Mechanistic Insights into LoRA:
Layer Sparsity for Adaptive Fine-Tuning via Path Patching

NeurIPS 2026

Fei Zuo*[1] Yizhou Huang✉*[2] Kezhi Wang✉[2]

[1] East China Normal University [2] Brunel University of London

8 of 28 layers achieve higher reasoning accuracy.

Qwen2.5-1.5B · rank 16 · same target projections

Adapted layers

Full-layer LoRA 28 / 28

Adaptive-LoRA **8 / 28**

3.5× fewer adapter parameters

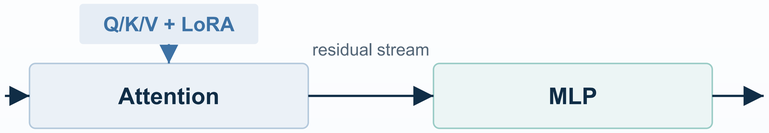
Commonsense accuracy

56.9% Full-layer LoRA → **59.1%** **Adaptive-LoRA** (+2.2 pp)

01 Where does LoRA act?

Path patching of trained adapters reveals concentrated, task-dependent effects.

Adapters and their downstream effects

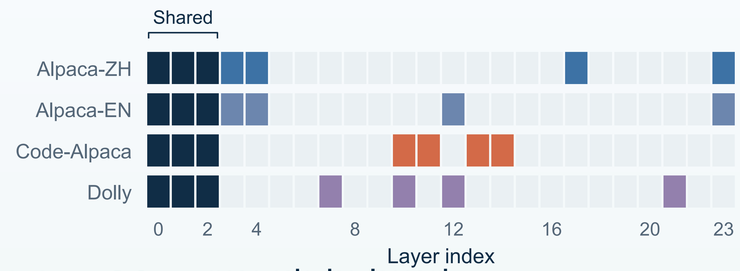


Normalized component effect



MLP blocks mediate the larger measured effect, even with adapters attached only to Q/K/V.

Important layers change with the task



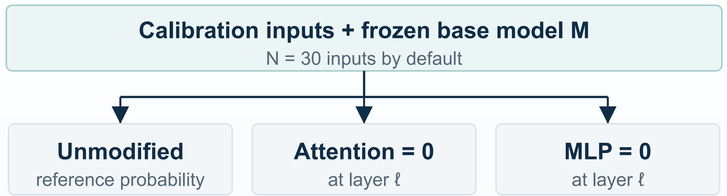
39.5% **mean pairwise Jaccard** between the four top-7 layer sets

Only layers 0–2 appear in every task's top-7 set. The remaining selected layers depend on the task.

Analysis: Qwen2.5-0.5B, 24 layers, Q/K/V adapters, rank 8.

02 How do we select layers?

Adaptive-LoRA calibrates on the frozen base. Layer selection requires no trained adapter.



$\delta_{\ell,c}(x)$: absolute change in the base model's top-token probability after zeroing output c.

$$I_\ell = \frac{1}{N} \sum_{x \in D_{\text{cal}}} (\delta_{\ell,a}(x) + \delta_{\ell,m}(x))$$

Select the top-k layers by importance score

Train LoRA only at the selected layers

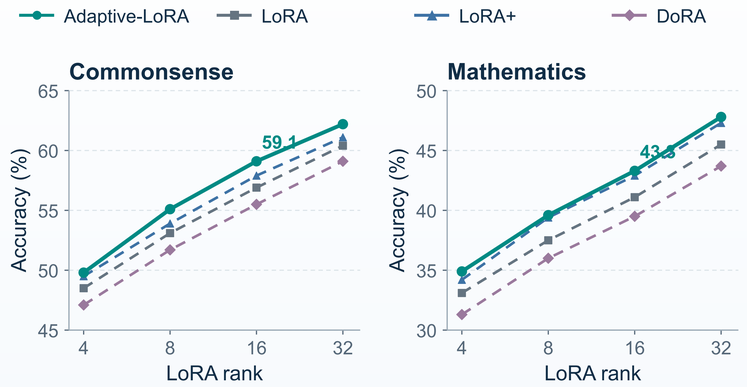
Base scores anticipate trained-adapter importance

Targets / rank	Spearman ρ	Top-7	Top-12
Q/K/V r = 8	0.630	5/7	9/12
Q/K/V r = 16	0.635	5/7	10/12
Q/V r = 4	0.893	5/7	10/12
Q/V r = 16	0.891	5/7	10/12
Q/V r = 32	0.867	5/7	10/12

Qwen2.5-0.5B, 100 inputs. Top-k recovery = intersection / k.

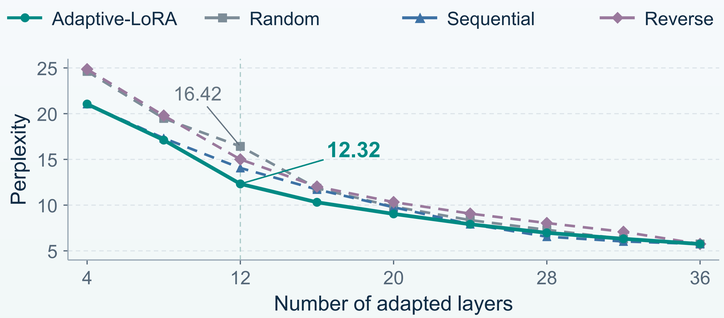
03 What does this buy?

Higher reasoning accuracy with ~30% of layers across all four tested ranks.



Qwen2.5-1.5B. 8 commonsense and 3 math tasks.
Adaptive: 8/28 layers. Baselines: all 28.
250 calibration inputs in this rank sweep.

Layer choice matters at the same budget



25% lower perplexity than random at 12/36 layers

Qwen2.5-3B, Alpaca-ZH, rank 16. Lower perplexity is better.

Mechanistic insight guides placement. Frozen-base calibration makes it usable before fine-tuning.

NEURIPS 2026 / PAPER 17827

* First authors
✉ Corresponding authors