eq. 2.

What the weights buy. Combining Lemmas 1, 2 and 3 at typical magnitude, the weight of component $i \le r$ is $w_i \asymp (K/M)\, a_i \|\mathbf{P}_T\mathbf{v}_i\|_2$, with $K = \|\mathbf{G}_T\|_F/\sqrt{sm}$ as in the proof of Theorem 1 and regime factor $a_i = 1/(\sigma_i c_i)$ when saturated and $a_i = t\sigma_i$ when transient, while $g_i \asymp (\|\mathbf{G}_T^\star\|_F/\sqrt{s})\,\|\mathbf{P}_T\mathbf{v}_i\|_2$. The overlap cancels in the ratio,

$$\frac{g_i}{w_i} \asymp \frac{C_\star}{a_i}, \qquad C_\star = \frac{\|\mathbf{G}_T^\star\|_F\, M}{\sqrt{s}\, K},$$

so $\Phi_0(\mathbf{w})^2 \asymp C_\star^2 \sum_i a_i^{-2}$ over the components with non-zero overlap: every such component is priced, and the price per unit of skill loading depends only on its spectral regime, never on whether it lies above or below a rank cut-off. A binary rule prices every component outside its set at zero, which makes Φ_ζ diverge as $\zeta \to 0$ unless the skill has no loading there; Corollary 1 shows that this divergence is not an artefact of the bound.

Rectangular layers. For $i > r$ (the null space of $\mathbf{W}$, non-trivial when $d_{\text{in}} > d_{\text{out}}$, as in down projections) the weight is $w_i = 0$, and by Lemma 3 $\mathbb{E}[g_i^2] > 0$ unless $\mathcal{U}_T \subset \mathrm{span}\{\mathbf{v}_1, \dots, \mathbf{v}_r\}$. Splitting the sum in the proof into $i \le r$ and $i > r$ then gives the first-order term $\Phi_0^{(\le r)}(\mathbf{w})\,\mathcal{L}_{\text{ortho}}^{1/2} + \|\mathbf{G}_T^\star\mathbf{\Pi}_\perp\|_F\,\varrho$, with $\mathbf{\Pi}_\perp$ the projector onto the null space of $\mathbf{W}$. The second term is a residual leak that no rule built on the singular vectors of $\mathbf{W}$ controls, whether graded or rank-fixed; for square layers ($d_{\text{in}} \le d_{\text{out}}$, as in gate and up projections) it is absent.

B.8 Corollaries

Detailed version of Corollary 1 (Rank-fixed protection as the binary case). Let $\mathcal{S}$ be a protected index set, $\mathbf{\Pi}_\mathcal{S} = \sum_{i\in\mathcal{S}} \mathbf{v}_i\mathbf{v}_i^\top$, and $\mathcal{A}_\mathcal{S} = \{\Delta\mathbf{W} \mid \Delta\mathbf{W}\mathbf{v}_i = \mathbf{0}\ \forall i \in \mathcal{S}\}$ the updates admitted by hard input-side protection. Then

$$\begin{aligned} \varrho\,\big\|\mathbf{G}_T^\star(\mathbf{I}-\mathbf{\Pi}_\mathcal{S})\big\|_F - \varrho\,\epsilon^\star - \tfrac{\ell}{2}\varrho^2 \;&\le\; \max_{\Delta\mathbf{W}\in\mathcal{A}_\mathcal{S},\,\|\Delta\mathbf{W}\|_F=\varrho} \Delta L_T \\ &\le\; \varrho\,\big\|\mathbf{G}_T^\star(\mathbf{I}-\mathbf{\Pi}_\mathcal{S})\big\|_F + \varrho\,\epsilon^\star + \tfrac{\ell}{2}\varrho^2, \end{aligned} \tag{15}$$

with $\|\mathbf{G}_T^\star(\mathbf{I}-\mathbf{\Pi}_\mathcal{S})\|_F^2 = \sum_{i\notin\mathcal{S}} g_i^2$.

Proof. *Upper bound.* Take $w_i = \mathbb{I}[i \in \mathcal{S}]$. Every $\Delta\mathbf{W} \in \mathcal{A}_\mathcal{S}$ has $\mathcal{L}_{\text{ortho}}(\Delta\mathbf{W}) = 0$, and $\Phi_\zeta(\mathbf{w})^2\zeta\varrho^2 = \varrho^2\big(\frac{\zeta}{1+\zeta}\sum_{i\in\mathcal{S}} g_i^2 + \sum_{i\notin\mathcal{S}} g_i^2\big) \to \varrho^2\sum_{i\notin\mathcal{S}} g_i^2$ as $\zeta \to 0$, so Theorem 2 gives the right-hand side of Eq. 15. *Lower bound.* Take $\Delta\mathbf{W} = \varrho\,\mathbf{G}_T^\star(\mathbf{I}-\mathbf{\Pi}_\mathcal{S})/\|\mathbf{G}_T^\star(\mathbf{I}-\mathbf{\Pi}_\mathcal{S})\|_F$, which satisfies

$\Delta\mathbf{W}\mathbf{v}_i = \mathbf{0}$ for $i \in \mathcal{S}$ and $\|\Delta\mathbf{W}\|_F = \varrho$. By ℓ-smoothness, $\Delta L_T \geq \langle \mathbf{G}_T^\star + \mathbf{N}^\star, \Delta\mathbf{W}\rangle_F - \frac{\ell}{2}\varrho^2$, with $\langle \mathbf{G}_T^\star, \Delta\mathbf{W}\rangle_F = \varrho\|\mathbf{G}_T^\star(\mathbf{I} - \mathbf{\Pi}_\mathcal{S})\|_F$ because $\mathbf{I} - \mathbf{\Pi}_\mathcal{S}$ is an orthogonal projector, and $\langle \mathbf{N}^\star, \Delta\mathbf{W}\rangle_F \geq -\varrho\,\epsilon^\star$ by Cauchy–Schwarz. Finally, $\|\mathbf{G}_T^\star(\mathbf{I} - \mathbf{\Pi}_\mathcal{S})\|_F^2 = \sum_{i\notin\mathcal{S}} \|\mathbf{G}_T^\star \mathbf{v}_i\|_2^2$ in the completed basis. □

The two sides of Eq. 15 share the leak $\big(\sum_{i\notin\mathcal{S}} g_i^2\big)^{1/2}$, so the graded bound is tight in the binary case up to the noise and curvature terms, and hard protection of $\mathcal{S}$ cannot remove a first-order degradation proportional to the skill loading it leaves outside $\mathcal{S}$, however strictly it is enforced. For $\mathcal{S} = \{1, \ldots, k\}$, Theorem 1 places $i^* \notin \mathcal{S}$ in $\mathcal{S}_T$, and by Lemma 3 with chance-level overlap $\mathbb{E}[g_{i^*}^2] \asymp \|\mathbf{G}_T^\star\|_F^2/n > 0$, so the leak is strictly positive in expectation. OPLoRA protects bilaterally, $\{\Delta\mathbf{W} \mid \Delta\mathbf{W}\mathbf{v}_i = \mathbf{0},\ \mathbf{u}_i^\top \Delta\mathbf{W} = \mathbf{0}^\top\ \forall i \leq k\}$; taking $\Delta\mathbf{W} \propto (\mathbf{I} - \mathbf{\Pi}_\mathcal{S}^U)\mathbf{G}_T^\star(\mathbf{I} - \mathbf{\Pi}_\mathcal{S})$, with $\mathbf{\Pi}_\mathcal{S}^U = \sum_{i\in\mathcal{S}} \mathbf{u}_i\mathbf{u}_i^\top$, in the same argument gives the leak $\|(\mathbf{I} - \mathbf{\Pi}_\mathcal{S}^U)\mathbf{G}_T^\star(\mathbf{I} - \mathbf{\Pi}_\mathcal{S})\|_F$, whose expectation by the second identity of Lemma 3, summed over the completed output basis, is $(1 - k/m)\sum_{i\notin\mathcal{S}} \mathbb{E}[g_i^2]$, so the conclusion is unchanged. PiSSA, MiLoRA, and CorDA fix their boundary only at initialization and impose no constraint during training, so their admissible set is all updates, $\mathcal{S} = \emptyset$, and the lower bound of Eq. 15 holds with the full loading $\|\mathbf{G}_T^\star\|_F$.

Corollary 2 (Role of the orthogonality weight)**.** *If Phase 2 starts from $\mathbf{B}_l = \mathbf{0}$ and returns parameters with $\mathcal{L}_{total} \leq \mathcal{L}_{total}(0)$, then $\sum_l \mathcal{L}_{ortho,l} \leq \big(\mathcal{L}_{SFT}(0) - \mathcal{L}_{SFT}\big)/\lambda_{ortho} \leq \mathcal{L}_{SFT}(0)/\lambda_{ortho}$, and Eq. 4 holds with $\mathcal{L}_{ortho}(\Delta\mathbf{W})$ replaced by $\big(\mathcal{L}_{SFT}(0) - \mathcal{L}_{SFT}\big)/\lambda_{ortho}$. Under the support condition of Theorem 2 the first-order term therefore decays as $\lambda_{ortho}^{-1/2}$, whereas the leak of Corollary 1 does not depend on the strength of the constraint.*

Proof. With $\mathbf{B}_l = \mathbf{0}$ at initialization, $\Delta\mathbf{W}_l = \mathbf{0}$ and $\mathcal{L}_{\text{ortho},l} = 0$ for every l, so $\mathcal{L}_{\text{total}}(0) = \mathcal{L}_{\text{SFT}}(0)$. If the returned parameters satisfy $\mathcal{L}_{\text{total}} \leq \mathcal{L}_{\text{total}}(0)$, then $\mathcal{L}_{\text{SFT}} + \lambda_{\text{ortho}} \sum_l \mathcal{L}_{\text{ortho},l} \leq \mathcal{L}_{\text{SFT}}(0)$, i.e., $\sum_l \mathcal{L}_{\text{ortho},l} \leq (\mathcal{L}_{\text{SFT}}(0) - \mathcal{L}_{\text{SFT}})/\lambda_{\text{ortho}}$, and $\mathcal{L}_{\text{SFT}} \geq 0$ for the cross-entropy loss gives the second inequality. Substituting into the several-layer form of Theorem 2 gives

$$\Delta L_T \leq \Phi_\zeta\Big(\frac{\mathcal{L}_{\text{SFT}}(0) - \mathcal{L}_{\text{SFT}}}{\lambda_{\text{ortho}}} + \zeta\varrho^2\Big)^{1/2} + \varrho\,\epsilon^\star + \tfrac{\ell}{2}\varrho^2.$$

□

Here $\mathcal{L}_{\text{SFT}}$ is the objective of Eq. 2 on the full Phase 2 training set. The descent condition $\mathcal{L}_{\text{total}} \leq \mathcal{L}_{\text{total}}(0)$ is not implied by stochastic optimization, but it is a property of the returned checkpoint alone and can be verified directly on it. The bound makes the trade-off of Section 4.4 explicit: the constraint weight enters only through $(\mathcal{L}_{\text{SFT}}(0) - \mathcal{L}_{\text{SFT}})/\lambda_{\text{ortho}}$, the product of the facts acquired and the inverse weight, so a larger λ_{ortho} tightens the first-order skill term only as long as it does not reduce acquisition by the same factor, and it cannot tighten the terms $\zeta\varrho^2$, $\varrho\,\epsilon^\star$, and $\frac{\ell}{2}\varrho^2$ at all.

Corollary 3 (SVF as a diagnostic probe)**.** *In the saturated regime, a component with a small singular value receives a larger score than a leading component with the same skill overlap; in the transient regime the ordering is reversed, so interpreting the weights $\mathbf{w}$, and $\mathcal{I}_{crit}$ as their binary case, requires the regime of the expert.*

Detailed version of Corollary 3 (SVF as a diagnostic probe). In the saturated regime, Eq. 12 gives $|z_i^* - 1| \asymp \|\mathbf{P}_T\mathbf{v}_i\|_2/(\sigma_i c_i)$ up to a common prefactor, so the probe score couples skill overlap with the *inverse* singular value. A component with a small raw singular value and non-negligible skill overlap therefore receives a larger score than a leading component with the same overlap, up to the curvature spread κ. In the transient regime, Eq. 11 gives $|z_i - 1| \approx t\,\sigma_i|\gamma_i|$ and the ordering is reversed. Which regime the expert occupies must therefore be known in order to interpret the weights $\mathbf{w}$, and the set $\mathcal{I}_{\text{crit}}$ that binarizes them. The same regime factor a_i is what remains of g_i/w_i in Theorem 2.

Corollary 4 (Multi-skill union)**.** *For M skill-critical sets of k indices per layer with mean pairwise overlap fraction p, $Mk\big(1 - \frac{(M-1)p}{2}\big) \leq \big|\bigcup_m \mathcal{S}_{T_m}\big| \leq Mk$, with equality on the left at $M = 2$. For graded weights combined as $w_i^\cup = \max_m w_i^{(m)}$, $\max_m \mathcal{L}_{ortho}(\Delta\mathbf{W}; \mathbf{w}^{(m)}) \leq \mathcal{L}_{ortho}(\Delta\mathbf{W}; \mathbf{w}^\cup) \leq \sum_m \mathcal{L}_{ortho}(\Delta\mathbf{W}; \mathbf{w}^{(m)})$, and $\Phi_\zeta(\mathbf{w}^\cup) \leq \Phi_\zeta(\mathbf{w}^{(m)})$ for the loadings of every skill m.*

Detailed version of Corollary 4 (Multi-skill union). Let $\mathcal{S}_{T_1}, \ldots, \mathcal{S}_{T_M}$ be the skill-critical sets of M skills, each of size k per layer, with mean pairwise overlap fraction $p = \binom{M}{2}^{-1} \sum_{a<b} |\mathcal{S}_{T_a} \cap$

$\mathcal{S}_{T_b}|/k$. By the Bonferroni inequalities, truncating inclusion–exclusion after the pairwise term gives a lower bound,

$$Mk\Big(1 - \tfrac{(M-1)p}{2}\Big) \;\le\; \Big|\bigcup_{m=1}^{M} \mathcal{S}_{T_m}\Big| \;\le\; Mk, \tag{16}$$

with equality on the left at $M = 2$, where inclusion–exclusion is exact, and with the gap on the left at $M = 3$ equal to the number of indices shared by all three sets. The upper bound Mk is attained only when the sets are disjoint, $p = 0$. The corollary thus requires no incoherence assumption between skill subspaces, only the measured quantity p: Appendix D finds $p \approx 0.24$ for our three skills, an order of magnitude above the chance level $k/r \approx 0.016$, so the skill-critical sets are coherent rather than incoherent.

Graded version. Let $w_i^{\cup} = \max_m w_i^{(m)}$. For every i, $\big(w_i^{(m)}\big)^2 \le \big(w_i^{\cup}\big)^2 \le \sum_{m'} \big(w_i^{(m')}\big)^2$; multiplying by $\|\Delta\mathbf{W}\mathbf{v}_i\|_2^2 \ge 0$ and summing over i gives

$$\max_m \mathcal{L}_{\text{ortho}}(\Delta\mathbf{W}; \mathbf{w}^{(m)}) \;\le\; \mathcal{L}_{\text{ortho}}(\Delta\mathbf{W}; \mathbf{w}^{\cup}) \;\le\; \sum_{m=1}^{M} \mathcal{L}_{\text{ortho}}(\Delta\mathbf{W}; \mathbf{w}^{(m)}). \tag{17}$$

Moreover $w_i^{\cup} \ge w_i^{(m)}$ implies $\Phi_\zeta(\mathbf{w}^{\cup}) \le \Phi_\zeta(\mathbf{w}^{(m)})$ for the loadings $g_i^{(m)}$ of every skill m, so Theorem 2 holds for each skill under the joint penalty with a constant no larger than under its own. The union of top-k sets in Eq. 16 is the binary case of $\mathbf{w}^{\cup}$, and Eq. 17 is its graded counterpart: the joint penalty costs at most the sum of the individual ones and, when the weight profiles overlap as the index sets do, much less.

B.9 Connection to Prior Work

Our theoretical framework provides a unifying geometric perspective on several recent PEFT methods. **SVFT** (Lingam et al., 2024) learns sparse combinations of singular vectors, with the insight that fine-grained control over expressivity comes from choosing *which* singular components to update. This is the same degree of freedom that Theorem 1 identifies. **PiSSA** (Meng et al., 2024), **MiLoRA** (Wang et al., 2025), and **CorDA** (Yang et al., 2024) instead fix that choice in advance, by principal, minor, or context-oriented rank, and **OPLoRA** (Xiong & Xie, 2026) enforces a fixed top-k choice during training. While these choices are suitable for general fine-tuning, Corollary 1 implies that a rank-fixed rule is sub-optimal for decoupled adaptation whenever it leaves skill loading outside the protected set, since in the saturated regime the components engaged by a skill are ranked by overlap divided by singular value and need not coincide with either end of the spectrum. Theorem 2 places all of these rules and PALoRA on one scale: each corresponds to a weight vector $\mathbf{w}$ (binary, and for initialization-only methods effectively zero during training), and the constant $\Phi_\zeta(\mathbf{w})$ measures how much skill loading the rule leaves unpriced.

This interpretation is consistent with our measurements. PiSSA falls below unconstrained LoRA in skill retention and MiLoRA stays level with it (Table 3), which indicates that selecting components by singular-value rank, at either end of the spectrum, does not by itself protect a distributed skill.

B.10 Geometric Interpretation and Limitation

The proof formalizes a simple geometric picture. Pre-training aligns the principal singular directions of the weight matrix with directions of high variance in the corpus, and these directions are largely task-agnostic. A reasoning skill, by contrast, activates a task-specific subspace that can be substantially misaligned with those principal directions. As a result, skill-relevant information may live in singular components that are not among the largest singular values.

Four limitations bound the reach of the results. The first is the misalignment condition of Assumption 2: if a skill aligns strongly with dominant corpus directions, then μ decreases and the detection margin narrows, although in the saturated regime Theorem 1 still separates provided the spectrum decays fast enough between k and i^*. The second is Assumption 4: a skill gradient with vanishing diagonal loading is invisible to any singular-value probe regardless of misalignment, and a probe over singular vectors rather than singular values would be required to cover it. The third is the training regime: the conclusions reverse between the transient and the saturated regime of Lemma 2, and the

regime of each component depends on its curvature c_i, which we do not estimate. The fourth concerns the injection guarantee: Theorem 2 is a first-order bound under the smoothness and gradient structure of Assumption 6, it bounds a worst case over updates rather than predicting the degradation of a particular run, it leaves the null-space leak of rectangular layers uncontrolled, and Corollary 2 relies on the descent condition of the returned checkpoint. The theorems should therefore be interpreted as a structured explanation of when SVF-guided weighting can reach beyond the leading singular directions and of what the graded penalty controls, tested behaviorally in Section 4.2 and Section 4.4, rather than as a universal statement for every task.

C Sequential Knowledge Injection

To test the continual-adaptation setting invoked in our motivation, we split the 1000-fact set into four disjoint rounds of 250 facts and carry a single LoRA adapter across rounds. Each round is initialized from the checkpoint of the preceding round, while the frozen SVF expert remains fixed throughout, and each round trains for 10 epochs. Table 5 in Section 4.5 compares two unconstrained controls, plain sequential LoRA (A) and SVF + sequential LoRA (B), and the orthogonal-subspace baseline SVF + OPLoRA (O) with PALoRA (D) under this protocol. Configuration (A) carries no expert, so it isolates what sequential injection does when nothing protects the skill. (O) protects the top-128 singular directions of the expert weights, computed once before the first round. (D) uses the same orthogonality weight as row (D) of Table 4(a). All rows are means over three seeds; the standard deviation across seeds is at most 3.7 GSM8K and 4.7 recall points for (A), (B), and (D), and at most 6.8 and 7.3 for (O).

Measured from the model each starts from, both unconstrained controls lose a similar amount of skill: (A) ends 8.0 GSM8K points below the base model, and (B) ends 9.2 points below the expert and 3.9 below the base model. The expert therefore does not slow the degradation, it only raises the level at which it begins. OPLoRA's fixed top-k projection slows the loss to 3.5 points but still ends 2.1 points below the base model, with the largest spread across seeds. PALoRA loses 1.5 points, remains above the base model at every round, and varies least across seeds; from round 2 on it has the highest GSM8K accuracy, and at round 4 it leads (B) by 5.0 points and (O) by 3.2. What changes the trajectory is the constraint rather than the expert.

The constraint does reduce recall in the early rounds. PALoRA's average recall is below that of (B) in the first two rounds (by 2.8 and 2.9 points), level with it at round 3, and 1.6 points higher at round 4. OPLoRA recalls the most at every round, 4.5 points more than PALoRA at round 4, although its seed-to-seed spread (up to 7.3 points) is larger than that gap. Measured with the SVF-weighted penalty PALoRA optimizes, the energy of the unconstrained update in (B) grows steadily with the rounds ($49.9 \rightarrow 238.6$), whereas PALoRA's grows far more slowly ($33.4 \rightarrow 94.3$), about 40% of (B)'s at round 4.

D Toward Multi-Skill Preservation

Corollary 4 bounds the size of the union of skill-critical index sets in terms of their pairwise overlap p, and the joint graded penalty in terms of the individual ones. This raises the practical question of how quickly the protected subspace grows when several skills must be preserved at once.

The natural extension of PALoRA to M skills is to take the elementwise maximum of the per-skill protection weights and to deploy an equally weighted mixture of the corresponding experts; the union of the top-k index sets is the binary case of that maximum, and it is what Eq. 16 bounds, while Eq. 17 bounds the graded penalty itself. Its growth can be measured directly from the frozen experts we already have, without additional training. Table 9 reports this growth for our three skills on Llama-3.1-8B, with $k = 64$ per layer over 96 adapted MLP layers of rank 4096.

The union grows sublinearly. The second skill costs 48.7 additional directions per layer and the third only 39.5, because the per-skill index sets overlap substantially. Pairwise, the experts share between 14.8 and 16.2 indices per layer, or 23 to 25% of k, against a chance level of 1.0 for independent uniform draws of 64 out of 4096. In addition, 5.9 directions per layer are selected by all three experts, against a chance level of 0.016, and no layer exhibits zero overlap for any pair. Protecting all three skills therefore consumes 3.71% of the spectrum, rather than the 4.69% that a disjoint union would

Table 9: Growth of the protected subspace under the union of SVF-identified critical indices across skills. Pairwise values are means over the three pairs.

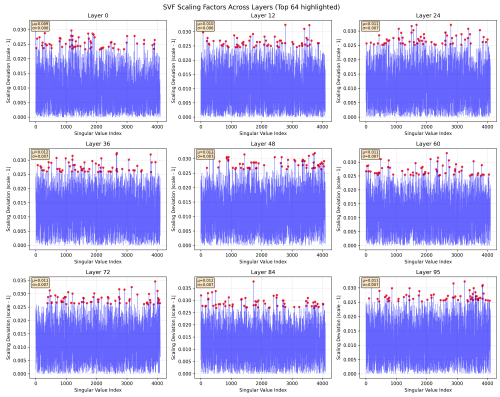
Skills in the union	Dir./layer	If disjoint	Of spectrum	Of disjoint bound
1 (single skill)	64.0	64	1.56%	100.0%
2 (mean over pairs)	112.7	128	2.75%	88.1%
3 (GSM8K, MBPP, AI2-ARC)	152.2	192	3.71%	79.2%

require. These values match Corollary 4: the mean pairwise overlap of 15.3 directions per layer gives $p \approx 0.24$, the bound of Eq. 16 is exact at $M = 2$, and at $M = 3$ it gives 146.1 directions per layer, below the measured 152.2 by the 5.9 directions shared by all three experts, up to rounding.

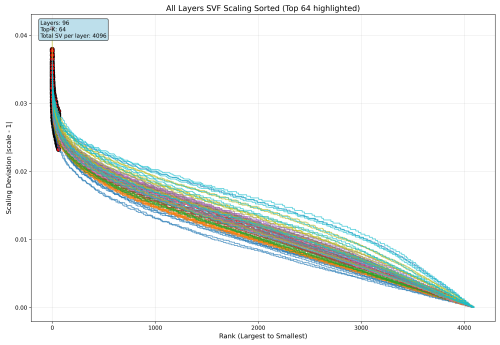
We emphasize that this is a statement about the geometry of the protected subspace rather than a retention result. It shows that the plasticity budget is eroded sublinearly in the number of preserved skills, and that multi-skill protection remains viable at our choice of k. Verifying end-to-end retention under joint training is left to future work.

E Additional SVF Adapter Visualizations

We provide the same SVF visualizations as in Figure 2 for all skill experts across both model families. For each SVF expert, we show (i) the sorted scaling magnitudes $|z_i^* - 1|$ across MLP layers and (ii) the top-64 SVF-selected components, sorted by $|z_i^* - 1|$ from largest to smallest.

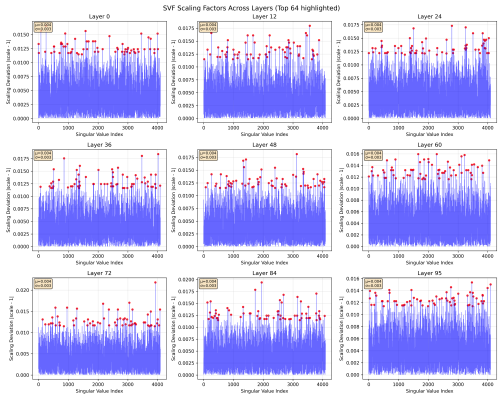


(a) SVF scaling magnitudes $|z_i^* - 1|$ over singular-value index.

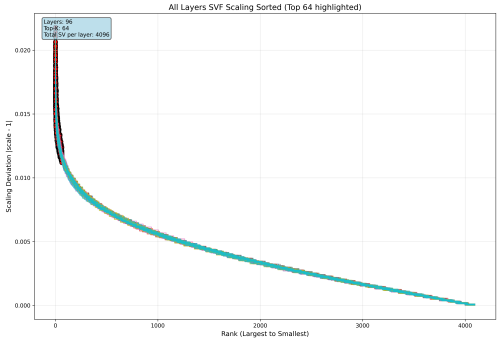


(b) Top-64 SVF scaling magnitudes $|z_i^* - 1|$, sorted from largest to smallest.

Figure 4: SVF visualization for the Llama-3.1-8B-Instruct GSM8K expert.

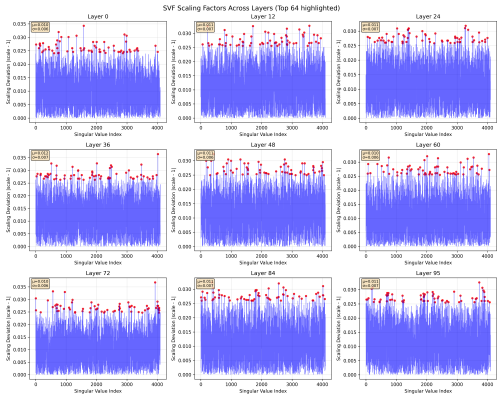


(a) SVF scaling magnitudes $|z_i^* - 1|$ over singular-value index.

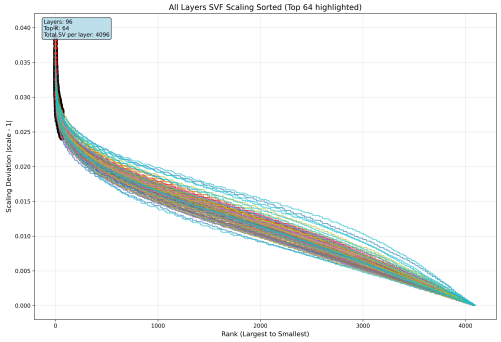


(b) Top-64 SVF scaling magnitudes $|z_i^* - 1|$, sorted from largest to smallest.

Figure 5: SVF visualization for the Llama-3.1-8B-Instruct MBPP expert.

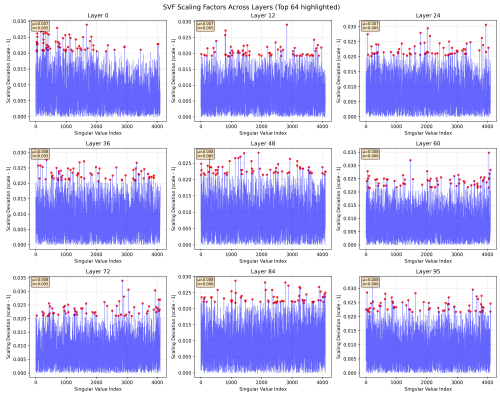


(a) SVF scaling magnitudes $|z_i^* - 1|$ over singular-value index.

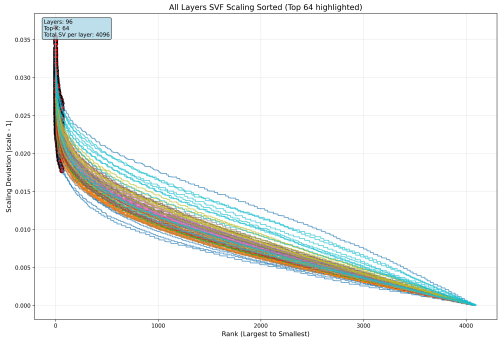


(b) Top-64 SVF scaling magnitudes $|z_i^* - 1|$, sorted from largest to smallest.

Figure 6: SVF visualization for the Llama-3.1-8B-Instruct AI2-ARC expert.

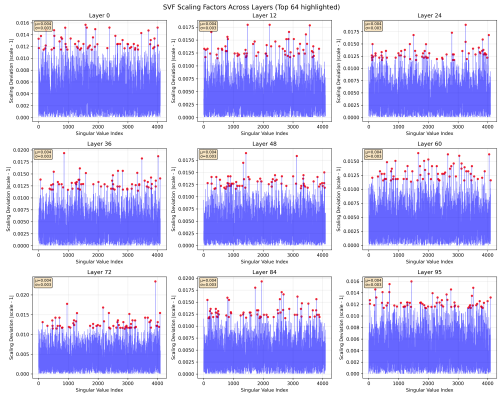


(a) SVF scaling magnitudes $|z_i^* - 1|$ over singular-value index.

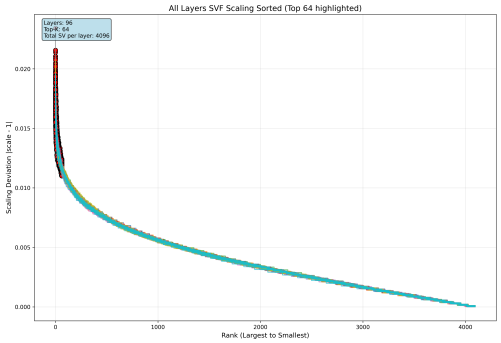


(b) Top-64 SVF scaling magnitudes $|z_i^* - 1|$, sorted from largest to smallest.

Figure 7: SVF visualization for the Mistral-7B-Instruct-v0.3 GSM8K expert.

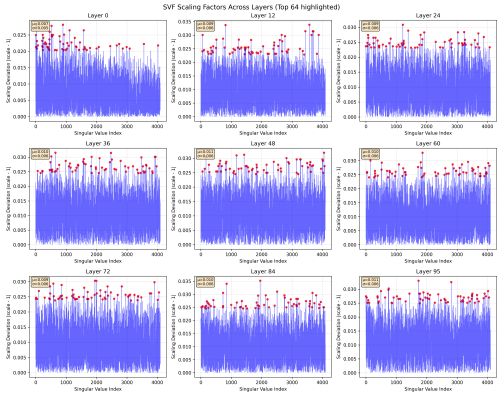


(a) SVF scaling magnitudes $|z_i^* - 1|$ over singular-value index.

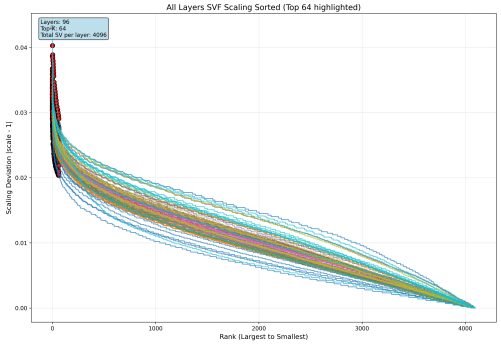


(b) Top-64 SVF scaling magnitudes $|z_i^* - 1|$, sorted from largest to smallest.

Figure 8: SVF visualization for the Mistral-7B-Instruct-v0.3 MBPP expert.



(a) SVF scaling magnitudes $|z_i^* - 1|$ over singular-value index.



(b) Top-64 SVF scaling magnitudes $|z_i^* - 1|$, sorted from largest to smallest.

Figure 9: SVF visualization for the Mistral-7B-Instruct-v0.3 AI2-ARC expert.

F DATASETS, TRAINING, AND EVALUATION PROTOCOL

This section summarizes the datasets used in PALoRA and clarifies their roles in the two-phase training and evaluation pipeline. We distinguish between *skill* datasets, which define the capabilities to be preserved and are used to train the SVF experts in Phase 1, and *knowledge* data, which provide the factual supervision used for LoRA-based knowledge injection in Phase 2. This separation is central to PALoRA: the skill data identify the subspace to protect, while the knowledge data define the new information to be acquired. Table 10 consolidates the experimental configuration summarized in Section 4.1.

All experiments were conducted in the same hardware environment with NVIDIA A100 SXM GPUs (80GB VRAM each), 16 vCPUs, and 117GB system memory. PALoRA Phase 1 used 2 GPUs, while all other experiments were run on a single GPU. Table 10 summarizes the experimental setup.

Table 10: Experimental setup.

Models	Llama-3.1-8B-Instruct (main experiments and ablations) and Mistral-7B-Instruct-v0.3 (cross-architecture validation).
Skills	GSM8K (math, final-answer accuracy), MBPP (code generation, pass@1), AI2-ARC Challenge (scientific reasoning, accuracy).
Knowledge injection	The `Unknown` split of the corpus of Pletenev et al. (2025), released for these base models. By construction, these facts are not recalled before fine-tuning, so all measured recall is acquired during injection. The split is multi-domain and covers all 13 domains of the taxonomy reported in Appendix F.3.2. We inject sets of size 100, 250, 500, and 1000; recall is the fraction of previously unknown facts answered correctly after fine-tuning.
Baselines	LoRA, PiSSA (Meng et al., 2024), MiLoRA (Wang et al., 2025), OPLoRA (Xiong & Xie, 2026), and CorDA (Yang et al., 2024). Every baseline is trained on and deployed with the same Phase 1 SVF expert as PALoRA, so the comparison isolates the injection rule (Appendix G).
Defaults	We use adapter rank $r = 16$, scaling $\alpha = 32$, orthogonality weight λ_{ortho}, which controls how much plasticity is allowed while stability is maintained (Section 4.4), batch size 8, LoRA dropout 0.05, and the AdamW optimizer. PALoRA has no protected-subspace size k: the constraint is graded over the full spectrum. For Llama-3.1-8B-Instruct, we use a learning rate of 2×10^{-4} and 10 epochs. For Mistral-7B-Instruct-v0.3, we use a learning rate of 7×10^{-5} and 5 epochs. SVF experts are trained for 10 iterations, each a single policy-gradient update over a batch of 256 sampled prompts, using the default hyperparameters of the original implementation.
Seeds	Every injection run is repeated under seeds 42, 1234, and 2024, and the values we report are means over the three.

F.1 OVERVIEW OF DATASET ROLES

In our experiments, GSM8K, MBPP, and AI2-ARC serve as skill datasets. Each of them is used to train a separate SVF expert in Phase 1, and each reappears after Phase 2 only as an evaluation benchmark for measuring skill preservation. They are therefore not used to optimize the knowledge-injection adapter.

The knowledge-injection data consist of TriviaQA-style factual question–answer pairs labeled according to whether the frozen base model can already answer them reliably. During Phase 2, these examples are used to supervise factual adaptation, whereas final knowledge recall is measured on the subset of facts labeled `Unknown`. Items labeled `HighlyKnown` are part of the Phase 2 training mixture only; they are not a target of factual acquisition and are not evaluated.

F.2 SKILL DATASETS

F.2.1 GSM8K

GSM8K (Cobbe et al., 2021) is a benchmark of grade-school mathematical word problems and serves as our mathematical reasoning task. In Phase 1, we train a dedicated SVF expert on the

GSM8K training split, using a lightweight held-out partition for checkpoint selection during the reinforcement-learning loop. GSM8K is not used during Phase 2 optimization.

After knowledge injection, skill preservation is evaluated with final-answer accuracy on the GSM8K benchmark. The model generates a free-form solution, and the prediction is scored by extracting the final numerical answer from the completion and comparing it against the ground-truth value. This metric isolates whether mathematical reasoning remains intact after factual updating.

F.2.2 MBPP

MBPP (Austin et al., 2021) is a benchmark for Python code generation from natural-language problem descriptions and serves as our coding-skill task. In Phase 1, we train an SVF expert on a shuffled subset of MBPP problems, with disjoint training and validation partitions used for learning and checkpoint selection. MBPP is not used in Phase 2 training.

After adaptation, we evaluate skill preservation with `pass@1` under execution-based evaluation. For each problem, the model generates a single program, which is then executed against the corresponding unit tests. A sample is marked correct only if the generated code passes the test suite, so the metric reflects functional program synthesis rather than surface-level similarity.

F.2.3 AI2-ARC

We use the AI2-ARC benchmark (Clark et al., 2018) as our science reasoning task. In Phase 1, the SVF expert is trained on ARC-Easy, again using a held-out partition for checkpoint selection. Phase 2 does not optimize on ARC data.

For evaluation, we report multiple-choice exact-match accuracy. Questions are presented with lettered answer options, and a prediction is counted as correct if the model selects the correct option after generation. This metric measures whether the scientific reasoning ability associated with the protected SVF subspace is preserved after factual adaptation.

F.3 Knowledge-injection dataset

F.3.1 Unknown facts

Knowledge injection is performed using TriviaQA-style factual question–answer pairs that are labeled according to whether the frozen base model can already answer them reliably (Pletenev et al., 2025). The primary target of adaptation is the subset labeled `Unknown`, which contains facts that are not robustly recalled before fine-tuning. These examples define the factual content to be injected during Phase 2.

For each experiment, we construct a knowledge-injection set by sampling a fixed number of `Unknown` items. In the main paper, we consider subsets of size 100, 250, 500, and 1000. Each subset defines an independent Phase 2 run, so the factual load can be varied while keeping the underlying protocol unchanged. The sets are nested, so varying the size varies the factual load and nothing else.

F.3.2 Domain composition

The injected facts are not concentrated in a single subject area. Labeling the `Unknown` pool against a 13-domain taxonomy, every domain is represented with at least 1% of the set: film, television and theatre (12.3%); geography and places (8.5%); sports and games (8.5%); literature and language (8.3%); music (6.4%); history and military (5.2%); business, technology and media (4.9%); society and everyday life (4.7%); science and nature (4.5%); politics, law and government (3.6%); food and drink (2.6%); art and architecture (2.1%); and religion, myth and philosophy (1.8%). The remainder is left unclassified by the lexicon.

The normalized entropy of the resulting domain distribution is 0.95, where 1.0 corresponds to a uniform distribution. The facts also vary in the kind of knowledge asserted, covering person, place, temporal, quantity, creative-work, organism, and open-ended attribute targets. We note that these domain labels are inferred rather than gold, since the source corpus provides no topic annotation.

F.3.3 HIGHLYKNOWN FACTS

In addition to the Unknown subset, we include examples labeled `HighlyKnown`, corresponding to facts that the base model already answers reliably before adaptation. During Phase 2, these examples are mixed with the Unknown facts at a ratio of 1:3 (one Unknown fact for three HighlyKnown facts). We adopt this mixture following the setting of Pletenev et al. (2025), with the goal of coupling factual injection with continued exposure to previously stable knowledge.

This design reflects the stability–plasticity trade-off studied in the main paper. Unknown facts define the new information to be injected, while the accompanying HighlyKnown examples act as a retention signal during training that reduces the risk of catastrophic forgetting during adaptation. Their role is confined to the training mixture: all reported recall is computed on the Unknown subset, since it captures whether the model successfully learns the injected facts.

F.4 TRAINING PROTOCOL BY PHASE

F.4.1 PHASE 1: SKILL EXPERT TRAINING

Phase 1 trains one SVF expert per skill dataset. Concretely, GSM8K defines the mathematical reasoning expert, MBPP defines the coding expert, and AI2-ARC defines the science reasoning expert. In each case, the trainable parameters are the SVF scaling vectors applied to the singular values of the frozen MLP weight matrices, while the singular vectors themselves remain fixed to those of the pre-trained model.

Following Transformer-Squared (Sun et al., 2025), skill specialization is learned with reinforcement learning rather than supervised imitation of reference completions. For each task, the model generates outputs on skill prompts, and a task-specific evaluator assigns a scalar reward based on correctness. The SVF parameters are then updated using this outcome-based signal, so that the learned scaling vector reflects the singular components most strongly engaged by the target skill. The budget is small: each iteration takes a single policy-gradient update over a batch of 256 sampled prompts, and 10 iterations suffice, so the stage fits one vector per layer rather than optimizing the model to convergence.

After training, the best-performing SVF expert is frozen and used as the geometric probe in PALoRA. Its scaling vector defines the per-direction protection weights $w_{l,i} = |z^*_{l,i} - 1| / \max_j |z^*_{l,j} - 1|$, which set how strongly each singular direction is preserved during factual adaptation. No direction is excluded from the protection and none is protected absolutely; the top-k index set $\mathcal{I}_{\text{crit}}$ used in the analysis of Section 3.3 and in the comparison of Section 4.2 is the special case in which these weights are binary.

F.4.2 PHASE 2: KNOWLEDGE INJECTION

In Phase 2, the base model and the Phase 1 SVF expert are kept frozen, and only a LoRA adapter is trained on the factual corpus, starting from the standard initialization $\mathbf{B}_l = \mathbf{0}$ so that the update is zero at the first step. For a given run, the protected subspace is defined by the SVF expert associated with a single target skill, so each Phase 2 experiment corresponds to a particular pairing of a target skill with a number of injected unknown facts.

The training objective combines the standard supervised fine-tuning loss on factual question–answer pairs with the PALoRA orthogonality penalty. The supervised term encourages the adapter to produce the target factual answer, while the orthogonality term charges the LoRA update for its projection onto each input singular direction in proportion to the squared protection weight of that direction, so the cost of using a direction rises with the reliance the expert revealed. Baseline methods are trained on the same deployed expert and use the same factual supervision, and differ only in how, or whether, they constrain the adaptation update.

Answer-only supervision. Each factual example is presented as a question followed by its target answer. The language-model loss is computed only on the answer span, while the question tokens are masked. The model therefore conditions on the full question at training time, but the optimization target is restricted to producing the desired factual completion rather than reconstructing the prompt itself.

F.5 Evaluation protocol

F.5.1 Skill evaluation

After Phase 2, we evaluate the adapted model on the skill benchmark corresponding to the SVF expert used in that run. GSM8K is scored with final-answer accuracy, MBPP with `pass@1`, and AI2-ARC with multiple-choice exact-match accuracy. These metrics quantify *skill preservation*, that is, the extent to which the target reasoning capability survives factual adaptation.

F.5.2 Knowledge evaluation

We additionally evaluate the adapted model on the `Unknown` portion of the factual dataset. At test time, the model is prompted with the question only, and a prediction is counted as correct if the generated answer matches one of the accepted aliases for the target fact. The resulting recall measures the fraction of previously unknown facts that are answered correctly after adaptation. The `HighlyKnown` items enter Phase 2 as part of the training mixture only and are not scored.

F.5.3 Protocol consistency across methods

All methods in our comparison use the same factual datasets, the same skill benchmarks, and the same evaluation rules for a given model and fact-set size. Differences between methods therefore arise from the adaptation mechanism itself rather than from changes in data, prompting, or scoring. This common protocol is essential for isolating the effect of SVF-guided orthogonal protection in PALoRA.

G Baseline Adaptation and Fairness Protocol

To ensure that the baseline comparisons are directly meaningful, we adapt each baseline to the same data interface, optimization setting, and evaluation protocol used for PALoRA, so that performance differences can be attributed to the adaptation principle itself rather than to mismatched experimental conditions.

Common protocol. All methods use the same base checkpoints, the same knowledge-injection datasets, and the same train–validation–test protocol. Crucially, the Phase 1 SVF expert is applied to the base model before the injection stage of *every* method, and before any method-specific preprocessing, so that each spectral baseline decomposes the same expert weights that PALoRA protects and every method is evaluated against the same starting skill level. Retention differences therefore reflect the injection rule alone, and not the presence of the expert in the deployed weights. For knowledge injection, we use the same answer-only supervised objective across methods, masking the prompt tokens and optimizing only the target answer tokens. Across both model families, we match the adapter configuration used in our main experiments, including adapter rank $r = 16$, scaling $\alpha = 32$, dropout 0.05, and batch size 8. The optimization schedule is model-specific: for Llama-3.1-8B-Instruct, we use a learning rate of 2×10^{-4} and train for 10 epochs, whereas for Mistral-7B-Instruct-v0.3, we use a learning rate of 7×10^{-5} and train for 5 epochs. Evaluation is likewise standardized: knowledge performance is measured with the same Unknown split, and skill preservation is assessed with the same task-specific evaluation pipeline for GSM8K, MBPP, and AI2-ARC.

LoRA-only. The LoRA-only baseline uses the same knowledge-injection setting as PALoRA, but removes the structural orthogonality constraint. It is therefore trained as an unconstrained adapter on the same target MLP projections, with the same adapter capacity and the same optimization schedule as PALoRA. This baseline isolates the effect of the constraint itself, while keeping the underlying adaptation budget fixed.

OPLoRA. For OPLoRA, we compute the protected subspace from the deployed expert weights, the same matrices PALoRA protects, and apply the method on the same MLP layers used by PALoRA. Following the original OPLoRA setting, we use a protected subspace size of $k = 128$. We otherwise keep the adapter rank, training objective, optimizer, and training schedule matched to our default

setting. As a result, the comparison focuses on the protection mechanism alone: both methods constrain the update on the same expert weights, but OPLoRA protects the k directions of largest singular value absolutely and leaves the rest free, whereas PALoRA grades the protection of every direction by the deviation of the expert's scaling.

CorDA. We implement CorDA in Knowledge-Preserved Mode and evaluate it under the same downstream knowledge-injection protocol as the other baselines. Its calibration statistics are computed from skill-specific training data using 256 calibration samples, collected after the expert is deployed so that the covariance describes the model the adapter is trained on. The subsequent fine-tuning stage follows the same optimizer family, training budget, and evaluation procedure as in the other LoRA-based comparisons, so the comparison remains aligned at the level of data and protocol.

For evaluation, we convert the trained CorDA adapter into a standard LoRA-form adapter so that it can be assessed through the same inference and reporting pipeline as the other methods. This conversion changes the effective adapter rank from 16 in the original CorDA parameterization to 32 in the converted LoRA form. We report CorDA under this converted setting to preserve a common evaluation interface while retaining its knowledge-preserved initialization mechanism.

PiSSA and MiLoRA. Both methods initialize the adapter from the singular value decomposition of the weights, PiSSA from the principal components and MiLoRA from the minor components, and neither uses skill data to decide which components to occupy: the boundary is a fixed rank in the raw spectrum. We compute the decomposition on the deployed expert weights, on the same MLP projections adapted by PALoRA, and we keep the training objective, optimizer, and adapter shape at our default setting. MiLoRA also keeps the default schedule, whereas PiSSA does not converge under it: at the default learning rate the injection degrades the skill rather than trading against it, so PiSSA is trained at 2×10^{-5} on Llama-3.1-8B-Instruct and 1.75×10^{-5} on Mistral-7B-Instruct-v0.3, in both cases with LoRA dropout 0.0. The comparison therefore isolates the choice of which singular components the update occupies. For evaluation, we convert the trained adapters to the standard LoRA form used by the rest of our pipeline.

Fairness of comparison. Under this protocol, all competing methods are aligned in model family, data, optimization budget, target layers, and evaluation metrics. The reported differences therefore reflect the injection rule alone—unconstrained LoRA, spectral initialization at either end of the spectrum, top-k protection, context-preserving initialization, or graded SVF-weighted protection—rather than artifacts of inconsistent training or evaluation pipelines.

Because the expert is deployed under every method, none of the comparison is confounded by the skill information PALoRA consumes in Phase 1: that information is present in all of the compared models. What separates the methods is where the injected update is allowed to go, and among the baselines only CorDA lets skill data shape that choice, while only OPLoRA constrains the update during training. Appendix I.1 reports the corresponding data and compute budget for every method.

H Seed Variability of the Main Results

Table 11 reports the sample standard deviation over the three seeds of every entry in Table 3, in the same layout. Averaged over the 24 configurations, the standard deviation of skill accuracy is 1.16 for PALoRA, against 1.75 for LoRA, 1.25 for PiSSA, 1.27 for MiLoRA, 1.02 for OPLoRA, and 4.56 for CorDA; that of recall is 2.37, 2.83, 1.67, 2.19, 2.14, and 6.40, respectively.

I Compute, Parameter Count, and Memory Analysis

This section analyzes the practical overhead of PALoRA from first principles. We distinguish three quantities: *trainable parameter count*, *additional stored state*, and *computational overhead*. The discussion is analytical and follows directly from the PALoRA formulation in Sections 3 and 3.2.

Trainable parameters. Let $l = 1, \ldots, L$ index the adapted MLP weight matrices, and let each matrix $\mathbf{W}_{\text{mlp}}^{(l)} \in \mathbb{R}^{d_{\text{out},l} \times d_{\text{in},l}}$ have singular-value dimension $r_l = \min(d_{\text{out},l}, d_{\text{in},l})$. In Phase 1, the

Table 11: Standard deviation over three seeds of the knowledge recall (Rec.) and skill accuracy (Acc.) reported in Table 3, both in percentage points.

		LoRA		PiSSA		MiLoRA		OPLoRA		CorDA		PALoRA	
Skill	**#F**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**	**Rec.**	**Acc.**
Llama-3.1-8B-Instruct													
GSM8K	100	1.0	4.01	2.0	0.89	4.5	0.80	2.5	3.42	1.7	4.91	2.0	0.86
	250	2.1	1.45	2.4	1.78	1.4	0.46	1.6	0.60	4.5	3.72	1.6	0.15
	500	0.3	1.49	1.9	2.81	1.0	1.14	3.4	0.91	3.4	1.43	2.9	0.80
	1000	1.0	2.85	1.8	0.77	1.8	0.35	3.0	1.45	4.3	0.76	0.8	1.01
MBPP	100	1.2	1.75	2.9	0.58	2.5	1.54	2.5	0.58	9.5	4.55	1.0	0.58
	250	4.7	0.58	1.6	2.10	1.1	2.54	3.7	1.01	11.9	5.08	1.2	1.75
	500	4.3	3.09	2.7	1.54	2.1	3.55	1.7	0.58	17.7	1.75	1.8	2.92
	1000	0.8	2.10	0.8	0.58	0.3	0.58	1.6	0.00	10.1	4.04	0.4	4.98
AI2-ARC	100	4.5	1.10	6.5	0.21	2.1	0.59	2.9	2.15	6.1	1.54	3.8	0.57
	250	2.7	0.46	1.1	0.54	1.7	0.22	1.7	0.23	5.9	2.08	2.0	1.25
	500	3.4	0.53	0.3	0.21	3.1	0.65	1.5	0.58	22.6	17.04	0.5	0.28
	1000	2.1	0.89	1.3	0.11	2.1	0.41	2.6	0.46	9.2	12.41	1.5	0.73
Mistral-7B-Instruct-v0.3													
GSM8K	100	2.0	2.23	1.0	1.54	1.5	2.08	0.6	2.47	3.1	3.86	2.1	1.19
	250	2.1	0.62	1.2	0.99	4.5	0.36	2.4	1.05	3.4	3.82	4.1	0.29
	500	6.9	7.18	0.5	2.55	2.6	2.65	3.5	0.57	5.5	1.40	4.7	1.31
	1000	3.1	1.03	0.8	0.79	2.5	3.68	0.6	0.61	4.9	6.37	3.4	0.78
MBPP	100	1.5	1.54	0.6	2.02	2.5	0.58	1.0	1.54	1.7	3.09	1.5	1.17
	250	4.0	2.54	3.1	2.10	3.0	1.54	1.4	1.54	7.3	2.92	3.5	2.92
	500	2.3	2.67	0.5	4.67	1.4	2.33	0.7	1.01	1.4	4.08	0.7	0.58
	1000	5.0	1.01	0.5	1.54	6.4	3.03	2.8	1.75	7.1	1.75	9.7	1.54
AI2-ARC	100	2.6	0.22	2.0	0.72	2.1	0.32	2.3	0.97	1.5	0.84	4.0	0.15
	250	2.4	0.14	2.0	0.45	0.6	0.05	2.9	0.18	3.8	11.92	1.2	0.65
	500	5.0	0.42	1.6	0.27	0.9	0.57	3.1	0.27	5.5	7.70	0.8	0.47
	1000	3.0	2.19	1.1	0.15	0.9	0.56	1.1	0.64	1.4	2.42	1.6	0.98

only trainable variables are the SVF scaling vectors $\{\mathbf{z}_l\}_{l=1}^{L}$, so the number of trainable parameters is

$$P_{\text{SVF}} = \sum_{l=1}^{L} r_l.$$

In Phase 2, PALoRA trains only the LoRA matrices $\{\mathbf{B}_l, \mathbf{A}_l\}_{l=1}^{L}$. For LoRA rank r, the corresponding trainable parameter count is

$$P_{\text{LoRA}} = \sum_{l=1}^{L} r\,(d_{\text{out},l} + d_{\text{in},l}),$$

which is the same parameterization as standard LoRA on the same target modules. Thus, PALoRA does not increase the number of trainable Phase 2 parameters relative to a matched LoRA baseline; its additional complexity comes from the frozen SVF expert and the orthogonality constraint.

Additional stored state. Relative to standard LoRA, PALoRA requires three additional objects. First, it stores the frozen SVF scaling vectors $\mathbf{z}_l^*$ obtained in Phase 1. Second, for each adapted layer l, it stores the protection weights

$$w_{l,i} = |z_{l,i}^* - 1| \,/\, \max_j |z_{l,j}^* - 1|,$$

as defined in the main text. Third, it stores the singular vectors of the protected side, since the Phase 2 orthogonality loss is computed by projecting the LoRA update onto every singular direction with these weights. Since Eq. 3 protects the input side, these are the right singular vectors $\mathbf{V}_l \in \mathbb{R}^{d_{\text{in},l} \times r_l}$, and the dominant additional storage is therefore linear in the size of the stored singular bases, namely on the order of $\sum_l r_l\, d_{\text{in},l}$; an output-side variant would store $\sum_l r_l\, d_{\text{out},l}$ instead, and a bilateral one the sum of the two. By contrast, the storage of $\mathbf{z}_l^*$ and $\mathbf{w}_l$ is comparatively small, and neither the singular vectors nor the weights are required at inference time.

Computational overhead. PALoRA introduces overhead in two places. The first is an offline cost incurred before Phase 2, consisting of the singular value decomposition of the adapted MLP weights

together with the training of the SVF expert in Phase 1. This cost is paid once per skill expert and is independent of the subsequent number of Phase 2 knowledge-injection runs that reuse the same frozen probe.

The second source of overhead appears during Phase 2 training. In addition to the standard supervised fine-tuning objective, PALoRA evaluates the orthogonality regularizer

$$\mathcal{L}_{\text{ortho},l} = \sum_{i=1}^{r_l} w_{l,i}^2 \, \|\mathbf{B}_l \mathbf{A}_l \mathbf{v}_{l,i}\|_2^2 = \|\mathbf{B}_l \mathbf{A}_l \mathbf{V}_l \, \mathrm{diag}(\mathbf{w}_l)\|_F^2$$

for each adapted layer. The sum runs over the full spectrum, but the factored form on the right is evaluated on the LoRA factors rather than on $\Delta\mathbf{W}_l = \mathbf{B}_l\mathbf{A}_l$, which keeps the per-step cost at $O(r\,d^2)$ per layer instead of the $O(d^3)$ that forming $\Delta\mathbf{W}_l\mathbf{V}_l$ would require. The resulting overhead therefore depends on the dimensions of the adapted MLP layers and on the adapter rank, rather than on any increase in trainable parameter count.

Inference-time considerations. The orthogonality term is used only during optimization in Phase 2 and does not introduce an additional objective at inference time. After training, deployment uses the adapted model parameters in the same way as other LoRA-style methods. Consequently, the practical inference footprint is determined by the final adapted weights rather than by the training-time regularization procedure.

Interpretation. This decomposition clarifies the cost profile of PALoRA. Phase 1 adds a one-time preprocessing stage, which in our protocol is shared with every baseline, while Phase 2 preserves the same LoRA parameterization as a matched baseline and adds only a structured regularization term tied to the protection weights. The main trade-off is therefore not an increase in trainable capacity, but the extra storage and computation required to score and preserve the SVF-weighted singular directions.

I.1 Measured Data and Compute Budget

Table 12 reports what each method estimates its subspace from, together with the measured wall-clock cost of each stage on Llama-3.1-8B, using the GSM8K expert, 1000 unknown facts, and 10 epochs, on an NVIDIA A100-SXM4-80GB. The injection column reports the training time, excluding model loading and evaluation.

Phase 1 is now common to the comparison, since every method is deployed on the same SVF expert, so we report it once rather than charging it to PALoRA: its 353 s cover the RL training of the expert and the decomposition it is written through, and this cost is paid once per skill and amortized over every subsequent injection run that reuses it. The preprocessing column holds what each method additionally needs before injection: the singular value decomposition of the expert weights for PiSSA and MiLoRA, the extraction of the protected top-k for OPLoRA, and covariance estimation for CorDA. PALoRA reuses the decomposition produced in Phase 1 and adds nothing here, paying instead in the injection stage.

Table 12: Data and wall-clock budget per method (Llama-3.1-8B, GSM8K, 1000 unknown facts, 10 epochs), on an A100-SXM4-80GB. The expert column is the same one-time cost for every row.

Method	Subspace estimated from	Expert (s)	Prep. (s)	Inject. (s)	Total (s)
LoRA	—	353	—	631	984
PiSSA	spectrum of $\mathbf{W}$	353	208	582	1143
MiLoRA	spectrum of $\mathbf{W}$	353	212	609	1174
OPLoRA	spectrum of $\mathbf{W}$	353	21	610	984
CorDA	256 calibration samples	353	389	594	1336
PALoRA	SVF deviation $\lvert\mathbf{z}^* - \mathbf{1}\rvert$	353	—	800	1153

Once the shared expert is accounted for, PALoRA's overhead is confined to the injection stage, which runs at 1.3–1.4$\times$ that of the baselines, with the additional cost arising from evaluating the orthogonality regularizer rather than from a larger trainable parameterization. End to end, a single

PALoRA run costs $1.2\times$ LoRA and $0.9\times$ CorDA, whose covariance estimation is the most expensive preprocessing in the comparison.