A-LoRA: Vocabulary-Boundary Adaptation through Shared Hidden-Space Transforms

Anonymous authors
Paper under double-blind review

Abstract

Vocabulary-boundary layers offer direct access to how a language model encodes tokens and predicts its next output, but their large vocabulary dimension makes conventional low-rank updates expensive. We present **A-LoRA**, which adapts these layers through a low-rank hidden-space transform and a frozen vocabulary matrix. Its $2dr$ trainable parameters are independent of vocabulary size; on tied models, one transform can serve both input and output while preserving a single mergeable matrix. We connect this parameterization to the geometry of 30 Base→Instruct model pairs and characterize the two computational paths activated by sharing. In five-seed boundary-only experiments on Qwen3-0.6B, shared A-LoRA reaches 50.90 CLUENER F1 and 59.14 WikiSQL execution accuracy with 32,768 parameters. At identical rank, selected learning rate, and parameter count, sharing improves on input-only adaptation by 5.83 and 4.24 points, with gains in five of five and four of five seeds. A separately tuned input-only control yields 47.37 versus 26.67 CLUENER F1 at near-matched budgets of 38,912 and 39,708 parameters. Evaluating the same checkpoints on unrelated tasks reveals distinct tradeoffs between source adaptation, answer preference, and generated output. A fixed-prefix logit analysis and decoding intervention explain when retained candidate rankings coexist with premature stopping. Together, the results identify pretrained vocabulary structure and input–output sharing as useful resources for small-budget adaptation.

1 Introduction

Low-rank adaptation usually modifies attention and feed-forward projections inside a language model (Hu et al., 2022; Mangrulkar et al., 2022). Its input embedding and output head provide two additional intervention points: one determines what enters the frozen computation, and the other determines how that computation becomes a token prediction. We study these *vocabulary-boundary layers* as a small-budget adaptation interface. The central question is how to allocate trainable capacity across the input, the output, and their shared structure.

Vocabulary layers differ from internal projections in two consequential ways. First, their $V \times d$ matrices often have $V \gg d$, so a conventional rank-r update spends rV parameters on vocabulary-indexed factors. Second, many language models tie the input embedding and output head. Updating one side, updating both independently, and sharing one update across both sides define different adapted functions and different parameter budgets. Boundary adaptation therefore calls for a joint treatment of parameterization and placement.

A-LoRA adapts the transform rather than learning new vocabulary-indexed factors. It writes $M_{\text{eff}} = M_0(I + sPQ)$, where M_0 is frozen and P, Q are hidden-space low-rank factors. This uses $2dr$ parameters and links token updates through their pretrained vectors. The structure has a concrete implication: linear relations among vocabulary rows are preserved by the update. On tied models, one transform acts on both the input representation and output readout, retaining a single matrix after merging. We derive these computational roles and relate the parameterization to measured post-training changes in 30 public Base→Instruct pairs.

Sharing gives a direct same-budget test of boundary placement. With the backbone frozen, shared A-LoRA on Qwen3-0.6B reaches 50.90 CLUENER F1 and 59.14 WikiSQL execution accuracy using 32,768 parameters. Compared with input-only A-LoRA at the same rank and selected

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.

learning rate, it improves the five-seed means by 5.83 and 4.24 points. The paired-seed results and complete input/output/bilateral comparisons reveal where the extra computational path helps. A separately tuned input-side comparison retains a 20.70-point F1 advantage over an active-row ordinary-LoRA control with 2.05% more parameters, distinguishing nominal vocabulary cost from the parameters activated in training.

Useful adaptation also requires understanding retained behavior. We evaluate the same checkpoints on unrelated science and inference questions, using candidate scores, full-vocabulary first-token readouts, and actual generation. These measurements distinguish the task an adapter learns from how it changes answer preferences, stopping, and output format. A cross-size output-only study connects large EOS-margin shifts to premature stopping; a uniform decoding intervention separates this mechanism from persistent repetitive generation. This turns retention evaluation into an analysis of adapter behavior rather than a single auxiliary score.

Our contributions are a vocabulary-independent boundary parameterization with explicit row coupling and tied sharing; a five-seed placement study demonstrating boundary-only task capacity and a same-budget sharing benefit; and a functional analysis linking boundary updates to source adaptation and retained generation. The resulting design space includes which boundary to adapt, how to use pretrained vocabulary structure, and when to share the learned transform.

2 A-LoRA: Parameterization, Computation, and Sharing

2.1 A Hidden-Space Transform for Vocabulary Matrices

Let $E_0, U_0 \in \mathbb{R}^{V\times d}$ be frozen input and output matrices. For either matrix M_0, dense vocabulary LoRA (W-form) and A-LoRA use

$$M_{\text{eff}} = M_0 + \Delta M, \tag{1}$$

$$\Delta M_W = sBC, \qquad B \in \mathbb{R}^{V\times r}, \quad C \in \mathbb{R}^{r\times d}, \tag{2}$$

$$\Delta M_{\text{aff}} = sM_0PQ, \qquad P \in \mathbb{R}^{d\times r}, \quad Q \in \mathbb{R}^{r\times d}, \tag{3}$$

where $s = \alpha/r$. Defining $T = I + sPQ$ gives

$$M_{\text{eff}} = M_0T, \qquad \underbrace{P_W = r(V+d)}_{\text{dense W-form}}, \qquad \underbrace{P_{\text{aff}} = 2dr}_{\text{A-LoRA}}. \tag{4}$$

At equal rank, the A-LoRA update is a structured W-form update with $B = M_0P$ and $C = Q$. Its parameter savings can be allocated to higher hidden-space rank. Both forms support exact weight merging in real arithmetic.

Pretrained rows couple token updates. For every vector $c \in \mathbb{R}^V$ satisfying $c^\top M_0 = 0$, A-LoRA satisfies $c^\top \Delta M_{\text{aff}} = 0$. Thus any linear relation among pretrained vocabulary rows also holds among their updates. An unseen training token can receive a nonzero update through its pretrained vector. This provides vocabulary-wide parameter sharing using only hidden-dimensional factors. Appendix A.2 derives this identity and contrasts it with row-local input LoRA.

2.2 One Transform, Two Computational Paths

Write the input embeddings for a prefix as columns of $X \in \mathbb{R}^{d\times n}$, and let $F(X) \in \mathbb{R}^d$ be its final frozen-backbone representation. With bias-free adapters, the four placements give

$$\begin{aligned}
&E\text{-only:} && z = U_0F(T_E^\top X),\\
&U\text{-only:} && z = U_0T_UF(X),\\
&\text{independent } E+U: && z = U_0T_UF(T_E^\top X),\\
&\text{shared } E+U: && z = W_0TF(T^\top X), \quad E_0 = U_0 = W_0.
\end{aligned} \tag{5}$$

Input adaptation changes the representation computed throughout the backbone; output adaptation changes the readout of its existing representation. Sharing uses the same factors on both paths. At $T = I + \epsilon R$, its first-order change is

$$\left.\frac{dz}{d\epsilon}\right|_0 = W_0Rh_0 + W_0J_F(X)[R^\top X], \qquad h_0 = F(X), \tag{6}$$

where $J_F(X)[\cdot]$ denotes a directional derivative. The two terms describe how one parameter set can jointly adjust token encoding and token prediction. Their empirical benefit is tested by comparing shared and input-only adapters at the same budget.

Shared placement is used only on tied bases. It preserves $E_{\text{eff}} = U_{\text{eff}} = W_0 T$, so the adapted model merges into one vocabulary matrix. Unilateral or independent adapters generally break effective tying even when the base tensor remains physically shared. Independent bilateral adapters have twice the parameter count at the same per-side rank; untied bases use unilateral or independent placement.

2.3 Dense and Active-Row Input Budgets

For input-only W-form with zero-initialized vocabulary factor B and zero weight decay, only training-observed token rows become nonzero. If S is that token set, the identical adapter can be stored with $r(|S| + d)$ parameters, rather than $r(V + d)$. Output softmax gradients are dense, so this compaction is specific to the input-only setting. A-LoRA instead shares the transform over all pretrained rows.

For Qwen3-0.6B on CLUENER, $|S| = 8{,}903$ and $d = 1{,}024$: dense rank-one W-form has 152,960 nominal parameters and 9,927 active-row parameters. A near-budget input control compares active-row W-form rank four (39,708) with A-LoRA rank 19 (38,912), a 2.05% difference. We distinguish these budgets throughout. All new comparisons disable bias. Historical variants and the softmax-invariant row-shared output bias are specified in Appendix A.13.

3 Geometry Motivating the Parameterization

3.1 Alignment of Observed Post-Training Changes

A-LoRA uses the pretrained matrix as a vocabulary basis. To assess this choice, we analyze corresponding Base and Instruct matrices $X, Y \in \mathbb{R}^{V \times d}$ and the change

$$\Delta W = Y - X \approx X(A - I) + \mathbf{1}_V b^\top. \tag{7}$$

For $Z = [X, \mathbf{1}_V]$, retain leading left singular vectors Q_K explaining 95% of its squared singular-value energy. The projected change fraction is

$$R^2_{\text{aff},K} = \|Q_K Q_K^\top \Delta W\|_F^2 / \|\Delta W\|_F^2. \tag{8}$$

This measures directional alignment of observed weight changes. Model lists, per-model values, exclusions, and the affine bias convention appear in Appendix A.9.

3.2 Structure and Affordable Rank

Across 30 retained pairs, output alignment exceeds input alignment in all 13 untied pairs (Table 1). Tied matrices have similar median alignment to output heads. The pattern persists beyond update magnitude: DeepSeek-V3's output change norm is approximately $0.4\times$ its input change norm, but its projected fraction is $2.7\times$ larger.

Table 1: Weight-space evidence across 30 Base→Instruct pairs. Alignment uses the 95%-energy affine subspace. Fit ratio compares offline explained gain at the nominal budget of W-form rank one.

Matrix group	Pairs	Alignment	Affine/W fit ratio	Affine wins
Untied input E	13	3.6%	0.34×	1/13
Untied output U	13	33.0%	2.18×	12/13
Tied $E = U$	17	32.2%	3.00×	15/17

Nominal parameter matching allocates a larger rank to the hidden-space transform:

$$r_{\text{aff}} \approx r_W (V + d)/(2d). \tag{9}$$

The analyzed V/d ratios of 15.7–242.5 yield roughly rank 8–122 A-LoRA per W-form rank one before rounding. Offline fits favor the affine form on most output and tied matrices at this budget. Together, the fits motivate spending parameters on a structured transform with higher affordable rank. Downstream placement additionally depends on computation: input perturbations propagate through F, while output perturbations act on its readout (Eq. 5). We therefore evaluate each placement directly rather than choosing it from weight alignment alone.

4 Experimental Protocol

Source-task adaptation. The primary placement study uses the tied Qwen3-0.6B Base model, with a frozen backbone and no hidden-layer adapters. CLUENER (Xu et al., 2020) is generated as typed entity spans in JSON; WikiSQL (Zhong et al., 2017) is generated as a structured query object. We report entity micro-F1 and query execution accuracy, respectively, on a 0–100 scale; normalized logical-form exact match is reported as a secondary WikiSQL metric. Each task uses 2,048 training examples, with 200/1,024 development examples and 1,343/1,024 held-out examples for CLUENER/WikiSQL. CLUENER's labeled public development set serves as our held-out evaluation, rather than its unlabeled official test set. Source metrics evaluate both task content and its required serialization.

Training and selection. We compare A-LoRA rank 16 with dense W-form rank one at all four placements, explicitly reporting unequal parameter counts. Both use scale $s = 8$, zero initial residual, AdamW with zero weight decay, effective batch 32, 64 steps, 3% warmup, cosine decay, and gradient clipping at 1. Training uses a bfloat16 backbone and float32 adapter/optimizer state; source and retention inference use float32. For each method and placement, learning rate is selected by mean source-development score over two seeds from $2 \times 10^{-4}\{1/16, 1/4, 1, 4\}$. A best-at-edge result triggers one finite, paired extension for both methods to $1/64$ and/or 16 on the relevant side. The selected configuration is then evaluated on five fresh seeds. Remaining boundary optima are recorded explicitly. Selected rates and seed ranges are in Appendix A.1.

Retention on other tasks. The same adapted checkpoints are evaluated without training on SciQ (Welbl et al., 2017) and ANLI round 1 (Nie et al., 2020), each with 1,000 fixed questions and all cyclic answer-order rotations (four/three). We distinguish (i) *pooled* candidate scoring, which normalizes probabilities over answer candidates in each rotation, aligns semantic choices, and averages these probabilities before choosing; (ii) *single-rotation* accuracy, averaged after making each rotated prediction; and (iii) actual greedy generation, scored on the same fixed 64-question subset for every method, with all rotations and a 32-token cap. Full-vocabulary first-token EOS rates are recorded separately. A uniform sensitivity condition requires at least one token before EOS; default decoding is always retained. Restricted candidate scoring is not equivalent to unconstrained generation.

Study organization and reproducibility. The primary study contains 80 adapters: two tasks, two parameterizations, four placements, and five fresh seeds. A near-budget input control adds ten adapters on CLUENER, using a fixed five-point rate grid $2 \times 10^{-4}\{1/16, 1/4, 1, 4, 16\}$, two development seeds, and the same five fresh CLUENER seeds. A separate fixed-rate study covers 45 output-only adapters across eight Qwen2.5/Qwen3 sizes. Historical stacked and instruction-loss experiments are reported separately in the appendix. We audit trainable counts, frozen base weights, zero-residual initialization, loss masks, checkpoint reloads, and saved-response scores. The anonymous supplement provides per-seed results, analysis code, and protocol details.

5 Boundary Adaptation with a Shared Parameter Budget

5.1 Boundary-Only Adaptation Has Substantial Task Capacity

Table 2 demonstrates substantial task learning using only vocabulary-boundary parameters. Input-only A-LoRA reaches 45.06 CLUENER F1 and 54.90 WikiSQL execution accuracy with 32,768 parameters and a frozen backbone. At the reported configurations, A-LoRA has higher mean source scores than dense W-form LoRA at every placement on both tasks. The source metric and the parameter count are reported together, with active-row input accounting in §2.3.

Table 2: Boundary-only source performance on Qwen3-0.6B. Every cell lists A-LoRA / dense W-form LoRA, averaged over five fresh seeds. A uses rank 16; W uses rank one; actual parameter counts are shown. Frozen base: CLUENER 0.59; WikiSQL 5.37.

Placement	Parameters (A / W)	CLUENER F1	WikiSQL EX
E	32,768 / 152,960	45.06 / 11.29	54.90 / 5.18
U	32,768 / 152,960	36.46 / 17.00	5.78 / 5.47
Independent $E + U$	65,536 / 305,920	49.09 / 20.28	55.78 / 4.71
Shared $E + U$	32,768 / 152,960	50.90 / 16.51	59.14 / 5.51

Placement has a pronounced functional effect. On WikiSQL, input-only A-LoRA reaches 54.90, while output-only A-LoRA remains near the base at 5.78 versus 5.37. Input adaptation can modify the representations computed throughout the frozen backbone; output-only adaptation operates on the existing representation. This contrast makes boundary placement a substantive modeling choice even within the same low-rank parameterization.

A complementary CLUENER check compares entity content on paired examples where both methods produce schema-valid outputs. A-LoRA retains a 19.94–33.55 point F1 advantage across placements on these intersections. This descriptive, post-hoc comparison shows a content distinction within valid outputs; subset construction and coverage are reported in Appendix A.1.

5.2 A Near-Budget Input Control

Dense vocabulary counts are not the only relevant budget for input adaptation. We therefore compare A-LoRA rank 19 with ordinary W-form rank four stored only on the 8,903 training-observed token rows. Each method independently selects its rate on two development seeds before the five fresh-seed runs. The ordinary control has 39,708 parameters, 2.05% more than A-LoRA's 38,912, and reaches 26.67 CLUENER F1 versus 47.37 for A-LoRA (Table 3). The mean paired gain is 20.70 points, positive in all five seeds, with ranges 24.91–29.57 and 46.82–48.47, respectively.

This comparison removes the large nominal vocabulary-factor budget as the explanation for the input-side result. It evaluates two distinct ways to spend a similar budget: row-local vocabulary factors and a higher-rank transform shared through pretrained rows. Their rank and initialization still differ. Retained generation, reported on the same checkpoints, makes the accompanying adaptation tradeoff explicit.

Table 3: Separately tuned input-only near-budget control on Qwen3-0.6B/CLUENER. Means over five fresh seeds. Gen: SciQ generated-answer accuracy on the fixed 64 questions/all rotations; Pool: candidate pooling over 1,000 questions.

Method	Rank	Parameters	Source F1	SciQ Pool	SciQ Gen
A-LoRA	19	38,912	47.37	82.78	65.31
Active-row W-form	4	39,708	26.67	86.24	79.14

5.3 Sharing Improves the Use of a Fixed Parameter Budget

Shared A-LoRA increases CLUENER F1 from 45.06 to 50.90 and WikiSQL execution accuracy from 54.90 to 59.14 relative to input-only adaptation. Both use 32,768 parameters, rank 16, selected learning-rate multiplier 16, and the same task-specific seeds. Input-factor initialization and training order are identical in all ten paired runs. Sharing wins in all five CLUENER seeds and four of five WikiSQL seeds (Figure 1). The mean paired gains are 5.83 and 4.24 points; the corresponding ranges are 3.59–7.46 and -0.29–12.01. These comparisons isolate the use of the additional output path at fixed trainable budget and selected rate.

Independent bilateral adapters use 65,536 parameters and achieve means of 49.09 and 55.78. Shared adaptation therefore reaches higher means with half the parameters. On WikiSQL, independent bilateral performance ranges from 46.97 to 60.94, whereas sharing ranges from 57.23 to 61.91; the latter's higher mean accompanies a narrower observed seed range. The paired independent/shared

comparison is positive in five CLUENER seeds and two WikiSQL seeds, emphasizing the full distribution rather than a universal run-level ordering.

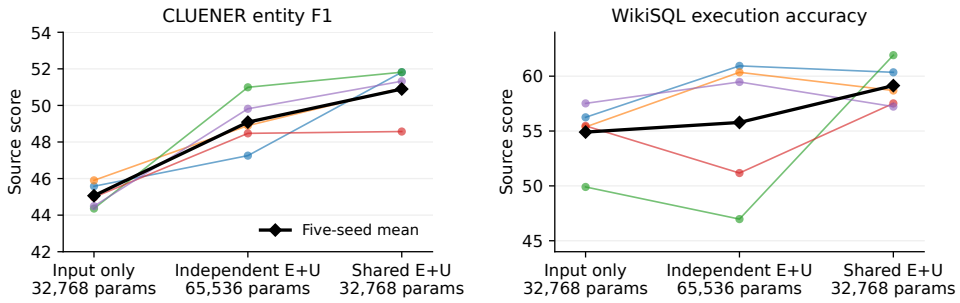

Figure 1: A-LoRA source performance across all five fresh seeds on Qwen3-0.6B. Thin lines connect the same seed across placements; black diamonds are means. Sharing uses the same parameter count as input-only and half that of independent bilateral adaptation. All shown configurations use rank 16 per transform and the same selected learning rate within each task.

5.4 Selecting Placement for Adaptation and Retention

Table 4: SciQ retention for the same five-seed checkpoints. Pool uses 1,000 questions; candidate and generation columns marked 64 use identical questions and rotations. The last column changes the source training task. Generation uses the default EOS rule and a 32-token cap.

		After CLUENER training			After WikiSQL
Placement	Method	Pool_{1000}	Candidate_{64}	Generation_{64}	Generation_{64}
Frozen base	–	86.80	83.98	83.20	83.20
E	A	84.10	74.06	70.70	66.02
E	W	86.22	81.41	79.92	80.23
U	A	87.32	70.39	21.95	83.20
U	W	86.90	79.06	76.48	83.44
Independent $E + U$	A	85.62	70.62	41.17	35.31
Independent $E + U$	W	86.24	77.81	71.09	82.73
Shared $E + U$	A	85.48	67.27	66.09	63.98
Shared $E + U$	W	79.12	58.28	40.08	83.12

Table 4 adds retained generation to the placement comparison. After CLUENER, shared A-LoRA improves both source F1 (50.90 versus 16.51) and SciQ generated-answer accuracy (66.09 versus 40.08) over dense W-form sharing. Within A-LoRA, input-only adaptation retains more SciQ generation accuracy (70.70) while shared adaptation achieves the higher source score. On WikiSQL, the corresponding shared-generation scores are 63.98 for A-LoRA and 83.12 for W-form, accompanying their different source scores of 59.14 and 5.51. These paired outcomes expose the adaptation–retention tradeoff that a source-only ranking leaves unresolved.

Output-only adaptation highlights a distinct evaluation issue. After CLUENER, A-LoRA has SciQ pooled candidate accuracy 87.32, compared with the base's 86.80, but generated-answer accuracy falls to 21.95 on the fixed 64-question diagnostic. Restricted candidate accuracy on those same 64 questions is 70.39; W-form output-only generation scores 76.48. Candidate ranking and generated answers therefore capture different aspects of the adapted model.

For CLUENER-trained A-LoRA, independent bilateral adapters have an 18.24% first-token EOS rate on the full 1,000-question SciQ set. Yet immediate EOS is only 1.33% for output-only A-LoRA, despite its lower generation score. Saved outputs include repeated source-format fragments, and requiring one non-EOS token changes the latter score only from 21.95 to 22.34. Premature stopping and repetitive or misformatted generation are distinct failure modes. ANLI results and all decoding conditions are reported in Appendix A.1.

5.5 Readout Analysis across Model Sizes

The fixed-rate historical study evaluates 45 output-only A-LoRA checkpoints across eight Qwen2.5/Qwen3 sizes, on both source tasks, at rank 16 and learning rate 2×10^{-4}. Full-vocabulary readouts reproduce saved candidate scores exactly. All eight bases have zero first-token EOS, but Qwen3-4B CLUENER adapters average 95.82% on SciQ and 92.41% on ANLI; Qwen3-8B averages 90.68% and 19.07%. The same 8B model after WikiSQL has zero on both. Source task and adaptation setting therefore matter alongside model size.

On the fixed 64-question SciQ subset, requiring one non-EOS token raises three historical 8B CLUENER adapters from 35.94/1.17/2.34 to 95.70/95.70/95.31; the base remains 95.70. This intervention identifies early stopping as the dominant failure in that setting. It does not recover the fresh 0.6B output-only adapters, separating stopping from repetitive generation. All models, tasks, seeds, and intervention results are reported in Appendix A.1.

5.6 Candidate Contrasts and Stopping Margins Can Decouple

For output-only adaptation, the hidden vector h at a fixed prefix is unchanged. The A-LoRA logit perturbation is $\delta z = sU_0PQh$. Let $\mathcal{C}$ contain the answer-letter tokens and $\bar{\delta}_{\mathcal{C}}$ be their mean logit perturbation. Candidate ranking depends on the base candidate logits and the centered perturbations $\delta z_a - \bar{\delta}_{\mathcal{C}}$; their common shift cancels, whereas stopping also depends on the EOS margin. In particular,

$$\delta(z_{\mathrm{EOS}} - z_a) = s\big(U_0[\mathrm{EOS}, :] - U_0[a, :]\big)PQh. \tag{10}$$

A small change among candidate contrasts therefore need not imply a small change in stopping propensity. This decomposition provides a direct diagnostic of output-side adaptation at a fixed prefix.

For historical Qwen3-4B CLUENER adapters on SciQ, the mean EOS shift relative to the candidate mean is 3.73 logits, while the mean RMS centered candidate perturbation is 0.18. The base EOS-minus-best-candidate margin averages -2.44 and becomes $+1.29$. Direct full-vocabulary evaluation, rather than this mean alone, confirms the 95.82% immediate-EOS rate. Checks of the historical training masks find no padding tokens incorrectly supervised as EOS. Appendix Figure 3 visualizes both logit shifts and the decoding intervention.

Design implications. The results connect three design choices: how the update uses pretrained vocabulary geometry, where the transform acts, and whether both boundaries share parameters. Shared A-LoRA is a particularly effective source-task configuration in the completed two-task comparison; input-only adaptation offers a different retention tradeoff, while output-only adaptation makes readout calibration central. Candidate-scoring blind spots have been identified previously (Wang et al., 2024); our analysis connects them to boundary placement and measured changes in stopping margins. Historical hidden-LoRA and stacked-adapter comparisons, including their mixed retention outcomes, are reported in Appendix A.1.

6 Related Work

Low-rank parameterization and placement. LoRA learns low-rank additive updates to frozen weights (Hu et al., 2022); QLoRA and AdaLoRA address quantized training and adaptive rank allocation (Dettmers et al., 2023; Zhang et al., 2023). PEFT implementations support a range of target modules (Mangrulkar et al., 2022; Hugging Face, 2026a). Multiplicative adaptation is also studied by LoRMA (Bihany et al., 2025). Our focus is the asymmetric $V \times d$ vocabulary matrix, where hidden-space factors avoid vocabulary-dependent trainable cost, and the interaction of this structure with input/output placement and sharing. Appendix A.13 gives the algebraic relation.

Vocabulary adaptation and tying. Selective token training addresses new or task-specific vocabulary (Hugging Face, 2026b). A-LoRA instead induces coupled updates across pretrained vocabulary rows. Weight tying shares token embeddings and classifiers (Press & Wolf, 2017; Inan et al., 2017); we study whether the *adaptation* should remain shared, be unilateral, or become independent. Concurrent anonymous work on architectural untying concerns pretraining (Anonymous, 2026), whereas our comparisons freeze the backbone and vary post-training adapter topology.

Evaluating retained behavior. Candidate probabilities and generated answers can disagree in instruction-tuned models (Wang et al., 2024). We connect this distinction to measured boundary-adapter perturbations: a common shift among candidate logits cancels in their comparison but can change their margins against EOS or other vocabulary tokens. The resulting readout diagnostics complement task accuracy and parameter-count comparisons.

7 Conclusion

A-LoRA uses pretrained vocabulary structure to make input and output adaptation independent of vocabulary size in trainable parameter count. Its shared form activates two computational paths with one hidden-space transform and preserves a mergeable tied matrix. Five-seed experiments demonstrate substantial boundary-only task learning and a same-budget advantage from sharing over input-only adaptation on two tasks. Geometry, active-row accounting, and retained-generation analysis explain different aspects of this design: which structure is reused, where the budget is spent, and how adaptation changes model behavior. Vocabulary-boundary parameterization and sharing are therefore productive design choices for parameter-efficient post-training.

8 Limitations

The fresh-seed primary study covers Qwen3-0.6B, two structured tasks, 2,048 training examples, and 64 steps. Its finite learning-rate search leaves several optima at the search boundary. A-LoRA and dense W-form comparisons differ in rank, budget, and initialization; active-row counts separately characterize input capacity. Source performance is not matched, so retention differences combine adaptation strength and parameterization. The same-budget shared/E-only comparison controls rank, selected rate, and count; shared/independent comparisons change count.

Generation diagnostics use 64 fixed questions, all answer rotations, a 32-token cap, and a conservative parser. Seed ranges are descriptive, rotations are correlated, and ANLI base performance is near chance. These measurements characterize retained answer generation rather than long-form quality. Public datasets may have appeared in pretraining. Weight-space fits are descriptive and include affine shifts that can be softmax-invariant on the output side. Historical fixed-rate and stacked studies have separate protocols; the former has no tuned W-form control. Extensions to larger models, longer training, and source-matched retention comparisons remain open.

AI USE STATEMENT

Generative AI tools assisted with implementation and debugging of experimental code, discussion of experimental design and interpretation of results, and translation and editing of manuscript material. They also assisted with organizing research materials and adapting the manuscript's LaTeX formatting. The authors are responsible for the final text, claims, experimental results, and accompanying artifacts.

ETHICS STATEMENT

This work performs method analysis and empirical evaluation on publicly available pretrained models and common downstream task datasets. We use public models, tokenizers, and datasets according to their released licenses and terms of use; we do not redistribute the original datasets or pretrained model weights. It does not involve human subjects or personally sensitive data. To the authors' knowledge, there are no additional ethical risks requiring disclosure.

REPRODUCIBILITY STATEMENT

§2 specifies parameterizations, parameter counts, and effective sharing. §4 and Appendix A.1 give the new training, selection, scoring, and seed protocols; the remaining appendix separately documents historical instruction-loss experiments and weight-space fits. Experimental artifacts retain frozen configurations, input and output hashes, raw generated responses, and audits of parameter freezing, loss masking, reload equivalence, and source/retention scores. The manuscript reports only completed comparison blocks, and does not present development selection scores as held-out test results.

The anonymous supplementary archive includes the complete manuscript source, per-seed tables, 94,680 source-example score records, 162,000 retention-question records, and 72,576 generated-answer records for the primary study and frozen base. CPU scripts independently recompute reported aggregates and answer-parser outcomes. Exact adapter components and numerical checks accompany the structural derivations; external model weights and task corpora are required to rerun model training or inference.

REFERENCES

AI@Meta. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Anonymous. Efficient untying of weight-tied language models via asymmetric geometric decoupling, 2026. Under review.

Harsh Bihany, Shubham Patel, and Ashutosh Modi. LoRMA: Low-rank multiplicative adaptation for LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10113–10133, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.527. URL https://aclanthology.org/2025.findings-acl.527/.

DeepSeek-AI. DeepSeek-V3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. DeepSeek-V3.1 model release. https://huggingface.co/deepseek-ai/DeepSeek-V3.1, 2025. Official model card; Accessed 2026-05-17.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

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.

Hugging Face. LoRA: PEFT developer guide. https://huggingface.co/docs/peft/en/developer_guides/lora, 2026a. Accessed 2026-05-17.

Hugging Face. Trainable tokens: PEFT documentation. https://huggingface.co/docs/peft/package_reference/trainable_tokens, 2026b. Accessed 2026-05-17.

Hakan Inan, Kevin Khassanchi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1A3CgZAb.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022. Software library.

Meta. Llama 3.2 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md, 2024. Official model card; Accessed 2026-05-17.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, 2020. URL https://aclanthology.org/2020.acl-main.441/.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, Valencia, Spain, 2017. Association for Computational Linguistics. doi: 10.18653/v1/E17-2025. URL https://aclanthology.org/E17-2025/.

Qwen Team. Qwen2.5 technical report, 2025a. URL https://arxiv.org/abs/2412.15115.

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

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Official model-family release and citation; Accessed 2026-05-17.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. URL https://arxiv.org/abs/1909.08053.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My Answer is C": First-token probabilities do not match text answers in instruction-tuned language models. *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://arxiv.org/abs/2402.14499.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017. URL https://aclanthology.org/W17-4413/.

Liang Xu, Yu Tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, Caiquan Liu, and Xuanwei Zhang. CLUENER2020: Fine-grained named entity recognition dataset and benchmark for chinese. *arXiv preprint arXiv:2001.04351*, 2020. URL https://arxiv.org/abs/2001.04351.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*, 2017. URL https://arxiv.org/abs/1709.00103.

A SUPPLEMENTARY MATERIAL

This appendix supplements experimental setup and per-model values for §3, §2, and §4, for separate consultation of post-training vector-space static analysis, A-LoRA budget probes, and downstream LoRA post-training experiments.

A.1 FRESH-SEED BOUNDARY-ONLY CONFIRMATION

This subsection documents the new structured-task comparison. It is distinct from the historical instruction-loss experiments later in this appendix. All 80 adapters below complete source and retention audits. The CLUENER fresh seeds are 260926100–260926104 and WikiSQL seeds are 260926200–260926204; the paired methods use the same task-specific seeds. Learning rates are chosen using two different development seeds, not these held-out runs. Scores and ranges below are descriptive; rotations are not independent training replicates.

Initialization and optimization. Both methods start with an exactly zero residual and scale $s = 8$. The frozen vocabulary matrix is retained. The zero factor in ordinary input LoRA is vocabulary-indexed, allowing exact active-row compaction for this input-only setting. Optimizer settings are AdamW $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay, gradient norm clip 1, global batch 32, and 64 steps. CLUENER/WikiSQL microbatches are 2/4. Loss is the mean over target tokens, including native EOS, with prompt and padding labels masked. Source decoding is greedy, with 512/256 new-token caps and batches 32/16 for CLUENER/WikiSQL. The retained task prompts and answer-order rotations are fixed across methods.

Evaluation definitions. For a question with K answer choices, each of its K cyclic rotations produces candidate log probabilities $\ell_{j,k}$. Within each rotation, we normalize over candidates: $p_{j,k} = \exp(\ell_{j,k}) / \sum_{a=1}^{K} \exp(\ell_{j,a})$. After mapping each rotated letter to its semantic choice, pooled scoring maximizes $K^{-1} \sum_j p_{j,k}$. Single-rotation scoring makes each choice before averaging correctness. Generation is unrestricted over the vocabulary and uses a fixed answer parser; unparsed outputs score zero. Immediate EOS is measured as the actual full-vocabulary first-token argmax, not as a candidate-restricted probability. A 32-token cap is a diagnostic limit: reaching it is not itself an error, since an answer may already have been stated. The 64 diagnostic question IDs were fixed before these new confirmation runs; all methods use the same subset.

Table 5: Selected learning-rate multiplier relative to 2×10^{-4}, source means, and five-seed ranges. A = A-LoRA rank 16; W = dense vocabulary LoRA rank one. CLUENER uses entity F1; WikiSQL uses execution accuracy. Several selected multipliers remain at the finite search boundary.

Task	Placement	Method	LR multiplier	Mean	Seed range
CLUENER	E	A	16	45.06	44.36–45.90
CLUENER	E	W	4	11.29	5.05–17.04
CLUENER	Independent $E + U$	A	16	49.09	47.25–50.99
CLUENER	Independent $E + U$	W	4	20.28	18.07–22.98
CLUENER	Shared $E + U$	A	16	50.90	48.57–51.82
CLUENER	Shared $E + U$	W	1	16.51	15.20–19.71
CLUENER	U	A	16	36.46	33.54–37.79
CLUENER	U	W	16	17.00	15.09–18.49
WikiSQL	E	A	16	54.90	49.90–57.52
WikiSQL	E	W	1	5.18	0.98–8.01
WikiSQL	Independent $E + U$	A	16	55.78	46.97–60.94
WikiSQL	Independent $E + U$	W	0.25	4.71	1.56–7.23
WikiSQL	Shared $E + U$	A	16	59.14	57.23–61.91
WikiSQL	Shared $E + U$	W	0.015625	5.51	5.27–5.96
WikiSQL	U	A	0.0625	5.78	5.47–6.05
WikiSQL	U	W	0.25	5.47	5.27–5.66

Budget scope. Dense rank-one W-form has 152,960 parameters per side versus 32,768 for rank-16 A-LoRA; independent bilateral counts double. Input-only W-form can be compacted to 9,927

Table 6: SciQ retention after CLUENER training. All values are percentages, averaged over five seeds. Pool/single/EOS use 1,000 questions; the final three columns use the same 64-question subset and all rotations. Min1 requires at least one non-EOS token.

Placement	Method	Pool	Single	EOS	Single_{64}	Gen_{64}	Min1_{64}
Base	–	86.80	85.33	0.00	83.98	83.20	83.20
E	A	84.10	79.95	0.00	74.06	70.70	70.70
E	W	86.22	83.99	0.00	81.41	79.92	79.92
Independent $E + U$	A	85.62	78.06	18.24	70.62	41.17	45.78
Independent $E + U$	W	86.24	81.16	0.00	77.81	71.09	71.09
Shared $E + U$	A	85.48	71.97	0.23	67.27	66.09	66.17
Shared $E + U$	W	79.12	59.97	0.01	58.28	40.08	40.08
U	A	87.32	73.66	1.33	70.39	21.95	22.34
U	W	86.90	80.61	0.01	79.06	76.48	76.48

Table 7: SciQ retention after WikiSQL training. All values are percentages, averaged over five seeds. Pool/single/EOS use 1,000 questions; the final three columns use the same 64-question subset and all rotations. Min1 requires at least one non-EOS token.

Placement	Method	Pool	Single	EOS	Single_{64}	Gen_{64}	Min1_{64}
Base	–	86.80	85.33	0.00	83.98	83.20	83.20
E	A	82.42	77.18	1.75	70.23	66.02	67.42
E	W	87.14	84.30	0.00	81.02	80.23	80.23
Independent $E + U$	A	81.82	72.70	12.39	66.33	35.31	40.31
Independent $E + U$	W	86.60	84.95	0.00	83.28	82.73	82.73
Shared $E + U$	A	85.04	75.66	1.60	70.23	63.98	64.53
Shared $E + U$	W	86.84	85.23	0.00	83.75	83.12	83.12
U	A	86.80	85.33	0.00	83.98	83.20	83.20
U	W	86.80	85.30	0.00	84.22	83.44	83.44

Table 8: ANLI round 1 retention after CLUENER training. All values are percentages, averaged over five seeds. Pool/single/EOS use 1,000 questions; the final three columns use the same 64-question subset and all rotations. Min1 requires at least one non-EOS token.

Placement	Method	Pool	Single	EOS	Single_{64}	Gen_{64}	Min1_{64}
Base	–	34.60	33.63	0.00	32.29	30.21	30.21
E	A	32.54	32.23	0.00	35.83	33.02	33.02
E	W	33.56	32.80	0.01	33.75	31.56	31.56
Independent $E + U$	A	32.44	32.79	22.19	34.27	19.17	24.17
Independent $E + U$	W	32.62	32.67	1.85	33.85	32.81	33.33
Shared $E + U$	A	33.10	32.66	4.78	33.33	24.58	26.15
Shared $E + U$	W	32.00	32.50	2.93	35.10	12.50	13.65
U	A	34.52	33.17	13.49	33.12	14.48	14.48
U	W	33.96	33.08	0.00	32.81	21.77	21.77

Table 9: ANLI round 1 retention after WikiSQL training. All values are percentages, averaged over five seeds. Pool/single/EOS use 1,000 questions; the final three columns use the same 64-question subset and all rotations. Min1 requires at least one non-EOS token.

Placement	Method	Pool	Single	EOS	Single_{64}	Gen_{64}	Min1_{64}
Base	–	34.60	33.63	0.00	32.29	30.21	30.21
E	A	32.20	32.11	19.69	36.56	27.19	33.12
E	W	34.00	33.16	0.00	36.04	33.85	33.85
Independent $E + U$	A	31.62	32.59	34.91	35.83	16.15	20.52
Independent $E + U$	W	34.98	33.78	0.00	33.85	30.31	30.31
Shared $E + U$	A	33.38	33.13	16.23	35.00	24.27	29.38
Shared $E + U$	W	34.58	33.75	0.00	32.92	31.04	31.04
U	A	34.58	33.61	0.00	32.29	30.21	30.21
U	W	34.58	33.71	0.00	32.29	30.21	30.21

Table 10: WikiSQL secondary metrics. LF is normalized logical-form exact match; empty denotes execution returning an empty list, not necessarily an error. Distinct EX checks distinct SQL placeholders for repeated-column conditions, in addition to the primary evaluator that follows the official DBEngine convention.

Placement	Method	LF	Valid query	Empty	Distinct EX
E	A	46.02	98.67	24.28	54.90
E	W	1.88	70.80	30.41	5.20
Independent $E + U$	A	47.60	97.97	23.57	55.76
Independent $E + U$	W	0.27	72.79	15.88	4.71
Shared $E + U$	A	50.10	98.75	21.02	59.14
Shared $E + U$	W	0.39	77.07	15.72	5.47
U	A	0.45	86.45	21.31	5.64
U	W	0.41	76.33	12.27	5.47

Table 11: Post-hoc CLUENER content check, restricted to the intersection of schema-valid outputs for each paired seed. Coverage is the mean fraction of held-out questions retained. The post-training selection of this subset prevents a causal format/content decomposition.

Placement	Coverage (%)	A entity F1	W entity F1
E	42.95	48.41	17.83
Independent $E + U$	75.19	52.52	24.04
Shared $E + U$	73.30	52.95	19.41
U	75.31	40.02	20.07

active-row rank-one parameters for the CLUENER training vocabulary. An independently tuned rank-four active-row control has 39,708 parameters versus 38,912 for rank-19 A-LoRA. The selected multipliers are 1 for active-row W-form and 16 for A-LoRA, chosen from the same fixed five-point grid using two development seeds (development means 26.96 and 42.66). The completed five fresh-seed means are 26.67 and 47.37 F1; per-seed source and retention CSVs accompany the supplement. The A-LoRA selection remains at the finite grid boundary.

Table 12: Complete retention means for the five-seed near-budget input control after CLUENER training. Pool/single/EOS use 1,000 questions; default/min1 generation uses the fixed 64-question subset and all rotations.

Task	Method	Pool	Single	EOS	Gen	Min1
SciQ	A-LoRA	82.78	77.20	0.01	65.31	65.31
SciQ	Active W	86.24	84.06	0.00	79.14	79.14
ANLI	A-LoRA	32.68	32.66	0.00	36.35	36.35
ANLI	Active W	33.72	33.99	0.00	30.62	30.62

Historical protocol boundaries. The historical output-only readout study below uses rank 16 and a fixed learning rate 2×10^{-4}, rather than the source-selected rates in the new comparison. It includes 45 adapters and eight bases, with SciQ and ANLI each evaluated on 1,000 questions and all rotations. All base EOS rates are zero. The saved candidate log probabilities match the original evaluation exactly. These full-vocabulary first-token measurements do not substitute for full generation, and there is no matched ordinary-LoRA comparator in this historical study.

Decoding intervention and its limit. For three historical Qwen3-8B CLUENER checkpoints, the same fixed 64 SciQ questions yield default generated-answer scores 35.938/1.172/2.344, versus 95.703/95.703/95.313 with one required non-EOS token. The base is 95.703 under both. ANLI scores are 67.188/44.271/59.375 by default and 67.188/64.063/67.188 with the intervention (base 65.104). All methods use the same rule, and default failures remain reported. This selected diagnostic was motivated by observed behavior and is not a prespecified general repair. In contrast, the newly confirmed Qwen3-0.6B CLUENER output-only A-LoRA changes only from 21.953 to 22.344 on SciQ, so disabling immediate EOS is insufficient for its repetitive outputs.

Table 13: Complete historical output-only A-LoRA first-token EOS matrix. Both training tasks and all eight models are retained, including zero-rate cells. Rates are averaged over the available seeds; the one/two-seed exceptions are explicit.

Model	Source task	Seeds	SciQ EOS (%)	ANLI EOS (%)
Qwen2.5-0.5B	CLUENER	3	0.000	0.000
Qwen2.5-0.5B	WikiSQL	3	0.000	0.000
Qwen2.5-1.5B	CLUENER	3	0.000	0.000
Qwen2.5-1.5B	WikiSQL	2	0.000	0.000
Qwen2.5-3B	CLUENER	3	0.000	0.000
Qwen2.5-3B	WikiSQL	1	0.025	0.000
Qwen2.5-7B	CLUENER	3	0.067	0.000
Qwen2.5-7B	WikiSQL	3	0.667	0.033
Qwen3-0.6B	CLUENER	3	0.000	0.000
Qwen3-0.6B	WikiSQL	3	0.000	0.000
Qwen3-1.7B	CLUENER	3	0.708	28.611
Qwen3-1.7B	WikiSQL	3	0.000	0.000
Qwen3-4B	CLUENER	3	95.817	92.411
Qwen3-4B	WikiSQL	3	48.175	46.156
Qwen3-8B	CLUENER	3	90.675	19.067
Qwen3-8B	WikiSQL	3	0.000	0.000

Learning-rate readout analysis (development checkpoints). A post-hoc CPU float64 analysis covers all 80 available output-only development checkpoints from the two-model, two-source learning-rate study. It uses frozen-prefix hidden representations, the EOS row, and candidate rows. Agreement with checked full-forward runs has maximum relative logit error 7.12×10^{-5} and no meaningful accuracy discrepancy. The fraction for which EOS exceeds every answer letter is only a candidate-comparison diagnostic: it is not the full-vocabulary EOS rate, because other tokens may have larger logits. No new generation is inferred from this calculation.

For Qwen3-0.6B CLUENER A-LoRA, increasing the learning-rate multiplier from 1 to 4 to 16 raises two-seed source-development F1 from 22.16 to 34.76 to 38.80, while SciQ single-rotation accuracy changes from 85.28 to 82.34 to 53.78; pooled scores remain 87.05, 87.40, and 87.55. For Qwen2.5-7B CLUENER A-LoRA at multiplier 4, SciQ single-rotation accuracy is still 94.20 while EOS exceeds all answer letters on 71.90% of prefixes. Ordinary output LoRA at its selected multiplier 16 shows 94.28 accuracy and a 92.53% EOS-versus-letters fraction. This supports an optimization/readout tradeoff worth testing, not a claim of A-LoRA-specific instability or a verified low-rate generation repair.

Historical stacked-adapter retention. An additional completed matrix evaluates 300 existing adapters: hidden LoRA alone, hidden plus A-LoRA, and hidden plus ordinary boundary LoRA, across 20 model–source-task cells and five seeds. Those runs use a different joint-training protocol. A-LoRA minus ordinary-LoRA pooled retention is positive in 11/20 cells and negative in 9/20 for each of SciQ and ANLI; no positive comparison survives Holm correction over the 40 A-LoRA-versus-ordinary comparisons (20 source–model cells by two retention tasks), using two-sided paired seed-level t-tests. One negative ANLI comparison, Qwen3-8B after CLUENER, has mean difference -1.32 points and Holm-adjusted $p = 0.0015$. These results do not support a universal retention advantage from stacking A-LoRA. A further descriptive rescore shows why pooling matters: for Qwen3-0.6B trained on Emotion, stacked A-LoRA versus ordinary LoRA has SciQ pooled accuracy 83.04 versus 81.32, but single-rotation accuracy 62.35 versus 75.01. Neither rescore is a new generation test.

A.2 Structural Properties of the Shared Transform

Vocabulary-row coupling. Write $R = sPQ$ and $\Delta M = M_0 R$. For any c with $c^\top M_0 = 0$, associativity gives $c^\top \Delta M = (c^\top M_0)R = 0$. More generally, if a row is represented as $M_0[i,:] = \sum_j a_j M_0[j,:]$, its update obeys the same relation. This statement holds for every learned P, Q; it does not require a small update or an invertible transform. It also does not imply preservation of angles or pairwise distances. At fixed rank this is a structural restriction relative to unrestricted BC,

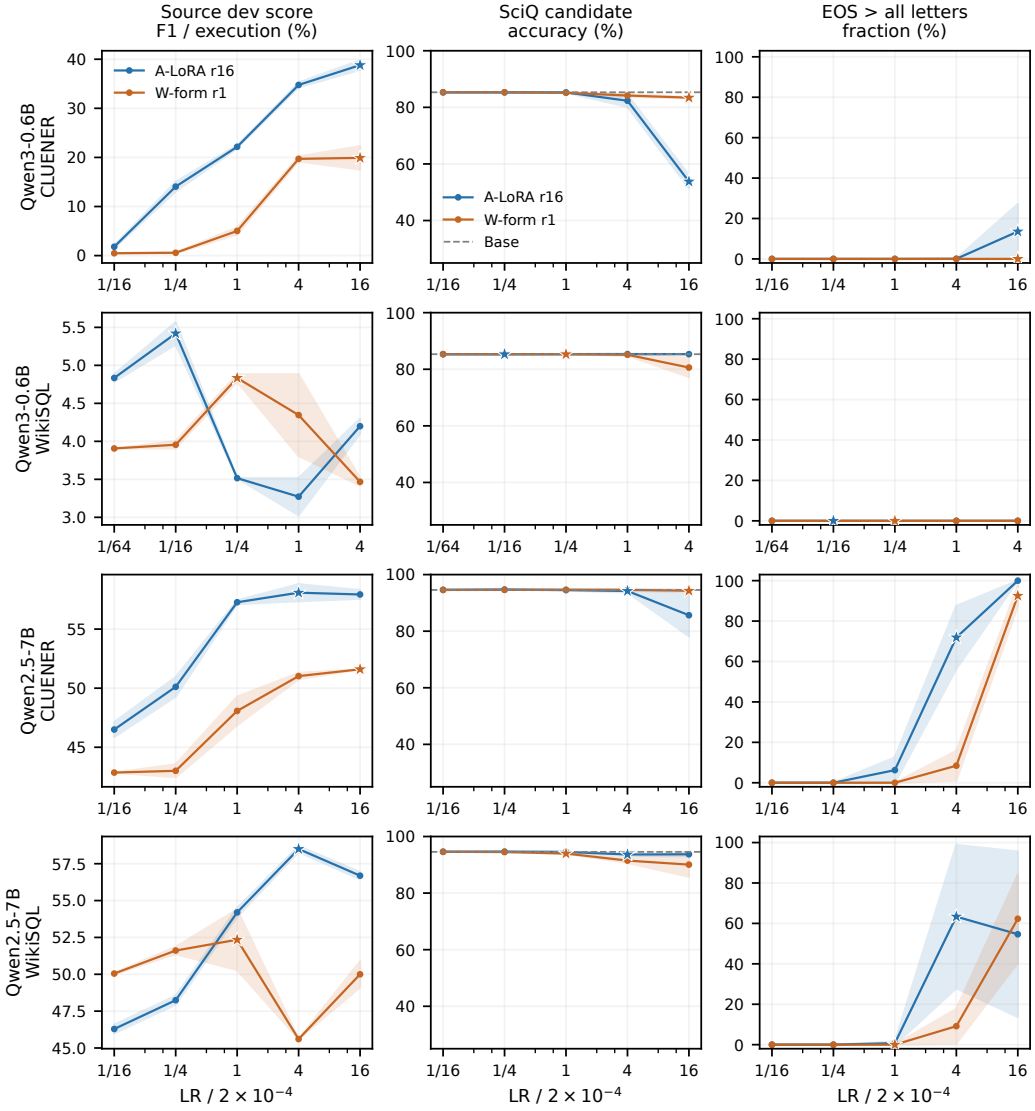

Figure 2: Complete output-only development learning-rate readout grid for the four studied model–task pairs (80 checkpoints, two seeds per point). Shading shows seed minima/maxima, not confidence intervals; stars mark frozen source-development selections. The right column compares EOS only with answer letters and must not be read as the actual full-vocabulary EOS rate. This post-hoc readout analysis performs no new generation and does not select on retention labels.

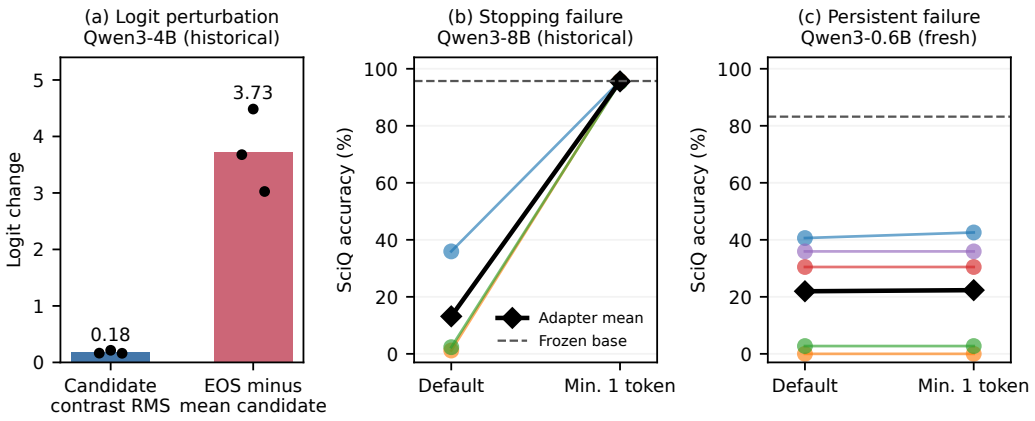

Figure 3: Different readout failures require different diagnoses. (a) Historical Qwen3-4B output-only adapters: three seed values and means of SciQ logit-change summaries, using 1,000 questions and all rotations. (b) The same minimum-length intervention largely restores a selected historical Qwen3-8B case. (c) It does not restore the newly confirmed Qwen3-0.6B case. Panels (b,c) use actual generation on the same fixed 64 questions, all rotations, and a 32-token cap; thin lines are seeds and black diamonds are means. Historical and fresh learning-rate protocols differ.

and the budget advantage comes from replacing V-dimensional trainable factors with d-dimensional ones.

Two paths under tied sharing. For a tied base, token embeddings are columns of X and $z(T) = W_0 T F(T^\top X)$. For $T = I + \epsilon R$ and differentiable F, the product and chain rules give

$$z(I + \epsilon R) = W_0 h_0 + \epsilon W_0 \{R h_0 + J_F(X)[R^\top X]\} + o(\epsilon).$$

The first term inside braces is the direct readout path; the second is the propagated input path. Sharing constrains both to the same direction R, whereas independent adapters can choose different directions. Either path can help or oppose the other for a particular objective. The five-seed comparison measures their net effect; we do not infer gradient alignment from the source scores.

Exact active-row input compaction. In W-form input LoRA, lookup touches $B[x, :]$ only for observed token IDs. With B initialized to zero, no weight decay, and zero initial optimizer moments, rows outside the training-defined set S remain zero. Retaining $B[S, :]$ and the common factor C therefore preserves the embedding function for every token, with zero update outside S. A-LoRA instead computes $M_0[x, :]PQ$ for any token, so training on S can induce changes elsewhere through the same transform. This difference concerns the support and sharing of learned updates; it does not establish a benefit on unseen tokens without a separate evaluation.

Candidate and full-vocabulary decisions. For a fixed prefix and candidate set $\mathcal{C}$, adding a common constant to $\{z_a : a \in \mathcal{C}\}$ leaves its restricted softmax and ordering unchanged. It can still change $z_{\mathrm{EOS}} - z_a$ and other non-candidate margins. For output-only A-LoRA, $\delta z = s U_0 P Q h$, yielding the EOS-margin expression in the main text. A margin against answer letters alone is not the full-vocabulary argmax; the latter is measured separately. The supplied numerical checks verify row coupling, merged/unmerged equivalence, the first-order expansion on a smooth toy backbone, and exact active-row compaction.

A.3 Parameter Budget Matching

A nominal dense budget can be approximately matched using the following rank rule. This algebra does not mean every experiment below uses matched budgets. Given Vocab rank r_W, take affine rank

$$r_{\mathrm{aff}}(r_W) = \left\lfloor r_W \cdot \tfrac{V+d}{2d} \right\rfloor \approx \left\lfloor r_W \cdot \tfrac{V}{2d} \right\rfloor. \tag{11}$$

For example, at $V/d \approx 31$, Vocab rank-1 corresponds to aff rank ≈ 15. Internal-layer baselines can likewise allocate rank under a specified total budget; details appear in Appendices A.12 and A.15.

The new main comparison reports actual unequal counts. Historical input-only versus bilateral W-form comparisons also differ in topology, and do not isolate parameterization. Active-row input accounting is treated separately in §2.3.

A.4 Validation Paths

The historical instruction-loss experiments were organized around the four paths in Appendix Table 15, rather than enumerating all adapter combinations. The first two paths test parameter efficiency and the adaptation value of vocabulary-boundary layers themselves; the last two distinguish standalone from hidden-LoRA-conditioned regimes and check shared mergeable form on tied models.

The historical variants include: hidden LoRA, vocab-only LoRA, A-LoRA-only input/`lm_head`/bilateral, hidden + input/`lm_head` A-LoRA, and mergeable tied A-LoRA. Except for full fine-tune sanity probe and frozen base, all experiments freeze base weights and train only the corresponding adapters; full definitions appear in Table 22.

A.5 Topology and Validation-Path Summaries

The following tables give the complete topology and validation-path summaries referenced in §2.2 and §A.4.

Table 14: Four vocabulary-boundary adapter topologies.

ID	Name	E_{eff}	U_{eff}	Role
T1	Input-only	$E_0+\Delta_E$	U_0	Independent input-side capacity
T2	Output-only	E_0	$U_0+\Delta_U$	Output-side structural hypothesis
T3	Shared	$W_0+\Delta$		Tied shared-matrix hypothesis
T4	Decoupled	$E_0+\Delta_E$	$U_0+\Delta_U$	Bilateral complement / redundancy

Table 15: How downstream experiments correspond to prior analysis and method design.

Question	Main comparison	Results
Can boundary-only adapters learn the source task?	Frozen base, A-LoRA, and W-form across four placements	§5.1
What parameter budget is actually active?	Nominal dense counts and training-defined active input rows	§2.3
Are unrelated-task answers preserved?	Candidate scores, default generation, and uniform minimum-length decoding	§5; Appendix A.1
Can candidate retention hide readout changes?	Full-vocabulary EOS and fixed-prefix logit decomposition	§5.5

A.6 NOTATION AND MAIN-TEXT CORRESPONDENCE

The main text uses V for vocabulary size and d for hidden dimension. Base/Instruct weights are denoted X and Y; $\Delta W = Y - X$ characterizes post-training change of vocabulary row vectors in hidden space. A-LoRA low-rank factors P, Q and Vocab LoRA factors B, C are independent; β is A-LoRA's hidden-dimensional bias.

Table 16: Correspondence between main-text objects and appendix experiment tables.

Main text	Question	Appendix section / table
§3 Table 1	Aggregate R^2_K, $A_{L,K}$	Table 18; full data in Appendix A.18
§3.2	aff/W equal budget	Table 19; full data in Appendix A.18
§2.2 Table 14	T1–T4 definitions	Appendix A.11
§A.4	Downstream fine-tuning	Appendix A.15

A.7 BI MODEL PAIRS AND EXCLUSION CRITERIA

Inclusion criteria: auditable public Base→Instruct weight pairs; full `embed_tokens` and `lm_head` (or tied shared) weights readable; V, d consistent after vocabulary alignment.

Main analysis group: non-excluded **30 pairs** = main 26 + extended 4 (DeepSeek-V3/V3.1, Qwen3-30B-A3B, Qwen3.5-35B-A3B), of which 17 tied and 13 untied. Full registry has 35 pairs; 5 excluded pairs appear in Appendix A.8. Table 25 lists Base/Instruct names, tier, and tied status for all 30 pairs. Model families include Qwen2.5/Qwen3/Qwen3.5 (Qwen Team, 2025a;b; 2026), Llama-3.x (AI@Meta, 2024; Meta, 2024), Gemma-2/3 (Gemma Team, 2024; 2025), and DeepSeek-V3/V3.1 (DeepSeek-AI, 2024; 2025).

Table 17: 30-pair BI main analysis group summarized by model family (n = pairs included from that family).

Family	n	tied	untied	Notes
Qwen3 / Qwen3.5	11	7	4	includes 2 MoE pairs
Qwen2.5	7	3	4	
Llama-3.x	4	2	2	
Gemma-2 / Gemma-3	6	6	0	
DeepSeek-V3 series	2	0	2	extended
Total	30	17	13	

A.8 EXCLUDED MODELS

The following 5 pairs are **not included** in the 30-pair main narrative:

- **Gemma-4** (4 pairs): row-norm gauge in weight files (Base E near-unit-sphere) + Per-Layer Embeddings (PLE) architecture, so raw E/U differ in convention from standard models; full-vocab affine R^2_E far below mainstream (0.67–0.78 vs. 0.99+).
- **Gemma-3-1B** (1 pair): abnormal full-vocab E R^2 (≈ 0.38), BI rewrite magnitude too large.

A.9 EFFECTIVE AFFINE SUBSPACE ANALYSIS (§3 DETAILS)

Construction: $Z = [X, \mathbf{1}_V]$; SVD on Z; take first K left singular vectors Q_K with cumulative energy $\eta = 95\%$. On raw $\Delta W = Y - X$, compute $R^2_{\text{aff},K} = \|Q_K Q_K^\top \Delta W\|_F^2 / \|\Delta W\|_F^2$.

Auxiliary metrics: sample-dimension SVD on ΔW yields σ_i, u_i; define $a_i(K)$, r_L, $A_{L,K}$, $R^{2,(L)}_{\text{aff},K} = r_L A_{L,K}$. Under centered convention, $D = Y_c - X_c$, $P = X_c(A - I)$, $P/D = \|P\|_F^2 / \|D\|_F^2$ (see Table 27).

Table 18: Untied 13 BI pairs: median effective subspace metrics ($\eta = 95\%$). $R^2_K \equiv R^2_{\text{aff},K}$.

Side	r_5	$A_{5,K}$	$R^{2,(5)}_K$	$R^{2,(10)}_K$	R^2_K	P/D
E	1.9%	0.048	0.07%	0.10%	3.6%	4.9%
U	22.2%	0.829	13.3%	16.0%	33.0%	31.9%

Key comparison: in all 13/13 untied pairs, $R^2_K(\text{U}) > R^2_K(\text{E})$ and P/D also holds 13/13; tied 17 pairs have $R^2_K \approx 32\%$, $A_{5,K} \approx 0.82$, on par with untied U. Per-model 43 rows appear in Table 26.

A.10 A-LoRA vs Vocab LoRA Budget Probe (§3 Details)

Parameter budget probe protocol: offline fitting simulation on observed post-training change, not adapter training. Given $D = Y_c - X_c$, optimal full affine yields $P = X_c(A - I)$, $R = D - P$. At W rank k, Vocab LoRA fits with k-rank W-form; A-LoRA fits with $r_{\text{aff}}(k)$-rank hidden factors per Eq. (11). Report $\text{aff}/\text{W} = \text{gain}_{\text{aff}}/\text{gain}_W$.

Table 19: A-LoRA vs Vocab LoRA grouped summary (median over 30 pairs).

Group	n	aff/W@1	win@1	aff/W@4	win@4	aff@W1
Untied U	13	2.18×	12/13	1.79×	10/13	15
Untied E	13	0.34×	1/13	0.33×	0/13	15
Tied E=U	17	3.00×	15/17	1.81×	14/17	37

Hybrid ($P + R$) adds limited incremental value to the U-side main narrative and is not primary evidence in the main text. 30-pair model list and per-model budget appear in Tables 25 and 27.

A.11 Topology Configuration Details

The four topologies in Table 14 combine with either parameterization. Shared effective matrices are restricted to tied bases; unilateral and independent bilateral adaptation apply to either base architecture. Topology selection must consider both task gains and retained behavior.

A.12 Parameter Budget Matching and Baselines

Nominal budget-matching rule (not used by every reported comparison): fix Vocab rank r_W; A-LoRA takes $r_{\text{aff}} = \lfloor r_W(V + d)/(2d) \rfloor$, ensuring $P_{\text{aff}} \approx P_W$ (Eqs. (2.1)–(11)).

Comparison with internal-layer LoRA: let internal LoRA act on standard modules (Q/K/V/O + MLP), with parameter count $P_{\text{int}} = L_{\text{mod}} \cdot r_{\text{int}}(V_{\text{mod}} + d)$. Matched-budget designs may impose $P_{\text{border}} + P_{\text{int}} = P_{\text{total}}$; other comparisons report unequal counts explicitly; specific ranks appear in Appendix A.15.

A.13 Method Details: Forward Pass, LoRMA$_+$ Relation, and Merge

Vocab LoRA forward. For embedding lookup,

$$e_x = M_0[x, :] + \frac{\alpha}{r} B[x, :] C; \tag{12}$$

for LM head,

$$z = M_0 h + \frac{\alpha}{r} BCh. \tag{13}$$

In Eq. (12), only row x of B is activated, yielding row-sparse gradients along vocabulary dimension; Eq. (13) produces dense gradients over all V rows.

A-LoRA forward. For token x,

$$e_x = M_0[x,:]\Big(I_d + \tfrac{\alpha}{r}PQ\Big) + \beta \tag{14}$$

$$= M_0[x,:] + M_0[x,:]\tfrac{\alpha}{r}PQ + \beta. \tag{15}$$

For hidden $h \in \mathbb{R}^d$,

$$z = M_0\Big(I_d + \tfrac{\alpha}{r}PQ\Big)h + \mathbf{1}_V\big(\beta^\top h\big) \tag{16}$$

$$= M_0 h + M_0\tfrac{\alpha}{r}PQh + (\beta^\top h)\,\mathbf{1}_V. \tag{17}$$

Here $\beta^\top h$ is identical for all vocabulary logits and cancels under softmax. This is a row-shared matrix shift, distinct from adding a hidden bias before an independent output projection, which would produce a vocabulary-dependent logit bias. The new comparison disables bias on both sides.

Relation to LoRMA$_+$. LoRMA$_+$ writes effective weights as $W_{\text{eff}} = (I + \frac{\alpha}{r}BA)W_0$ (Bihany et al., 2025); A-LoRA's multiplicative part $M_0(I_d + \frac{\alpha}{r}PQ)$ is formally similar. The difference is that A-LoRA targets $V \times d$ vocabulary matrices, exploits dimensional asymmetry $V \gg d$ by placing adapter parameters in hidden space, and additionally includes $\mathbf{1}_V\beta^\top$; the weight-space analysis in §3 is descriptive motivation, rather than a necessary or sufficient applicability condition.

Shared merge. For shared A-LoRA, the effective shared matrix can be merged offline as

$$W_{\text{merged}}[i,:] := W_0[i,:]\Big(I_d + \tfrac{\alpha}{r}PQ\Big) + \beta, \tag{18}$$

where $i \in \{1,\dots,V\}$. After merge, the adapter requires no additional operations relative to the same dense matrix computation; actual latency is not benchmarked; under decoupled topology, $E_{\text{eff}} \neq U_{\text{eff}}$ and merge to a single tied matrix is generally impossible.

A.14 Mergeable Tied A-LoRA Details

Mergeable tied A-LoRA addresses deployment constraints in tied-embedding models. It shares the same hidden-space affine adapter on input and `lm_head` sides, with transposed action on the output side so effective weights after training can still merge back into one tied embedding matrix. After merge per Eq. (18), we check element-wise difference between online adapter logits and merged-weight logits; maximum absolute error in the historical merge checks is about 3×10^{-5}–6×10^{-5}, within floating-point error.

Table 20 records similar historical instruction validation losses for several placements. This is compatible with overlapping adaptation capacity, but does not identify a causal redundancy mechanism or demonstrate retained generation.

Table 20: Parameter and effect calibration for mergeable tied A-LoRA (Qwen3-0.6B, main instruction-tuning set).

Config	Params	eval loss
hidden LoRA	5.05M	1.101
`affine_input + hidden`	5.06M	1.097
`affine_lm_head + hidden`	5.06M	**1.093**
`affine_input_lm_head + hidden`	5.08M	1.095
mergeable tied + hidden	5.06M	1.094

A.15 DOWNSTREAM LORA POST-TRAINING SETUP

This section records the historical instruction-loss protocol, which differs from the fresh-seed structured-task protocol in Appendix A.1. These results characterize validation loss under their recorded configurations, not preservation of unrelated-task generation.

Table 21: Unified downstream LoRA post-training configuration.

Item	Setting
Main data	`sft_t2t_mini_25k` subset from MiniMind instruction data, 24k train + 1k eval.
External validation	UltraChat 100k (100k train + 2k eval) and MiniMind random 100k (100k train + 2k eval).
Base models	Qwen2.5-0.5B/1.5B/3B/7B and Qwen3-0.6B/1.7B/4B/8B; covers tied and untied embedding structures.
Training	Except full fine-tune sanity probe, base weights frozen and only adapters trained; default 1 epoch, max sequence length 1024, effective batch size is typically 16, learning rate 2×10^{-4}, cosine decay, warmup ratio 0.03.
hidden LoRA	Default on q/k/v/o and MLP up/down/gate projections in every layer, rank 8, $\alpha = 16$; low-rank sweep uses rank 1/2/4.
A-LoRA	Default vocabulary-boundary hidden-space affine rank 16, $\alpha = 128$ ($s = \alpha/r = 8$); A-LoRA-only and tied alignment experiments additionally sweep $s \in \{16, 32, 64\}$ or rank 32.
Vocab LoRA	W-form LoRA on `embed_tokens`/`lm_head`, commonly rank 1/2/4, as vocabulary-dimensional parameter-efficiency control.
Seeds and metrics	Default seed 42; key placement, tied, and mergeable experiments add seeds 43/44. Uniformly report final validation loss (eval loss, lower is better) and trainable parameter count.

Table 22: Downstream LoRA / A-LoRA variant definitions.

Variant	Train location	Purpose
hidden LoRA	Transformer block projections	Standard PEFT main baseline
single-layer LoRA	Single decoder layer projections	Same-scale small-budget control
hidden + Vocab LoRA	hidden LoRA + `embed_tokens`/`lm_head` W-form	Vocab-dim LoRA parameter-efficiency control
A-LoRA-only input	`embed_tokens` A-LoRA	Standalone small-budget primary config
A-LoRA-only lm_head	`lm_head` A-LoRA	Affine-only placement attribution control
A-LoRA-only input+lm_head	`embed_tokens` + `lm_head` A-LoRA	Pure vocabulary-boundary bilateral control
hidden + input A-LoRA	hidden LoRA + input A-LoRA	Input-side redundancy check
hidden + lm_head A-LoRA	hidden LoRA + output A-LoRA	Primary config combined with hidden LoRA
hidden + input+lm_head A-LoRA	hidden LoRA + bilateral A-LoRA	Bilateral redundancy control
mergeable tied A-LoRA	tied shared A-LoRA, transposed on output	Mergeable tied config avoiding bilateral redundancy

A.16 DOWNSTREAM EXPERIMENT RAW RESULT TABLES

Tables 23–24 retain historical instruction-loss measurements. They are not the fresh-seed main comparison and do not constitute equal-topology or equal-budget efficiency evidence. Parameter counts follow actual configs: A-LoRA input counts unilateral `embed_tokens` adapter; Vocab LoRA counts bilateral `embed_tokens`+`lm_head` W-form adapters; all counts include trainable parameters and bias, rounded to thousands or millions.

These historical values are final validation loss after one epoch, lower is better. Their data, training length, optional bias, and evaluation differ from the new main-text structured-task comparison.

Table 23: Raw results for parameter efficiency and vocabulary-boundary value.

Evidence line	Model	Config	Params	eval loss
Parameter efficiency	Qwen3-0.6B	A-LoRA input r2	5.1k	**1.494**
Parameter efficiency	Qwen3-0.6B	Vocab LoRA r2	0.61M	1.529
Parameter efficiency	Qwen3-0.6B	A-LoRA input r4	9.2k	**1.469**
Parameter efficiency	Qwen3-0.6B	Vocab LoRA r4	1.23M	1.495
Parameter efficiency	Qwen2.5-1.5B	A-LoRA input r16	50.7k	**1.473**
Parameter efficiency	Qwen2.5-1.5B	Vocab LoRA r4	1.23M	1.538
Parameter efficiency	Qwen3-1.7B	A-LoRA input r16	67.6k	**0.981**
Parameter efficiency	Qwen3-1.7B	Vocab LoRA r4	1.24M	1.145
Boundary value	Qwen3-0.6B	A-LoRA input r16, s=64	34k	**1.419**
Boundary value	Qwen3-0.6B	single-layer LoRA qkvo, mid	41k	1.473
Boundary value	Qwen3-0.6B	single-layer LoRA q, layer 0	49k	1.626
Boundary value	Qwen3-0.6B	single-layer LoRA q, mid	49k	1.667

Table 24: Raw results for standalone placement, hidden-LoRA-conditioned placement, and mergeable tied.

Evidence line	Model / condition	baseline / input	lm_head / mergeable	Conclusion
standalone	Qwen2.5-0.5B	**1.816**	1.907	input stronger
standalone	Qwen3-0.6B	**1.460**	1.621	input stronger
standalone	Qwen2.5-1.5B	**1.502**	1.608	input stronger
standalone	Qwen2.5-7B untied	**1.258**	1.392	input stronger
+hidden LoRA	Qwen2.5-0.5B	1.297	**1.288**	lm_head stronger
+hidden LoRA	Qwen3-0.6B	1.097	**1.093**	lm_head stronger
+hidden LoRA	Qwen2.5-1.5B	1.050	**1.044**	lm_head stronger
+hidden LoRA	Qwen3-1.7B	0.7362	**0.7328**	lm_head stronger
+hidden LoRA	Qwen2.5-7B untied	0.918	**0.913**	lm_head stronger
multi-seed	Qwen2.5-0.5B seed 43/44	1.297 / 1.297	**1.289 / 1.288**	same direction
multi-seed	Qwen3-0.6B seed 43/44	1.096 / 1.097	**1.094 / 1.093**	same direction
mergeable tied	Qwen3-0.6B + hidden	hidden 1.101	**mergeable 1.094**	effect preserved
mergeable tied	Qwen3-0.6B seeds 43/44	—	**1.095 / 1.095**	similar loss
external	UltraChat 100k 1.5B / 1.7B	hidden 1.075 / 1.058	mergeable 1.075 / 1.058	equal rounded loss

A.17 Implementation Notes

A-LoRA implementation. Eq. (14) can compute $M_0[x,:]\frac{\alpha}{r}PQ + \beta$ after lookup, avoiding materializing full M_{eff}.

Vocabulary-parallel tensor parallelism. In vocabulary-parallel settings, vocabulary-related matrices can be partitioned along the V dimension (Shoeybi et al., 2019). Thus Vocab LoRA's $B \in \mathbb{R}^{V\times r}$ shards with the vocabulary dimension, whereas A-LoRA's P, Q, β contain no V dimension and can be replicated or sharded along hidden dimensions.

Merge and serving. T3 shared A-LoRA can merge per Eq. (18) for single-matrix serving; T4 decoupled requires loading E_{eff} and U_{eff} separately.

A.18 Per-Model Raw Data Tables

The following three tables provide per-model raw data for the main-text and preceding aggregate tables. They are placed at the end of the appendix for ease of reference.

Table 25: 30-pair BI model list (Base→Instruct).

Base	**Instruct**	**tier**	**tied**	d	V
DS-V3	DS-V3	ext.	N	7168	129280
DS-V3.1	DS-V3.1	ext.	N	7168	129280
G-2-27B	G-2-27B	main	Y	4608	256000
G-2-2B	G-2-2B	main	Y	2304	256000
G-2-9B	G-2-9B	main	Y	3584	256000
G-3-12B	G-3-12B	main	Y	3840	262208
G-3-27B	G-3-27B	main	Y	5376	262208
G-3-4B	G-3-4B	main	Y	2560	262208
L-3.1-70B	L-3.1-70B	main	N	8192	128256
L-3.1-8B	L-3.1-8B	main	N	4096	128256
L-3.2-1B	L-3.2-1B	main	Y	2048	128256
L-3.2-3B	L-3.2-3B	main	Y	3072	128256
Q2.5-0.5B	Q2.5-0.5B	main	Y	896	151936
Q2.5-1.5B	Q2.5-1.5B	main	Y	1536	151936
Q2.5-14B	Q2.5-14B	main	N	5120	152064
Q2.5-32B	Q2.5-32B	main	N	5120	152064
Q2.5-3B	Q2.5-3B	main	Y	2048	151936
Q2.5-72B	Q2.5-72B	main	N	8192	152064
Q2.5-7B	Q2.5-7B	main	N	3584	152064
Q3-0.6B	Q3-0.6B	main	Y	1024	151936
Q3-1.7B	Q3-1.7B	main	Y	2048	151936
Q3-14B	Q3-14B	main	N	5120	151936
Q3-30B-A3B	Q3-30B-A3B	ext.	N	2048	151936
Q3-4B	Q3-4B	main	Y	2560	151936
Q3-8B	Q3-8B	main	N	4096	151936
Q3.5-0.8B	Q3.5-0.8B	main	Y	1024	248320
Q3.5-2B	Q3.5-2B	main	Y	2048	248320
Q3.5-35B-A3B	Q3.5-35B-A3B	ext.	N	2048	248320
Q3.5-4B	Q3.5-4B	main	Y	2560	248320
Q3.5-9B	Q3.5-9B	main	N	4096	248320

Table 26: Per-model effective affine subspace metrics (43 rows, $\eta = 95\%$). Tied models: E=U row only; untied models: separate E/U rows. $R^2_K \equiv R^2_{\text{aff},K}$.

Base model	**Side**	r_5	$A_{5,K}$	$A_5/(K/V)$	R^2_K	$R^{2,(5)}_K$	$R^{2,(10)}_K$
Q3-0.6B	E	21.6%	0.749	137.1×	34.8%	16.20%	19.68%
Q3-1.7B	E	18.4%	0.790	67.7×	32.2%	14.50%	16.85%
Q3-4B	E	16.6%	0.817	58.3×	32.2%	13.52%	16.27%
Q3-8B	E	2.3%	0.088	4.9×	6.1%	0.20%	0.26%
Q3-8B	U	14.5%	0.802	34.6×	32.0%	11.64%	14.29%
Q3-14B	E	2.8%	0.079	4.9×	5.7%	0.22%	0.29%
Q3-14B	U	13.4%	0.852	29.6×	32.4%	11.42%	14.06%
Q3-30B-A3B	E	0.9%	0.057	5.6×	3.4%	0.05%	0.09%
Q3-30B-A3B	U	17.8%	0.747	66.6×	33.0%	13.28%	16.05%
Q3.5-0.8B	E	9.7%	0.523	189.4×	11.0%	5.09%	5.85%
Q3.5-2B	E	8.2%	0.684	118.3×	11.6%	5.64%	6.12%
Q3.5-4B	E	7.4%	0.667	91.8×	11.5%	4.91%	5.36%
Q3.5-9B	E	1.6%	0.029	2.2×	2.5%	0.05%	0.07%
Q3.5-9B	U	9.7%	0.731	55.3×	20.2%	7.12%	8.30%
Q3.5-35B-A3B	E	2.1%	0.020	3.8×	1.5%	0.04%	0.06%
Q3.5-35B-A3B	U	10.5%	0.665	105.1×	20.0%	6.98%	8.15%
Q2.5-0.5B	E	29.1%	0.878	271.7×	31.4%	25.53%	26.22%
Q2.5-1.5B	E	68.5%	0.756	92.6×	57.1%	51.75%	52.72%
Q2.5-3B	E	63.0%	0.752	67.5×	53.8%	47.37%	48.55%
Q2.5-7B	E	4.8%	0.015	1.6×	1.5%	0.07%	0.10%
Q2.5-7B	U	48.8%	0.839	48.5×	47.0%	40.98%	41.90%

Continued on next page

(continued from Table 26)

Base model	**Side**	r_5	$A_{5,K}$	$A_5/(K/V)$	R_K^2	$R_K^{2,(5)}$	$R_K^{2,(10)}$
Q2.5-14B	E	3.0%	0.023	1.6×	2.4%	0.07%	0.10%
Q2.5-14B	U	49.4%	0.834	29.5×	51.5%	41.20%	42.87%
Q2.5-32B	E	3.2%	0.021	1.6×	2.5%	0.07%	0.11%
Q2.5-32B	U	52.2%	0.855	30.3×	53.0%	44.60%	45.67%
Q2.5-72B	E	1.5%	0.032	1.3×	3.6%	0.05%	0.08%
Q2.5-72B	U	51.2%	0.889	21.1×	51.8%	45.52%	46.14%
L-3.2-1B	E	26.0%	0.884	69.3×	45.8%	22.99%	25.82%
L-3.2-3B	E	22.0%	0.886	45.6×	44.4%	19.47%	22.02%
L-3.1-8B	E	1.7%	0.050	2.2×	3.7%	0.09%	0.12%
L-3.1-8B	U	22.9%	0.829	32.7×	38.2%	18.97%	21.90%
L-3.1-70B	E	1.2%	0.068	1.5×	6.1%	0.08%	0.12%
L-3.1-70B	U	22.2%	0.844	17.0×	38.6%	18.78%	20.95%
G-2-2B	E	47.5%	0.660	90.2×	39.6%	31.35%	32.67%
G-2-9B	E	32.8%	0.944	85.4×	38.8%	30.97%	32.91%
G-2-27B	E	35.7%	0.871	65.0×	40.6%	31.11%	33.58%
G-3-4B	E	9.4%	0.879	111.1×	13.5%	8.27%	9.52%
G-3-12B	E	7.5%	0.836	71.9×	12.9%	6.27%	7.27%
G-3-27B	E	5.1%	0.866	53.4×	10.3%	4.39%	4.81%
DS-V3	E	1.9%	0.048	1.1×	4.8%	0.09%	0.14%
DS-V3	U	22.8%	0.366	8.3×	13.2%	8.36%	8.96%
DS-V3.1	E	1.3%	0.050	1.1×	4.9%	0.06%	0.10%
DS-V3.1	U	16.1%	0.465	10.6×	12.8%	7.48%	8.16%

Table 27: Per-model A-LoRA budget probe (30 pairs). P/D = centered linear share; aff/W@1 = aff/W explained-gain ratio at W rank=1; aff@W1 = parameter-matched aff rank on U side.

Base model	tier	tied	V/d	P/D (E)	P/D (U)	aff/W@1 (E)	aff/W@1 (U)	aff@W1
DS-V3	ext.	N	18.0	0.060	0.122	0.25×	0.75×	9
DS-V3.1	ext.	N	18.0	0.061	0.112	0.29×	1.15×	9
G-2-27B	main	Y	55.6	0.320	0.320	3.00×	3.00×	28
G-2-2B	main	Y	111.1	0.369	0.369	1.30×	1.30×	56
G-2-9B	main	Y	71.4	0.269	0.269	2.60×	2.60×	36
G-3-12B	main	Y	68.3	0.123	0.123	2.18×	2.18×	34
G-3-27B	main	Y	48.8	0.099	0.099	2.58×	2.58×	24
G-3-4B	main	Y	102.4	0.101	0.101	4.32×	4.32×	51
L-3.1-70B	main	N	15.7	0.082	0.478	0.36×	2.18×	8
L-3.1-8B	main	N	31.3	0.048	0.411	0.26×	2.82×	16
L-3.2-1B	main	Y	62.6	0.408	0.408	3.04×	3.04×	31
L-3.2-3B	main	Y	41.8	0.402	0.402	2.87×	2.87×	21
Q2.5-0.5B	main	Y	169.6	0.257	0.257	1.58×	1.58×	85
Q2.5-1.5B	main	Y	98.9	0.568	0.568	0.93×	0.93×	49
Q2.5-14B	main	N	29.7	0.047	0.486	0.23×	1.18×	15
Q2.5-32B	main	N	29.7	0.049	0.515	0.28×	1.09×	15
Q2.5-3B	main	Y	74.2	0.554	0.554	0.98×	0.98×	37
Q2.5-72B	main	N	18.6	0.073	0.481	0.34×	1.08×	9
Q2.5-7B	main	N	42.4	0.032	0.462	0.18×	1.16×	21
Q3-0.6B	main	Y	148.4	0.311	0.311	6.00×	6.00×	74
Q3-1.7B	main	Y	74.2	0.296	0.296	4.18×	4.18×	37
Q3-14B	main	N	29.7	0.079	0.319	0.52×	4.30×	15
Q3-30B-A3B	ext.	N	74.2	0.041	0.311	1.48×	5.17×	37
Q3-4B	main	Y	59.4	0.298	0.298	4.99×	4.99×	30
Q3-8B	main	N	37.1	0.073	0.308	0.54×	4.45×	19
Q3.5-0.8B	main	Y	242.5	0.119	0.119	4.27×	4.27×	121
Q3.5-2B	main	Y	121.2	0.115	0.115	3.45×	3.45×	61
Q3.5-35B-A3B	ext.	N	121.2	0.020	0.207	0.59×	5.25×	61
Q3.5-4B	main	Y	97.0	0.120	0.120	3.04×	3.04×	49
Q3.5-9B	main	N	60.6	0.029	0.210	0.35×	5.04×	30