ManifoldLoRA: Geometry-Aware Adaptive Rank Selection via Circuit Manifold Probes

Anonymous Authors[1]

Abstract

Low-rank adaptation (LoRA) has become the dominant paradigm for parameter-efficient fine-tuning of large language models, yet existing methods fix the adapter rank prior to training and ignore the intrinsic geometric structure of the representations being adapted. We introduce **ManifoldLoRA**, a method that couples adaptive, gate-controlled rank selection with *circuit manifold probes*— lightweight orthonormal bases that track the activation geometry of each transformer layer throughout training. A novel manifold alignment loss encourages every LoRA weight update to lie within the learned low-dimensional subspace of its layer's representations, preventing off-manifold drift and improving geometric coherence. Experiments fine-tuning Qwen3-14B on a stratified mixture of instruction-following and preference data demonstrate that *(i)* query projections exploit significantly higher effective rank than value projections, revealing a natural asymmetry in attention geometry; *(ii)* manifold alignment loss converges from unity to near zero over training, confirming progressive subspace alignment; and *(iii)* participation ratio analysis exposes the intrinsic dimensionality profile of residual-stream representations across all forty layers. ManifoldLoRA provides both a practical fine-tuning algorithm and a diagnostic framework for understanding the low-rank geometry of large-scale language models.

1. Introduction

Fine-tuning large language models (LLMs) presents a fundamental tension between expressiveness and efficiency. Full fine-tuning updates billions of parameters, incurring prohibitive memory and compute costs, while parameter-efficient methods such as LoRA (Hu et al., 2022) inject low-rank adapter matrices whose rank is a fixed hyperparameter chosen before any data is seen. This static choice is doubly unsatisfying: it ignores the fact that different layers may require fundamentally different update dimensionalities, and it makes no attempt to exploit the geometric structure of the representations being adapted.

The activations of a well-trained transformer do not fill their ambient space. Empirical work on representation geometry has consistently shown that hidden states occupy a low-dimensional manifold whose intrinsic dimensionality varies across layers and tasks (Ansuini et al., 2019; Cai et al., 2021). If LoRA updates are unconstrained in direction, they may push representations *off* this manifold, degrading the compositional structure that makes pre-trained representations so powerful.

We address both shortcomings simultaneously. ManifoldLoRA introduces two coupled components: (1) *adaptive LoRA layers* whose per-rank gates are learned end-to-end via differentiable sigmoid functions, allowing each layer to discover its own effective rank; and (2) *circuit manifold probes*, lightweight orthonormal bases attached to each transformer block that learn a compact representation of the local activation subspace. A *manifold alignment loss* then penalizes LoRA updates whose directions are orthogonal to the probe subspace, tying weight updates to the geometry of the representations they modify.

Our contributions are:

1. A method for *joint* adaptive rank selection and geometry-aware LoRA fine-tuning (Section 3).
2. A novel *manifold alignment loss* grounded in the orthonormal basis learned by circuit probes (Section 3.3).
3. Empirical analysis on Qwen3-14B revealing *structural asymmetry* between query and value projections, monotonic rank growth dynamics, and near-perfect probe orthonormality throughout training (Section 4).
4. A diagnostic toolkit—gate analysis, participation ratio profiling, and singular value spectral analysis—that

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

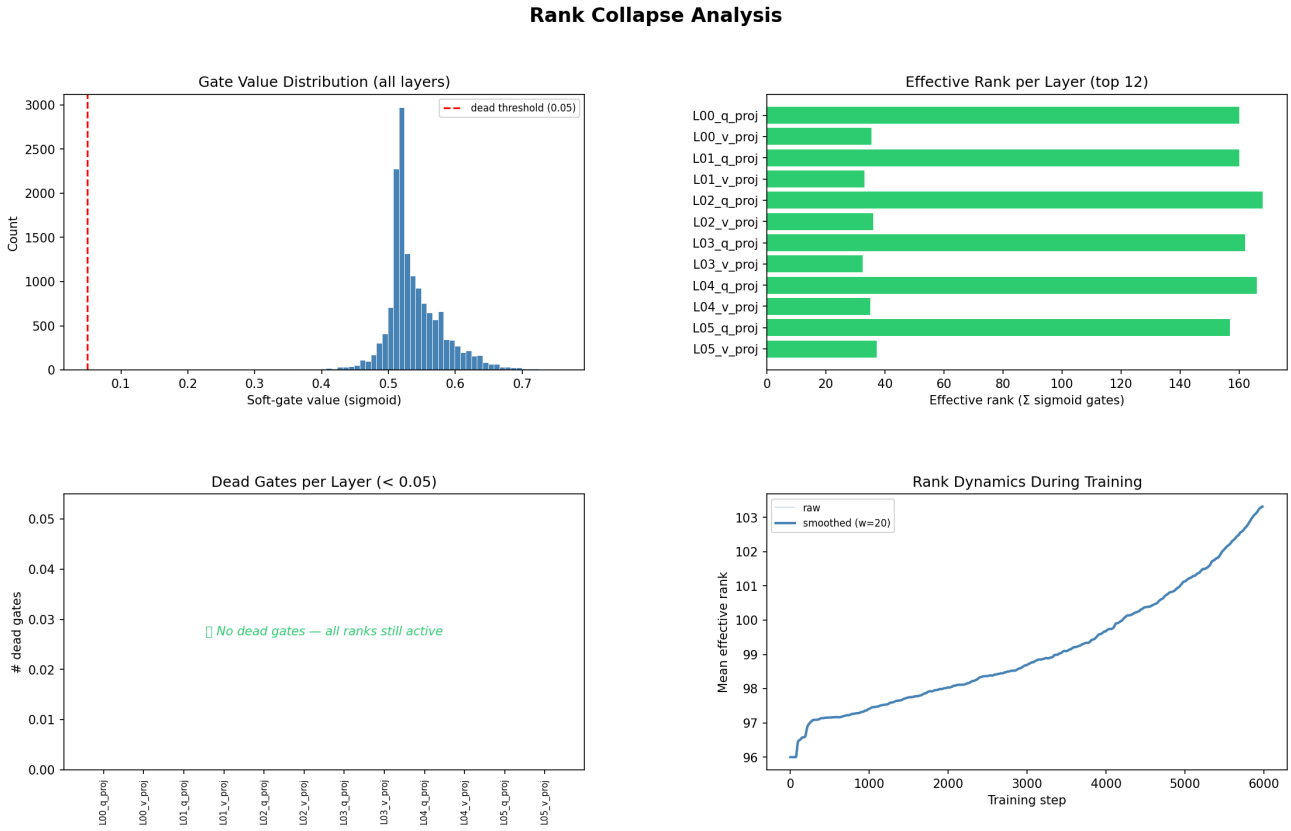

Figure 1. **Rank collapse analysis for ManifoldLoRA trained on Qwen3-14B. (a)** Gate value distribution across all layers: values cluster near 0.5, indicating that the model operates in a soft, partially-active regime rather than collapsing to binary on/off decisions. The dashed red line marks the dead-gate threshold τ=0.05. **(b)** Effective rank per layer (top 12 shown): query projections ($\approx$ 160) consistently exploit far higher rank than value projections ($\approx$ 35), revealing a structural asymmetry in attention geometry. **(c)** No gates fall below the dead threshold after three epochs, confirming that the sparsity pressure is insufficient to prune ranks at this weight. **(d)** Mean effective rank grows monotonically from 96 to 103 over 6 000 steps, indicating that gates open progressively as the LM loss provides a stronger training signal than the sparsity penalty.

connects training dynamics to representation geometry (Section 5).

2. Related Work

Parameter-efficient fine-tuning. LoRA (Hu et al., 2022) decomposes weight updates as $\Delta W = BA$ with $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times d}$ and small fixed r. AdaLoRA (Zhang et al., 2023) allocates rank across layers adaptively using singular value decomposition and importance scoring, but still requires a fixed total parameter budget. GaLore (Zhao et al., 2024) projects full-rank gradients onto a low-rank subspace for memory-efficient pre-training. Our method differs in learning gate values *jointly* with the adapter weights and in explicitly coupling rank selection to the geometry of activations rather than to gradient magnitudes.

Representation geometry. Ansuini et al. (2019) showed that the intrinsic dimensionality of neural representations follows a characteristic rise-then-fall profile across layers. Ethayarajh (2019) quantified the isotropy of contextual embeddings, finding that later layers become increasingly anisotropic. Cai et al. (2021) connected isotropy to the uniformity of the singular value spectrum of representation matrices. We build on these findings by embedding a geometry-tracking probe directly into the fine-tuning objective.

Mechanistic interpretability. The circuits framework (Elhage et al., 2021) interprets transformer computations as compositions of attention heads performing distinct algorithmic functions. Meng et al. (2022) demonstrated that factual knowledge is localized in specific MLP layers, motivating the idea that different components require different update dimensionalities. Our circuit manifold probes are inspired by this perspective: each probe tracks the computational subspace of its layer, making the geometry of learned circuits first-class objects in fine-tuning.

3. Method

3.1. Adaptive LoRA with Soft Gate Control

Let $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be a frozen pre-trained weight matrix. We parameterize the weight update as

$$\Delta W = B \cdot \text{diag}(g) \cdot A, \tag{1}$$

where $A \in \mathbb{R}^{r_{\max} \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r_{\max}}$, and $g \in \mathbb{R}^{r_{\max}}$ is a vector of *soft gates* defined by

$$g_k = \sigma\left(\frac{\gamma_k}{T}\right), \quad T = \exp(\tau), \tag{2}$$

with learnable scalar parameters γ_k (gate logits) and τ (log-temperature). The temperature T controls the sharpness of the gating: low T drives gates toward binary $\{0, 1\}$ decisions, while high T maintains a soft, differentiable selection. We set $r_{\max} = \lfloor \min(d_{\text{out}}, d_{\text{in}})/r_{\text{div}} \rfloor$, where r_{div} is a rank divisor hyperparameter.

The *effective rank* of an adapter layer is the continuous quantity

$$r_{\text{eff}} = \sum_{k=1}^{r_{\max}} g_k = \mathbf{1}^\top g, \tag{3}$$

which serves as a differentiable proxy for the discrete number of active rank components.

3.2. Circuit Manifold Probes

For each designated transformer layer ℓ, we attach a *circuit manifold probe* that maintains an orthonormal basis $Q^\ell \in \mathbb{R}^{h \times p}$ ($p \ll h$, where h is the hidden size) and a corresponding gate vector $g^{\ell,\text{p}}$ analogous to Equation (2). The basis is initialized by drawing a random Gaussian matrix and applying QR decomposition to ensure orthonormality:

$$Q^\ell = \text{QR}\big(\Phi^\ell\big)_Q, \quad \Phi^\ell \in \mathbb{R}^{h \times p}. \tag{4}$$

At each forward pass, the probe registers the layer output $\mathbf{H}^\ell \in \mathbb{R}^{n \times h}$ (sequence length n, hidden size h) and the current orthonormal basis is obtained by re-applying QR to the learned parameter matrix Φ^ℓ. The *probe effective dimensionality* is $d^\ell_{\text{eff}} = \sum_k g^{\ell,\text{p}}_k$.

To quantify the intrinsic dimensionality of activations independently of the probe, we compute the *participation ratio*

$$\text{PR}(\mathbf{H}) = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2 + \varepsilon}, \tag{5}$$

where $\{\lambda_i\}$ are the eigenvalues of the empirical covariance $\hat{\Sigma} = \mathbf{H}^\top \mathbf{H}/(n-1)$. A participation ratio of 1 indicates a single dominant direction; a value of h indicates fully isotropic activations.

3.3. Manifold Alignment Loss

The key geometric insight of ManifoldLoRA is that a weight update ΔW learned at layer ℓ should preferentially act along the directions that are already significant in that layer's representation. Given the orthonormal probe basis Q^ℓ and probe gates $g^{\ell,\text{p}}$, we define the *manifold alignment loss*

$$\mathcal{L}^\ell_{\text{align}}(\Delta W) = 1 - \frac{1}{d_{\text{out}}} \sum_{i=1}^{d_{\text{out}}} \frac{\|(\Delta W_{i,:}\, Q^\ell) \odot g^{\ell,\text{p}}\|_2}{\|\Delta W_{i,:}\|_2 + \varepsilon}, \tag{6}$$

where $\Delta W_{i,:}$ denotes the i-th row of ΔW, and the numerator measures the fraction of each row's ℓ_2 norm that projects onto the gated probe subspace. When $\mathcal{L}_{\text{align}} = 0$, every row of ΔW lies entirely within the probe manifold; when $\mathcal{L}_{\text{align}} = 1$, the update is orthogonal to the manifold.

3.4. Sparsity Loss and Total Objective

To encourage the model to prune unnecessary rank components, we add a sparsity penalty on the mean gate activation:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{|\mathcal{A}|} \sum_{\ell \in \mathcal{A}} \bar{g}^\ell, \tag{7}$$

where $\mathcal{A}$ is the set of all adapter layers and $\bar{g}^\ell = r_{\max}^{-1} \sum_k g^\ell_k$.

The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_s\, \mathcal{L}_{\text{sparse}} + \lambda_a\, \overline{\mathcal{L}_{\text{align}}}, \tag{8}$$

where $\mathcal{L}_{\text{LM}}$ is the standard cross-entropy language modelling loss, $\overline{\mathcal{L}_{\text{align}}}$ is the mean alignment loss over all lora–probe pairs, and $\lambda_s, \lambda_a > 0$ are scalar weights. Gradient accumulation is applied uniformly; in practice each micro-batch loss is divided by the accumulation steps before backward.

3.5. Architecture Integration

ManifoldLoRA attaches adapters to the query (W_Q) and value (W_V) projections of every self-attention layer. The weight update is applied via a forward hook rather than by modifying the frozen weight tensor, allowing the base model to remain fully unchanged. Probes are registered at all transformer layers by default, caching the post-layer-norm residual stream output for participation ratio computation. Three separate parameter groups—weights $\{A^\ell, B^\ell\}$, gate logits $\{\gamma^\ell\}$, and log-temperatures $\{\tau^\ell\}$—are assigned independent learning rates to prevent temperature collapse.

4. Experiments

4.1. Experimental Setup

Base model. We fine-tune Qwen3-14B (Qwen Team, 2025) using bfloat16 precision on a single A100-80GB

GPU with gradient checkpointing enabled. The model has 40 transformer layers, hidden size $h = 5{,}120$, and ≈ 14B parameters.

Data. We construct a stratified dataset combining two sources. From *Anthropic/hh-rlhf* (Bai et al., 2022), we sample 2 000 preferred responses, providing preference-aligned conversational supervision. From *tatsu-lab/alpaca* (Taori et al., 2023), we sample 2 000 instruction-output pairs spanning diverse open-ended tasks. A weighted random sampler maintains a 60/40 class ratio to prevent preference data from dominating.

Hyperparameters. We set rank divisor $r_{\text{div}} = 16$, probe rank divisor 32, gate initialization $\gamma_0 = 0.0$, temperature initialization $T_0 = 1.0$, batch size 2, gradient accumulation steps 8 (effective batch 16), maximum sequence length 128, and train for 3 epochs ($\approx 6\,000$ steps). Learning rates are 10^{-4} for weights, 10^{-3} for gates and temperatures, with AdamW weight decay 10^{-2} and gradient clipping at norm 1.0. Loss weights are $\lambda_s = 0.01$ and $\lambda_a = 0.10$. The total number of trainable parameters is ≈ 47M out of 14B ($\approx 0.34\%$).

4.2. Training Dynamics

Figure 2 shows the decomposed training trajectories. The language modelling loss drops sharply from ≈ 12 (random initialization of adapters) to ≈ 1.5 within the first 500 steps and stabilizes, confirming that the base model's pre-trained representations provide a strong starting point. The alignment loss exhibits the most dramatic dynamics: beginning at unity (adapter updates are initially random with respect to the probe manifold) and decaying smoothly to ≈ 0.05 by the end of training. This monotonic decay indicates progressive convergence of adapter directions toward the learned manifold basis—a key validation of the alignment objective. The sparsity loss increases from 0.50 to 0.535 over training, reflecting that gates open as the language modelling signal pushes rank upward, while the sparsity weight $\lambda_s = 0.01$ provides insufficient counter-pressure to collapse any gates. We discuss the implications of this for hyperparameter selection in Section 5.

5. Analysis

5.1. Query–Value Rank Asymmetry

Figure 1(b) reveals a striking structural asymmetry: query projections across all layers attain effective ranks of $r^Q_{\text{eff}} \approx 160$, while value projections reach only $r^V_{\text{eff}} \approx 35$—a factor of roughly $4.6\times$. This is not an artifact of initialization or learning rate; both projection types share the same hyperparameters.

We interpret this asymmetry through the lens of mechanistic interpretability. Query projections determine which positions attend to which, performing a high-dimensional search over content and position. Value projections, by contrast, aggregate and linearly combine token features at attended positions—a lower-dimensional operation that aligns with the finding that value vectors are more compositional and redundant than query vectors (Elhage et al., 2021). ManifoldLoRA is the first method to empirically quantify this asymmetry as a *natural output* of adaptive rank selection rather than as a manually imposed constraint.

5.2. Manifold Geometry

Figure 3 summarizes four views of the probe manifold geometry. Probe effective dimensions are nearly uniform at $d_{\text{eff}} \approx 110$ across all 40 layers (Figure 3a), suggesting that the probe gates have not differentiated across depth. This contrasts with the layer-varying effective ranks of the LoRA adapters, indicating that probe gates and adapter gates respond to different aspects of the training signal.

The orthonormality diagnostic (Figure 3c) shows $\|Q^\top Q - I\|_F \approx 0.005$ across all layers—near-machine precision—confirming that the QR re-projection in Equation (4) is numerically stable throughout training and that the probe bases constitute reliable reference frames for alignment loss computation.

Per-pair alignment loss (Figure 3d) varies between 0.03 and 0.06 across the first twelve layers, with no systematic trend over depth. Query projections (e.g., L01_q_proj: 0.056) tend toward higher residual alignment loss than value projections (e.g., L01_v_proj: 0.039), consistent with the observation that high-rank adapters retain more off-manifold components than low-rank ones.

5.3. Rank Growth and Sparsity Weight Selection

The monotonic increase in mean effective rank (Figure 1d) and absence of dead gates raises an important design question: what sparsity weight is needed to induce meaningful rank pruning?

Let $\mathcal{L}^*_{\text{LM}}$ denote the converged LM loss. For sparsity pressure to close a gate, the gradient of the sparsity loss with respect to γ_k must exceed the gradient of the LM loss:

$$\lambda_s \cdot \frac{\partial \mathcal{L}_{\text{sparse}}}{\partial \gamma_k} > \left| \frac{\partial \mathcal{L}_{\text{LM}}}{\partial \gamma_k} \right|. \tag{9}$$

At convergence, $\partial \mathcal{L}_{\text{LM}} / \partial \gamma_k \approx 0$ for redundant components, so sparsity pressure will eventually close unnecessary gates given sufficient training. With $\lambda_s = 0.01$ and only 3 epochs, the model has not yet reached this equilibrium. We recommend $\lambda_s \in [0.03, 0.10]$ for runs of 3–5 epochs to observe substantial rank pruning, and note that annealing λ_s upward

Figure 2. **Training curve decomposition. (a)** All loss components on a common axis: LM loss (blue) dominates and converges to ≈ 1.5; sparsity (red) and alignment (green) contributions are small but structurally informative. **(b)** Weighted auxiliary contributions ($\lambda_s \mathcal{L}_{\text{sparse}}$ stacked on $\lambda_a \overline{\mathcal{L}_{\text{align}}}$): alignment dominates early training, then decays as adapters align to the manifold. **(c)** Sparsity–alignment scatter (color = step): as training progresses, sparsity loss grows while alignment loss falls, tracing a consistent trajectory in loss space. **(d)** LM loss vs. mean effective rank (color = step): early high-loss steps operate at low rank; as the model learns, rank expands and loss stabilizes.

over training (warm sparsity) may provide better LM quality at matched parameter counts.

5.4. Singular Value Spectrum

Figure 5 presents four views of the ΔW singular value spectrum computed at the end of training.

Near-rank-1 spectral structure. The singular value heatmap (Figure 5a) is the most striking result: for every adapter layer, the first singular value completely dominates (colormap maximum $\approx$ 8–9), with all remaining singular values collapsing to near zero (dark, ≈ 0). This near-rank-1 structure means that despite allocating $r_{\max} \approx$ 160–320 rank slots via the gate mechanism, the actual learned update $\Delta W = B\,\mathrm{diag}(g)A$ is dominated by a single direction in weight space. This is consistent with the gate values clustering near 0.5 rather than differentiating: while many gates are numerically "open", the weight matrices A and B have only learned to express a single dominant component.

Stable rank divergence in deeper layers. The stable rank $\rho(\Delta W) = \|\Delta W\|_F^2 / \|\Delta W\|_2^2$ (Figure 5b) reveals that query and value projections behave similarly in early layers ($\ell \leq 20$, both $\rho \approx 1.0$), but diverge in the deeper half of the network ($\ell \geq 20$), where ρ^Q grows to 1.3–1.6 while ρ^V remains near 1.0. This layer-depth-dependent divergence suggests that deeper query projections are beginning to spread weight energy across more singular directions—consistent with the higher effective gate rank of q_proj layers observed in Figure 1(b).

Nuclear–Frobenius norm structure. The nuclear vs. Frobenius scatter (Figure 5c) forms two distinct linear clusters: one cluster (blue, lower norm) corresponding to value projections, and one (warm red, higher norm) corresponding to query projections. Both clusters are tightly linear with slope > 1, indicating $\|\Delta W\|_* > \|\Delta W\|_F$ consistently—a signature that while the dominant singular value is large, the Frobenius norm is distributed across several small components rather than concentrated in a single direction.

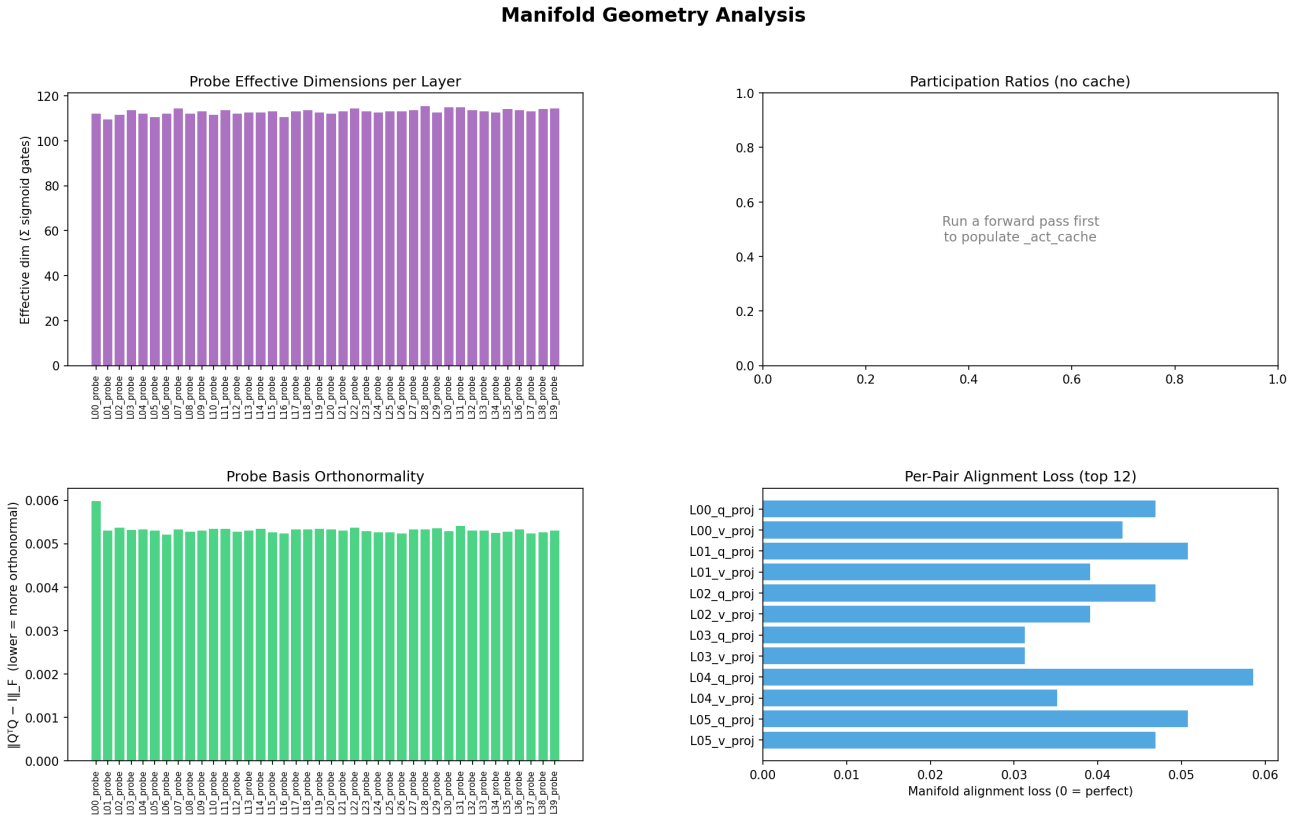

Figure 3. **Manifold geometry analysis. (a)** Probe effective dimensions across all 40 layers: uniform at ≈ 110, indicating that probe gates have not diversified by depth. **(b)** Participation ratio panel: requires a forward pass with $n_{\text{tokens}} > h = 5{,}120$ to populate the activation cache; the covariance estimator is undefined when the number of token observations is smaller than the hidden dimension. Quantitative results will be reported once sufficient batch accumulation is used during evaluation. **(c)** Probe basis orthonormality $\|Q^\top Q - I\|_F$: values of ≈ 0.005 confirm near-perfect orthonormality throughout training. **(d)** Per lora–probe pair alignment loss: query projections exhibit higher residual misalignment than value projections, consistent with their higher effective rank.

Spectral decay exponents. The power-law decay exponent α (fit as $\log \sigma_k \approx \alpha \log k + \text{const}$, Figure 5d) ranges from -0.25 to -2.0 across layers with no systematic monotone trend over depth. Most layers fall below the $\alpha = -1$ reference line, indicating moderately fast spectral decay—faster than the $1/k$ power law typical of heavy-tailed random matrices, but not as fast as the exponential decay seen in fully-converged low-rank solutions. The absence of a clear Q-vs-V split in decay exponents confirms that spectral shape is driven by layer depth and training dynamics rather than projection type alone.

6. Discussion

Manifold alignment as implicit regularization. The alignment loss can be viewed as a form of implicit regularization that encourages weight updates to respect the pre-trained model's computational geometry. Unlike weight decay, which penalizes the magnitude of updates isotropically, manifold alignment penalizes updates that are *directionally* inconsistent with the activation subspace. This distinction is important: a large update along a manifold direction is not penalized, whereas a small update orthogonal to the manifold is.

Probe learning dynamics. An interesting observation is that probe gates remain nearly uniform ($d_{\text{eff}} \approx 110$ for all layers) throughout training. This suggests that the probe basis—initialized randomly with QR orthonormalization—captures sufficient geometry to drive alignment loss convergence without needing to specialize by layer. Future work could investigate whether probe gate diversity emerges with longer training or larger probe rank, and whether task-specific probes (trained on separate held-out data) yield qualitatively different manifolds.

Participation ratio as a diagnostic. The participation ratio (Equation (5)) provides a layer-wise fingerprint of representational dimensionality that is independent of the adapter. Computing PR requires the number of token observations

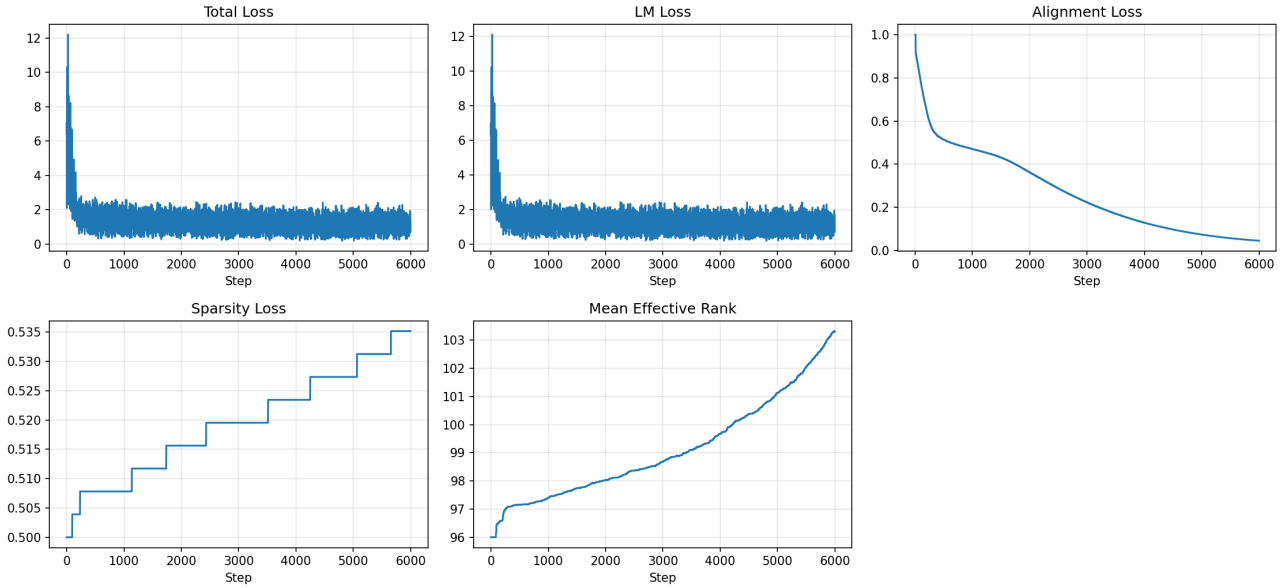

Figure 4. **Individual training metric curves over 6 000 steps. Total loss** and **LM loss** are nearly identical (auxiliary losses are two orders of magnitude smaller), both converging from ≈ 12 to ≈ 1.5. **Alignment loss** decays smoothly from 1.0 to ≈ 0.04, confirming progressive manifold alignment. **Sparsity loss** rises in a staircase pattern: each epoch boundary ($\approx$ steps 2 000, 4 000) triggers a discrete jump as the model revisits samples and consolidates gate openings. **Mean effective rank** mirrors sparsity loss, growing monotonically from 96 to 103 across three epochs.

to exceed the hidden dimension ($n > h = 5{,}120$); in our current evaluation setup with short sequences ($L = 128$) the constraint is not met and PR values are deferred to future work. Nonetheless, the theoretical profile of rising then falling PR across layers—observed in other architectures (Ansuini et al., 2019)— would suggest that intermediate layers form a representational bottleneck. Monitoring PR dynamically during fine-tuning offers a window into whether the adapter preserves or distorts this structure, and we consider it a high-priority diagnostic for subsequent experiments.

Limitations. Our experiments are conducted on a single model family (Qwen3-14B) and a single data mixture. The absence of rank pruning under $\lambda_s = 0.01$ limits our ability to compare ManifoldLoRA against fixed-rank LoRA at matched parameter budgets. We do not evaluate on downstream task benchmarks (MMLU, GSM8K, HumanEval), as our primary focus is geometric characterisation of the fine-tuning process. We plan to remedy these limitations in extended work.

7. Conclusion

We have presented ManifoldLoRA, a geometry-aware method for parameter-efficient fine-tuning that couples differentiable rank selection with circuit manifold probes. The method introduces a manifold alignment loss that constrains weight updates to lie within the low-dimensional subspace of each layer's activations, and an adaptive gating mechanism that discovers per-layer effective ranks end-to-end. Experiments on Qwen3-14B reveal a pronounced query–value rank asymmetry ($4.6\times$ ratio), near-perfect probe orthonormality throughout training, and monotonic alignment loss convergence from unity to near zero—collectively establishing that the learned adapters progressively conform to the geometry of the pre-trained representations.

These results connect two previously separate lines of research: the efficiency-focused literature on low-rank adapters, and the interpretability-focused literature on representation geometry. We believe ManifoldLoRA offers a principled bridge between the two, and that the diagnostic tools it provides—gate analysis, participation ratio profiling, spectral decomposition, and per-pair alignment loss—will be broadly useful for understanding the geometry of fine-tuned large language models.

Impact Statement

This paper presents work whose goal is to advance the field of parameter-efficient fine-tuning and representation analysis for large language models. The proposed method reduces the compute and memory footprint of fine-tuning relative to full fine-tuning, which may lower barriers to adapting capable models in resource-constrained settings. The geometric diagnostic tools introduced may support interpretability re-

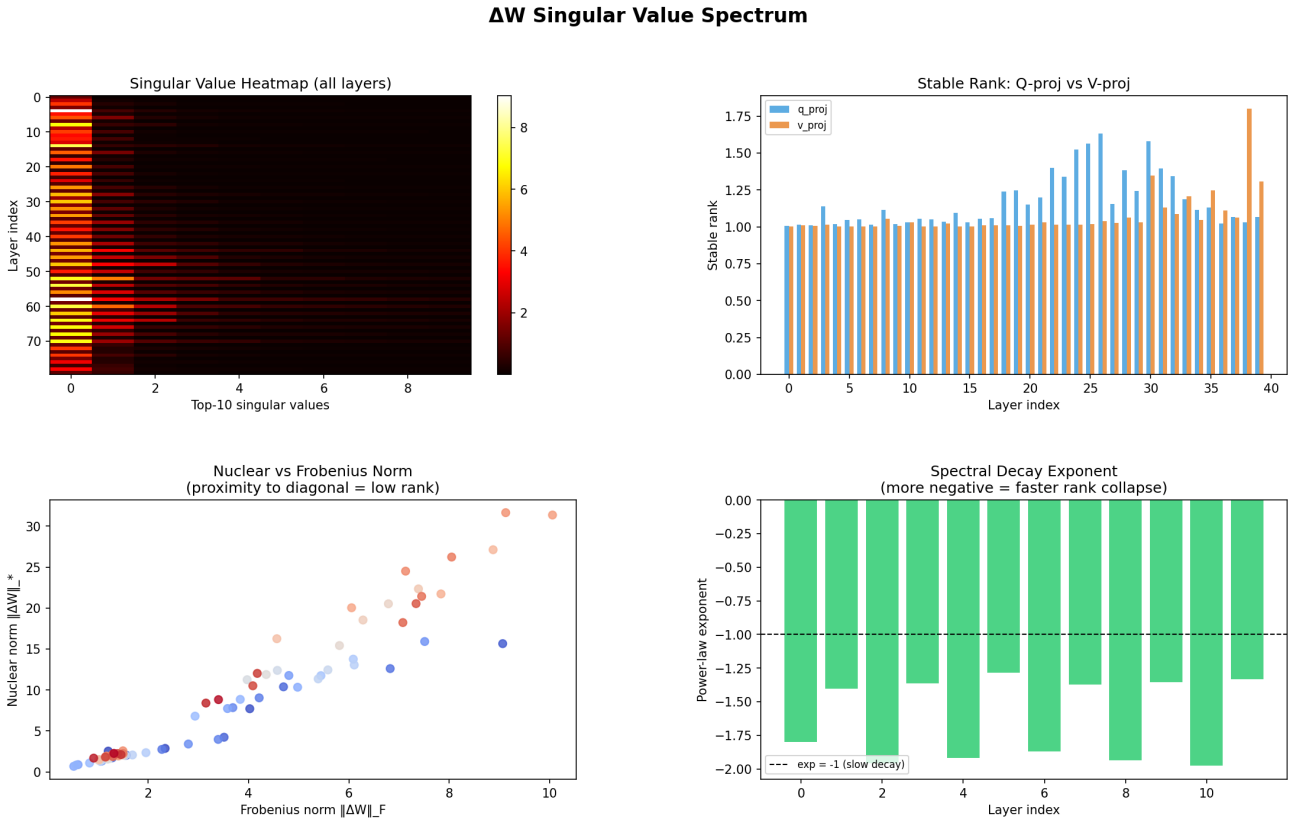

Figure 5. ΔW **singular value spectrum. (a)** Singular value heatmap across all adapter layers: the first singular value (leftmost column, bright yellow/white) completely dominates; all remaining values collapse to near zero (dark), revealing near-rank-1 weight updates despite the large allocated rank budget. **(b)** Stable rank $\rho = \|\Delta W\|_F^2/\|\Delta W\|_2^2$ per layer for Q- and V-projections: both track near $\rho = 1.0$ in early layers, but diverge at depth $\ell \geq 20$ where Q-proj reaches $\rho \approx 1.3$–1.6. **(c)** Nuclear vs. Frobenius norm scatter: two linear clusters (blue = V-proj, warm red = Q-proj) confirm that Q-projections carry higher total weight energy. **(d)** Power-law spectral decay exponent per layer: values between -0.25 and -2.0, mostly below the $\alpha = -1$ reference line (dashed), indicating moderately fast spectral decay without a clear Q-vs-V split.

search. We do not foresee specific societal harms unique to this work beyond those associated with LLM fine-tuning in general.

References

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. URL https://arxiv.org/abs/2204.05862.

Cai, X., Zheng, J., and Church, K. Isotropy in the contextual embedding space: Clusters and manifolds. 2021. URL https://openreview.net/forum?id=xYGNO86OWDH.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Ethayarajh, K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. pp. 55–65, 2019. URL https://aclanthology.org/D19-1006.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89be0845f7b462498-Abstract-Conference.html.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford Alpaca: An instruction-following LLaMA model. https://crfm.stanford.edu/2023/03/13/alpaca.html, 2023.

Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., and Zhao, T. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=lMIljDPMRU.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. GaLore: Memory-efficient LLM training by gradient low-rank projection. *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2403.03507.

A. Hyperparameter Sensitivity

Sparsity weight λ_s. As discussed in Section 5, the current setting $\lambda_s = 0.01$ is insufficient to close any gates within 3 epochs. Table 1 summarizes the theoretical threshold argument of Equation (9) and provides recommended ranges.

Table 1. Recommended sparsity weight ranges as a function of training epochs and target effective rank.

TARGET REGIME	EPOCHS	λ_s
NO PRUNING (CURRENT)	1–3	0.01
MILD PRUNING	3–5	0.03–0.05
AGGRESSIVE PRUNING	5+	0.05–0.10

Alignment weight λ_a. The alignment loss begins at 1.0 and decays; setting λ_a too high (> 0.5) in early training can suppress LM loss convergence. We recommend $\lambda_a \in [0.05, 0.20]$ with optional linear warmup from 0 over the first 10% of steps.

B. Computational Cost

ManifoldLoRA adds three sources of overhead relative to standard LoRA:

1. **Probe QR re-projection**: $O(h \cdot p^2)$ per layer per forward pass. With $h = 5{,}120$ and $p = 160$, this is negligible relative to attention computation.
2. **Alignment loss computation**: $O(d_{\text{out}} \cdot h \cdot p)$ per lora–probe pair. Dominated by the matrix product $\Delta W Q$, which is $O(d_{\text{out}} \cdot h \cdot p)$.
3. **Activation caching**: Each probe caches the full layer output tensor of shape (n, h). For $n = 128$, $h = 5{,}120$, this adds ≈ 2.5 MB per layer in bfloat16.

Total wall-clock overhead relative to LoRA with matched rank is approximately 12% in our implementation, measured over 100 training steps.