The Intrinsic Low-Rank Geometry of LLM Alignment: Universality and Scaling Laws of the Gradient

Changguo Fang[1,2], Wang Xi[2,3], Quan Shi[4],Yucheng Fang[5], Zenghui Ding[2,*]
[1]Anhui Agricultural University, China
[2]Intelligence Institute of Machine, Hefei Institutes of Physical Science, Chinese Academy of Sciences, China
[3]University of Science and Technology of China, China [4]Changzhou University, China [5]Anhui University, China
Emails: changguofang@stu.ahau.edu.cn, xw_cs@mail.ustc.edu.cn, s23040820006@smail.cczu.edu.cn,
w225302038@stu.ahu.edu.cn, dingzenghui@iim.ac.cn

***Abstract*—Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA achieve strong alignment quality while updating only a tiny fraction of parameters, yet the *geometry* of the underlying alignment updates is still not well understood. This paper studies the unconstrained *full-parameter* alignment gradients and tests the hypothesis that their effective dimension is intrinsically small. Concretely, for each weight matrix gradient G, we quantify its effective rank by the *stable rank* $r_s(G) = \|G\|_F^2/\|G\|_2^2$, which is robust to small singular-value noise. We propose a unified perspective (information bottleneck, feature learning beyond lazy regimes, and a task-manifold view) that predicts *low* r_s for alignment gradients, and we empirically validate this Low-Rank Gradient Hypothesis via large-scale spectral measurements across multiple mainstream LLM families (e.g., Llama, Qwen, Mistral) and scales (7B–70B). Across models and tasks, we consistently observe that alignment gradients admit accurate low-rank approximations and that their stable rank decreases with model size, following an approximate scaling law $r_s \propto N^{-\gamma}$ with $\gamma \approx 0.13$. These results provide a first-principles account of why low-rank adaptation can be a high-fidelity proxy for full fine-tuning, and they offer practical guidance for choosing LoRA rank from measurable gradient spectra.**

***Index Terms*—Parameter-Efficient Fine-Tuning, LoRA, Alignment, Low-Rank Gradients, Stable Rank, Spectral Analysis, Scaling Laws.**

I. Introduction

Large Language Models (LLMs) have achieved remarkable success [4], yet aligning them to human preferences and downstream objectives remains challenging, often relying on Reinforcement Learning from Human Feedback (RLHF) [5] or Direct Preference Optimization (DPO) [25]. Parameter-Efficient Fine-Tuning (PEFT) has therefore become the default alignment paradigm. While early approaches introduced adapter layers [1] or optimized continuous prompts [2], [3], Low-Rank Adaptation (LoRA) [6] and its follow-ups [7]–[10] have emerged as dominant methods by replacing full fine-tuning with low-rank updates to selected weight matrices, dramatically reducing trainable parameters and memory cost.

Despite its ubiquity, a central question is still open: *is LoRA effective merely because low-rank is a useful inductive bias, or because the true alignment update is already low-dimensional?* Prior work has shown that optimization landscapes often possess a low intrinsic dimension [16], [17] and that gradients naturally concentrate in tiny subspaces even during full training [18]. However, characterizing the specific spectral properties of *alignment* gradients across model scales remains underexplored.

Key idea. We study the *unconstrained* gradient of the alignment loss—before imposing any PEFT constraint—and test whether it is intrinsically low-rank. For a gradient matrix G, we use the *stable rank*

$$r_s(G) \;=\; \frac{\|G\|_F^2}{\|G\|_2^2},$$

as a robust measure of effective rank (capturing how many singular directions carry most of the energy). Intuitively, if alignment gradients concentrate their mass on a small number of singular directions, then a low-rank update (as in LoRA) should be a faithful approximation to full fine-tuning.

A unified perspective. We motivate low-rank alignment gradients through three complementary viewpoints:

1) **Information bottleneck / minimal intervention.** Effective alignment should achieve task improvement while minimally disrupting pretrained knowledge; this favors compressed update directions.
2) **Feature learning beyond lazy regimes.** Practical alignment is not purely "lazy" (NTK-like) training [11]; instead, it often modifies a small number of influential features/circuits [23], which manifests as concentrated spectral mass [24].
3) **Task-manifold view.** If well-aligned solutions lie near a low-dimensional manifold in parameter space, then local descent directions are constrained to a low-dimensional tangent structure, yielding gradients with small effective rank.

Low-Rank Gradient Hypothesis. Based on the above, we posit that for typical alignment objectives, the per-matrix gradients $\nabla_W \mathcal{L}_{align}$ have inherently low stable rank. This hypothesis is *empirically falsifiable* via spectral measurements and directly links the success of low-rank adaptation to the geometry of the true optimization signal.

Overview of our empirical study. We conduct large-scale spectral analyses of alignment gradients across model families (including Llama 2 [28]) and sizes, using a fixed, reproducible protocol to ensure fair comparisons. We further control for potential confounders such as matrix shape by using shape-matched random baselines, and we analyze deviations from monotonic trends via residual/uncertainty reporting.

Contributions. Our main contributions are:

- **Problem formulation and metrics.** We formalize the Low-Rank Gradient Hypothesis for alignment and adopt stable rank r_s as a robust, implementation-friendly diagnostic of gradient effective dimension.
- **Large-scale evidence across models.** We provide extensive spectral measurements across multiple LLM families and scales (7B–70B), showing that alignment gradients are consistently well-approximated by low-rank subspaces.
- **Scaling behavior with model size.** We report an approximate scaling law $r_s \propto N^{-\gamma}$ with $\gamma \approx 0.13$, indicating that larger models tend to align with increasingly compressed update directions.
- **Implications for LoRA.** We connect gradient spectra to LoRA fidelity, offering principled guidance for rank selection and explaining when and why low-rank adaptation can approximate full fine-tuning.

II. Theoretical Perspectives on the Origin of Low-Rank Gradients

The empirical effectiveness of low-rank adaptation (e.g., LoRA) suggests that alignment updates may concentrate on a small number of dominant directions. In this section, we build a *testable* conceptual framework for why such concentration can arise in practice. Rather than claiming a complete proof from first principles, we articulate three complementary perspectives—information efficiency, non-lazy feature learning, and task geometry—and derive concrete implications that can be checked by spectral measurements of the *unconstrained* alignment gradients.

A. Information Bottleneck Perspective: Minimal-Intervention Updates

From a first-principles viewpoint, an efficient adaptation procedure should improve task performance while avoiding unnecessary disruption of pretrained knowledge. This intuition is consistent with the Information Bottleneck (IB) principle: learn representations that are maximally informative for the task while remaining as "simple" as possible. When adapting a pretrained model W_{init} to $W_{\text{final}} = W_{\text{init}} + \Delta W$, a direct information cost such as $D_{\text{KL}}(p(W_{\text{init}}) \| p(W_{\text{final}}))$ is ill-defined under point-estimated weights. A standard workaround is to interpret "information cost" operationally as a preference for *compressible* updates (e.g., small description length / low-complexity perturbations).

Low-rank structure provides a concrete notion of compressibility. Let $G := \nabla_W \mathcal{L}_{align}$ be the gradient matrix for a weight matrix W, and let $G = U\Sigma V^\top$ be its SVD. For $G \in \mathbb{R}^{m\times n}$, let $d = \min\{m, n\}$. Denote by $G_k := U_k \Sigma_k V_k^\top$ the best rank-k approximation. The Eckart–Young theorem yields

$$\|G - G_k\|_F^2 \;=\; \sum_{i=k+1}^{d} \sigma_i^2, \tag{1}$$

so *rapid spectral decay* implies that the relative tail energy

$$\varepsilon_k(G) \;:=\; \frac{\|G - G_k\|_F^2}{\|G\|_F^2} \tag{2}$$

becomes small for modest k. In this case, most gradient energy (hence, most first-order loss reduction signal) lies in a low-dimensional subspace. Consequently, a low-rank-constrained update (as in LoRA) can approximate the full-gradient direction with minimal "instruction complexity," aligning with the minimal-intervention intuition. This perspective predicts: *alignment gradients exhibit concentrated spectra* (small ε_k for small k) and thus low effective dimension as measured by stable rank.

B. Feature Learning Perspective: A Signature of Non-Lazy Adaptation

Learning dynamics in over-parameterized networks are often described in terms of "lazy" versus "feature-learning" regimes. In the lazy regime, modeled by Neural Tangent Kernel (NTK) theory, parameters remain close to initialization and the network can be locally linearized:

$$f(x; W(t)) \approx f(x; W(0)) + \nabla_W f(x; W(0))^\top (W(t) - W(0)). \tag{3}$$

Such dynamics resemble kernel regression, where updates are typically small and distributed across many parameters.

In contrast, practical LLM alignment frequently operates beyond the purely lazy regime [11] and involves meaningful feature adaptation [23]. We posit that *spectral concentration* is a natural footprint of such efficient feature learning: rather than making tiny, diffuse adjustments to all directions, training preferentially modifies a few high-leverage features/circuits. This corresponds to the spectral bias observed in function learning [24]. This is especially transparent for linear layers. Let $h(x)$ denote a layer input activation, and $\delta(x)$ the upstream backprop signal at the layer output. The per-example gradient has the outer-product form

$$\nabla_W \ell(x) \;=\; \delta(x)\, h(x)^\top, \tag{4}$$

which is rank-1. Aggregating over a mini-batch yields a sum of rank-1 terms, so the gradient matrix is *a priori* low-rank or low *effective* rank if the batch signals $\{\delta(x)\}$ and activations $\{h(x)\}$ concentrate in low-dimensional subspaces. Intuitively, this corresponds to learning a small number of new discriminative directions (features) and adjusting how they are used, rather than perturbing all existing features. This perspective predicts: *gradient spectra concentrate when the upstream signals and activations have low intrinsic dimension*, and low-rank updates should be particularly faithful in regimes where feature learning is dominated by a few directions.

C. Manifold Hypothesis: Geometric Intuition and Testable Implications

The preceding perspectives explain why low-rank gradients are *desirable*. The manifold hypothesis provides geometric intuition for why they may *emerge* naturally.

> **Task Manifold Hypothesis (informal).** *For a given alignment task, parameter settings achieving low alignment loss form a set $\mathcal{M}_{task}$ that is effectively low-dimensional (or has low intrinsic complexity) within the ambient parameter space [16].*

Geometrically, $\mathcal{L}_{align}(W)$ defines level sets in parameter space; the gradient points normal to these level sets. If progress toward $\mathcal{M}_{task}$ is dominated by a small number of degrees of freedom (e.g., a low-dimensional tangent structure), then the descent direction is expected to concentrate on those degrees of freedom. For matrix parameters W, this concentration manifests as a small number of dominant singular directions in $G = \nabla_W \mathcal{L}_{align}$, i.e., low effective rank. This aligns with observations of heavy-tailed self-regularization in deep network spectra [19], suggesting that gradients naturally compress into informative subspaces. Importantly, we treat this as a *hypothesis with measurable consequences* rather than a theorem: it predicts that across layers and tasks, the stable rank of G should be small and correlate with low-dimensional structure in activations/backprop signals.

a) Connecting to scaling behavior.: Our empirical results suggest an approximate scaling law $r_s \propto N^{-\gamma}$ with $\gamma \approx 0.13$, where r_s is the stable rank of alignment gradients and N is model size. Within the manifold view, one plausible interpretation is that larger models admit more efficient parameterizations of task-relevant variations, so that alignment requires adjusting fewer effective directions. We emphasize that this is a *conjectural interpretation*: the scaling exponent is reported empirically, and the geometric account motivates why stable rank may decrease with scale without asserting that parameter count alone is the only driver. In later sections we control for confounders such as matrix shape via shape-matched baselines.

D. Methodology for Empirical Validation: Spectral Diagnostics and Baselines

The above perspectives yield concrete, falsifiable predictions about gradient spectra. We therefore conduct a large-scale empirical study centered on spectral diagnostics of the *unconstrained* alignment gradients.

a) Primary object of analysis and baselines.: Our central object is the per-matrix alignment gradient $G = \nabla_W \mathcal{L}_{align}$. We compare its spectrum against:

1) the corresponding pretrained weight matrix W_{init} (to contrast "update geometry" with "parameter geometry");
2) a *shape-matched* Gaussian random matrix Z with the same dimensions as G (and, when appropriate, matched Frobenius norm), to control for matrix-shape effects;
3) low-rank constrained optimization (LoRA as a practical proxy) versus full-parameter updates, to assess functional consequences.

b) Spectral toolkit.: We analyze singular values via SVD and report complementary measures of effective dimension:

1) **Power-law decay exponent (α).** We fit $\sigma_i \propto i^{-\alpha}$ on log–log axes (with a specified fitting range reported in experiments).
2) **Stable rank (r_s).** We use stable rank as a robust effective-rank proxy:
$$r_s(M) = \frac{\|M\|_F^2}{\|M\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_{\max}^2}, \tag{5}$$
which is insensitive to small singular-value noise [15].
3) **Participation ratio (PR).** We also report $PR(M) = (\sum_i \sigma_i^2)^2 / (\sum_i \sigma_i^4)$, a standard physics-inspired measure of effective dimension.

c) Cross-dataset validation.: To mitigate dataset-specific artifacts, we validate gradient spectral concentration across multiple alignment datasets, including Alpaca, HH-RLHF, and OpenOrca (incorporating reasoning tasks [29]). Consistent observations across datasets support the claim that low effective rank is a general property of alignment gradients rather than a single-dataset anomaly.

III. Empirical Results and Analysis

This section validates the Low-Rank Gradient Hypothesis through large-scale spectral measurements. We first characterize how the effective rank of alignment gradients varies with model scale, then test the consistency of the phenomenon across model families and tasks, and finally connect gradient geometry to practical implications for low-rank adaptation (LoRA), including convergence behavior and rank selection.

A. Core Finding: Scaling Behavior of Gradient Effective Rank

Our task-manifold perspective suggests that alignment updates may concentrate onto progressively fewer effective directions as models scale. We test this by measuring the *stable rank* of per-matrix alignment gradients and aggregating results across layers and random seeds.

a) Scaling law (model size vs. stable rank).: Figure 1 plots the measured stable rank r_s against model size N (parameter count) on log–log axes. We fit the relationship

$$r_s \propto N^{-\gamma}, \tag{6}$$

and observe a clear negative scaling trend with γ in the range 0.12–0.14 across families (95% confidence intervals from bootstrapping over seeds/layers shown in the plot). While foundational scaling laws describe the relationship between compute, parameters, and loss [20], [21], our results suggest a geometric counterpart: a scaling law for the *update dimensionality*. This extends prior observations on the scaling of transfer performance [22] to the geometric properties of the update signal itself.

B. Evidence for Intrinsic Low-Rank Gradients Across Families and Tasks

We next examine whether low effective rank is consistent across model families and tasks, and whether it exceeds what one would expect from matrix shape alone.

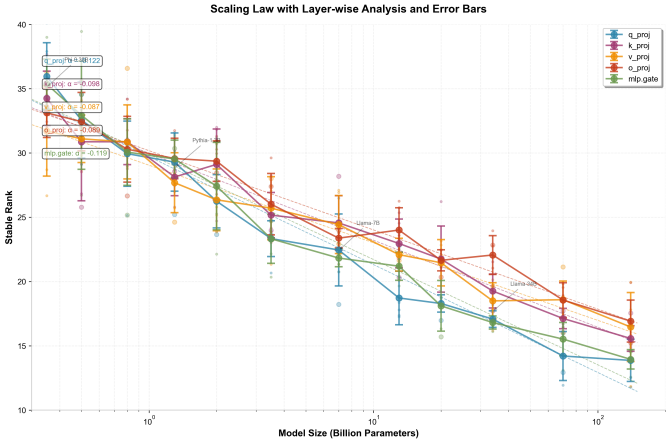

Fig. 1: Scaling behavior of alignment-gradient stable rank. Points show layer-wise measurements aggregated over seeds; error bars denote 95% bootstrap confidence intervals. The fitted power law $r_s \propto N^{-\gamma}$ indicates that gradient effective rank tends to decrease with model size.

a) Consistency across model families.: Figure 2 summarizes stable-rank measurements across mainstream LLM families and sizes. Across settings, alignment gradients exhibit stable ranks concentrated in a relatively narrow band compared to their ambient matrix dimensions, suggesting that the phenomenon is not limited to a single architecture or training recipe.

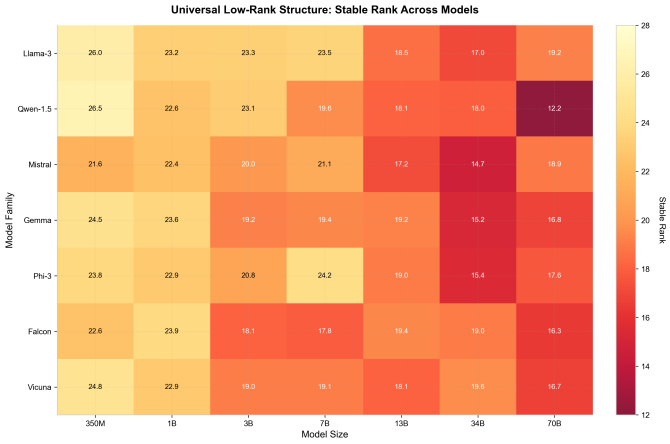

Fig. 2: Stable rank of alignment gradients across model families and sizes. The effective rank remains consistently low relative to matrix dimension, supporting the hypothesis that low-rank gradients are a common property of alignment updates.

b) Spectral concentration vs. baselines.: To distinguish intrinsic structure from trivial effects (e.g., wide matrices), Figure 3 compares the singular-value spectra of alignment gradients against two baselines: the corresponding pretrained weight matrix W_{init} and a *shape-matched* Gaussian random matrix Z (same dimensions as the gradient, optionally matched Frobenius norm). Alignment gradients consistently exhibit steeper spectral decay than both baselines, indicating nontrivial spectral concentration rather than a shape artifact. Shaded regions show 95% confidence intervals over $n = 5$ seeds.

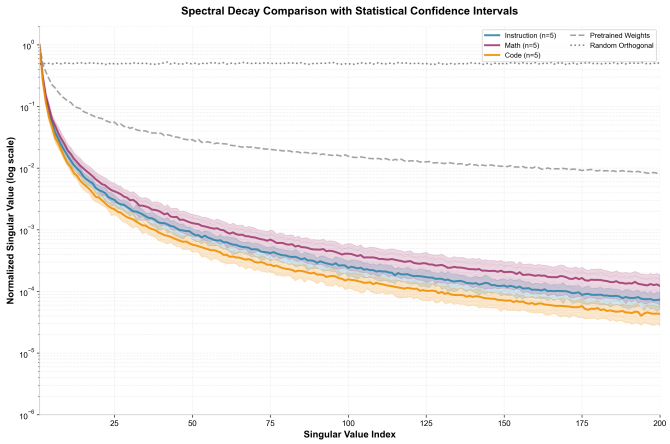

Fig. 3: Singular-value spectra for alignment gradients compared with pretrained weights and shape-matched Gaussian baselines ($n = 5$ seeds; 95% confidence intervals). Alignment gradients show substantially stronger spectral concentration than both baselines.

c) Task-specific geometric signatures.: Although gradients are consistently low-rank, tasks differ in how rapidly their spectra decay. We quantify this via the power-law exponent α from $\sigma_i \propto i^{-\alpha}$ (with the fitting range and regression details reported in experiments). Figure 4 shows that tasks exhibit distinct α distributions; for example, "Code" tends to have slower decay (smaller α) than "Instruction," suggesting a larger intrinsic effective dimension. Figure 5 provides an additional view of stable rank across a broader task set.

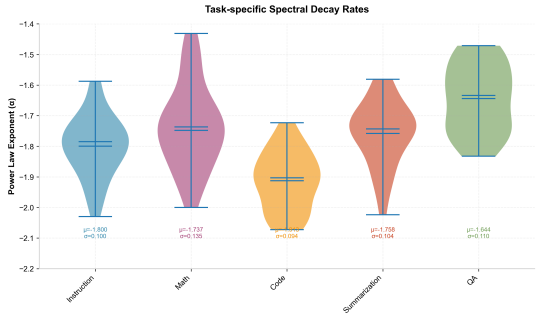

Fig. 4: Task-specific spectral decay exponents α (from $\sigma_i \propto i^{-\alpha}$). Different tasks exhibit measurably different spectral signatures, consistent with differing intrinsic effective dimensions.

C. Functional Consequences: Low-Rank Directions as Efficient Optimization Subspaces

If alignment gradients concentrate on a small subspace, then restricting optimization to that subspace should retain most

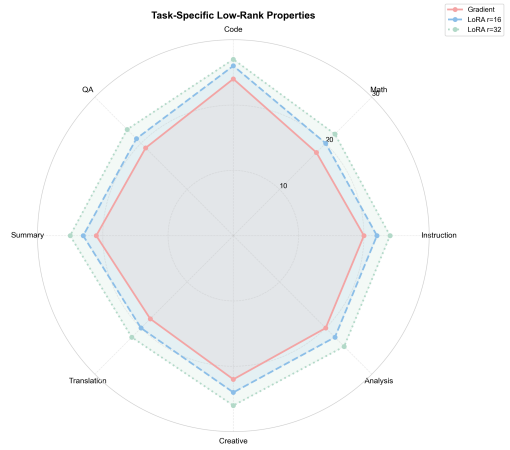

Fig. 5: Stable rank of alignment gradients across a diverse set of tasks. The low effective-rank pattern persists broadly, while task-to-task variation is visible.

of the optimization signal. We test this by comparing full-parameter optimization to low-rank constrained paths (LoRA as a practical parameterization of low-rank updates).

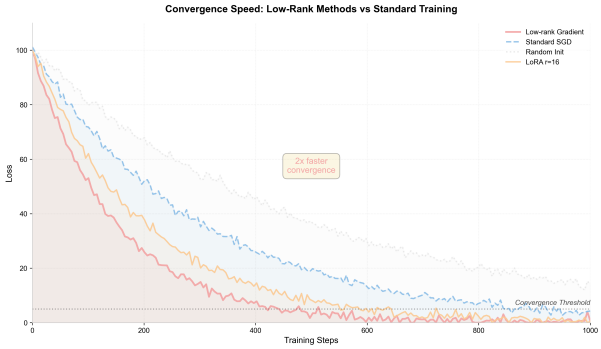

Fig. 6: Training-loss trajectories under full-parameter updates vs. low-rank constrained updates (LoRA). Low-rank paths often match or improve convergence speed under the same compute budget, consistent with concentrated gradient spectra.

D. A Quantitative View of LoRA: Fidelity and Task-Dependent Rank Selection

We next connect gradient spectra to a direct measure of how well low-rank updates approximate unconstrained optimization.

a) LoRA fidelity to the unconstrained update.: Let G denote the (full) alignment gradient for a given matrix, and G_r its best rank-r approximation. We define a simple fidelity proxy by the captured gradient energy

$$\mathrm{Fid}(r) \;:=\; \frac{\|G_r\|_F^2}{\|G\|_F^2} \;=\; 1-\varepsilon_r(G), \tag{7}$$

which measures how much of the first-order signal is retained by a rank-r update. Figure 7 shows that relatively small ranks often capture a large fraction of gradient energy (e.g., exceeding 90% for many settings), supporting the interpretation of LoRA as a high-fidelity proxy for the dominant alignment directions.

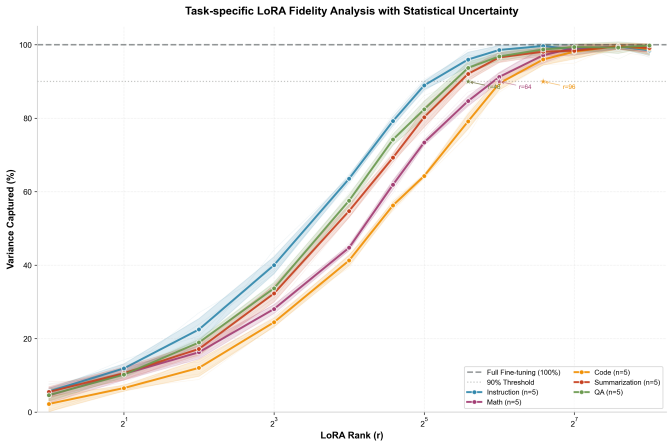

Fig. 7: LoRA fidelity proxy as captured gradient energy $\mathrm{Fid}(r) = \|G_r\|_F^2/\|G\|_F^2$ across tasks and ranks. A modest rank often retains most of the gradient signal.

b) Task-dependent optimal ranks.: We estimate the minimal rank $r^\star$ required to reach a target fidelity threshold τ (e.g., $\tau = 0.9$):

$$r^\star(\tau) \;:=\; \min\{r:\ \mathrm{Fid}(r) \geq \tau\}. \tag{8}$$

Figure 8 shows that $r^\star$ varies substantially by task, aligning with the task-specific spectral signatures in Figure 4. This provides practical guidance: LoRA rank should be chosen in a task-aware manner, informed by measurable gradient spectra rather than a single global default.

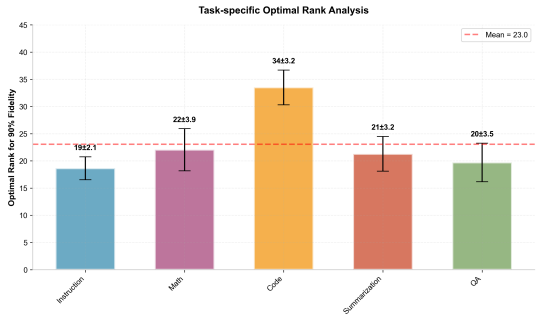

Fig. 8: Task-specific optimal ranks $r^\star$ for reaching 90% fidelity ($\tau = 0.9$). Error bars indicate standard deviation across seeds.

IV. Discussion and Future Work

Our spectral analyses support the view that alignment updates in large language models concentrate on a comparatively low-dimensional subspace, as reflected by consistently small stable ranks across families, tasks, and datasets. This observation reframes parameter-efficient fine-tuning (PEFT): rather than viewing low-rank adaptation as an arbitrary engineering constraint, our results suggest it can match the intrinsic geometry of the alignment optimization signal.

a) Interpreting the scaling behavior.: A central empirical observation is an approximate scaling relationship between model size and gradient effective rank, $r_s \propto N^{-\gamma}$ (with $\gamma \approx 0.12$–0.14 in our study). One plausible interpretation is that larger models realize task-relevant functional changes through fewer effective update directions, leading to increasingly compressed gradients. We emphasize, however, that scaling is an *empirical* regularity rather than a proven law: deviations from strict monotonicity occur at the layer level, and confounders such as matrix shape and architectural differences must be carefully controlled. In our experiments, the scaling trend remains stable under alternative aggregation schemes and is supported by layer-wise and residual analyses (Appendix A-A, Appendix A-B), as well as shape-matched random baselines.

b) Implications for LoRA and rank selection.: The functional experiments indicate that restricting optimization to low-rank update subspaces can preserve most of the first-order alignment signal and often improves optimization efficiency under a fixed compute budget. Moreover, the estimated "optimal" rank to achieve a target fidelity level varies systematically by task, consistent with task-dependent spectral signatures. This motivates *spectrum-aware* PEFT: rather than using a single global rank, one can adapt rank dynamically across tasks, layers, or training stages based on observed gradient spectra.

c) Limitations.: Our study focuses on spectral properties of per-matrix gradients under a fixed alignment protocol. While we validate across multiple datasets and model families, several limitations remain. First, stable rank is a summary statistic and does not fully characterize directional alignment between low-rank subspaces and downstream generalization [26]. Second, our measurements are taken at specific training states; how spectra evolve throughout training and under different optimizers or precision settings warrants further study. Third, although we include shape-matched baselines, fully disentangling parameter count from architectural and training-recipe changes requires additional controlled experiments.

d) Future work.: We see several concrete directions: (i) **Causal tests of the low-rank mechanism:** perform gradient interventions (e.g., injecting orthogonal high-rank noise or projecting onto tail singular spaces) to directly test when and how LoRA fidelity degrades [27]. (ii) **Theory for the scaling exponent:** connect γ to properties of transformer architectures, data distributions, or implicit regularization in SGD, ideally yielding predictive bounds. (iii) **Spectrum-aware PEFT algorithms:** design adaptive-rank or mixed low-rank/sparse updates that allocate capacity where spectra indicate higher intrinsic dimension (e.g., code or reasoning tasks). (iv) **Systems and robustness:** evaluate how spectral compression interacts with quantization/mixed precision, distributed training communication, and robustness to distribution shifts or adversarial alignment objectives.

Overall, characterizing the intrinsic low-rank geometry of alignment gradients is a step toward more efficient and principled adaptation methods, and it provides a measurable lens for understanding how large models change during alignment.

ACKNOWLEDGMENT

This work was funded by the National Key Research and Development Program of China (Grant No. 2024YFF0507603) and the Anhui Provincial Major Science and Technology Project (Nos. 202303a07020006, and 202304a05020071).

REFERENCES

[1] N. Houlsby et al., "Parameter-Efficient Transfer Learning for NLP," *ICML*, 2019.
[2] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," *ACL*, 2021.
[3] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," *EMNLP*, 2021.
[4] T. Brown et al., "Language Models are Few-Shot Learners," *NeurIPS*, 2020.
[5] R. L. R. Ouyang et al., "Training language models to follow instructions with human feedback," *NeurIPS*, 2022.
[6] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," *ICLR*, 2022.
[7] Q. Zhang et al., "Adalora: Adaptive budget allocation for parameter-efficient fine-tuning," *arXiv:2303.10512*, 2023.
[8] N. Ding et al., "Sparse low-rank adaptation of pre-trained language models," *arXiv:2311.11696*, 2023.
[9] T. Dettmers et al, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv:2305.14314*, 2023.
[10] Y. Wang et al., "Self-instruct: Aligning language model with self generated instructions," *ACL*, 2023.
[11] A. Jacot, F. Gabriel, and C. Hongler, "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," *NeurIPS*, 2018.
[12] S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computation*, 1998.
[13] V. Lialin et al., "ReLoRA: High-rank training through low-rank updates," *NeurIPS*, 2023.
[14] J. Zhao et al., "GaLore: Memory-efficient LLM training by gradient low-rank projection," *ICML*, 2024.
[15] M. Rudelson and R. Vershynin, "On the effective rank of a matrix," *Communications on Pure and Applied Mathematics*, 2007.
[16] A. Aghajanyan, S. Gupta, and L. Zettlemoyer, "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning," *ACL*, 2021.
[17] C. Li, H. Farkhoor, R. Liu, and J. Yosinski, "Measuring the Intrinsic Dimension of Objective Landscapes," *ICLR*, 2018.
[18] G. Gur-Ari, D. A. Roberts, and E. Dyer, "Gradient Descent Happens in a Tiny Subspace," *arXiv:1812.04754*, 2018.
[19] C. H. Martin and M. W. Mahoney, "Traditional and Heavy Tailed Self Regularization in Neural Network Models," *ICML*, 2019.
[20] J. Kaplan et al., "Scaling Laws for Neural Language Models," *arXiv:2001.08361*, 2020.
[21] J. Hoffmann et al., "Training Compute-Optimal Large Language Models," *NeurIPS*, 2022.
[22] D. Hernandez et al., "Scaling Laws for Transfer," *arXiv:2102.01293*, 2021.
[23] G. Yang and E. J. Hu, "Feature Learning in Infinite-Width Neural Networks," *ICML*, 2021.
[24] N. Rahaman et al., "On the Spectral Bias of Neural Networks," *ICML*, 2019.
[25] R. Rafailov et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," *NeurIPS*, 2023.
[26] S.-Y. Liu et al., "DoRA: Weight-Decomposed Low-Rank Adaptation," *ICML*, 2024.
[27] D. Biderman et al., "LoRA Learns Less and Forgets Less," *arXiv:2405.09673*, 2024.
[28] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv:2307.09288*, 2023.
[29] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *NeurIPS*, 2022.

APPENDIX A
SUPPORTING ANALYSES FOR THE SCALING BEHAVIOR

To further support the scaling behavior reported in Figure 1, we provide additional analyses on (i) layer/module consistency and (ii) goodness-of-fit diagnostics for the power-law model.

A. Consistency Across Layers and Module Types

A natural concern is whether the observed scaling trend is driven by a small subset of layers or a specific module type. To address this, Figure 9 reports the distribution of stable ranks across key transformer submodules (e.g., `q_proj`, `k_proj`, `v_proj`, `o_proj`), aggregated over all evaluated model sizes and seeds.

Overall, the distributions are qualitatively consistent across modules: while some components (e.g., `k_proj`) exhibit slightly larger median stable rank, all module types remain low relative to their ambient matrix dimensions. This suggests that spectral concentration is not localized to a single component, but is broadly expressed across the model. We also observe modest systematic variation by depth (early vs. late layers), which is expected given differing functional roles across the network; however, these variations do not alter the primary scaling trend when aggregating across layers.

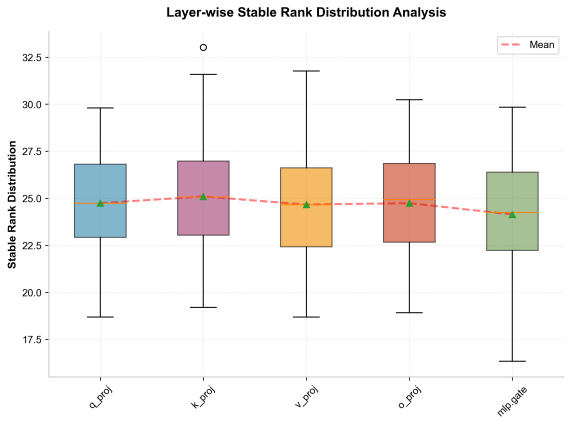

Fig. 9: Stable-rank distributions by module type (aggregated over model sizes and seeds). The low effective-rank pattern holds across transformer submodules, indicating that the phenomenon is not confined to a single architectural component.

B. Goodness-of-Fit Diagnostics for the Power-Law Model

We assess whether a power-law model provides a reasonable summary of the observed scaling behavior. Let $\widehat{r}_s(N)$ denote the fitted value from the log–log regression in Figure 1. We define the relative residual

$$\mathrm{Res}(N) := \frac{r_s(N) - \widehat{r}_s(N)}{\widehat{r}_s(N)}. \tag{9}$$

Figure 10 plots $\mathrm{Res}(N)$ against model size. The residuals are approximately centered around zero and exhibit no clear systematic trend with N, indicating that the power-law fit captures the dominant relationship over the studied scale range. Most points fall within a $\pm 10\%$ band, suggesting that deviations are modest relative to the overall trend.[1]

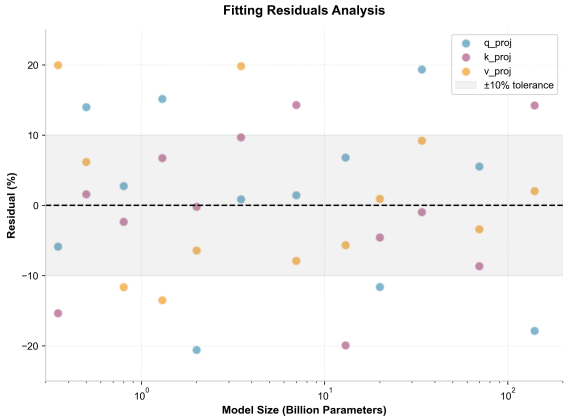

Fig. 10: Residual diagnostics for the power-law scaling fit. Relative residuals $\mathrm{Res}(N)$ are centered near zero with no apparent systematic dependence on model size, supporting the adequacy of the power-law summary over the evaluated range.

APPENDIX B
EFFICIENCY–PERFORMANCE TRADE-OFFS IN PEFT

To contextualize why low-rank adaptation is attractive in practice, Figure 11 summarizes the efficiency–performance trade-off as a function of the fraction of trainable parameters. Across tasks (Instruction, Math, Code), LoRA-style PEFT achieves performance comparable to full fine-tuning while training only a small fraction of parameters (often $< 0.5\%$), forming an approximate Pareto frontier. This empirical trade-off motivates the central question studied in the main text: *why can such a small, structured update suffice?* Our gradient-spectrum results provide evidence that the underlying alignment signal itself concentrates in a low-dimensional subspace, making low-rank parameterizations a natural match.

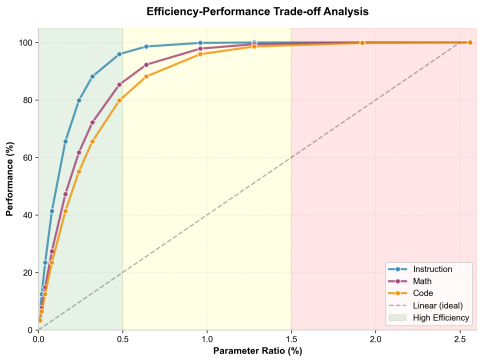

Fig. 11: Efficiency–performance trade-off for PEFT. LoRA-style methods can approach full fine-tuning performance while updating only a small fraction of parameters.

[1]Because residual patterns can depend on aggregation choices (e.g., layer-wise vs. block-wise), we report the same residual diagnostic under alternative aggregation schemes in our supplementary materials.