WHICH LAYERS SHOULD SHARE? FUNCTIONAL SUBSTITUTABILITY FOR LANGUAGE MODEL COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-layer parameter sharing compresses a pretrained language model without shortening its computation graph, but similarity between components does not establish whether one can perform another's role. We introduce SUBSHARE, which selects shared bases through directed functional substitution. For each compatible donor–target pair, SUBSHARE measures the next-token loss change caused by substitution in the intact model, converts the resulting asymmetric matrix into medoid group costs, and reuses it across budgets with Greedy or mixed-integer planning. Our local-to-global analysis relates selection quality to plan-dependent residual variation rather than absolute-loss prediction error. Across two Qwen3.5 scales and two Gemma models, SUBSHARE achieves the best recovered loss among all tested assignment controls in 15 of 16 model–budget settings. The gains extend to zero-shot tasks and transfer to Basis Sharing and exact tying. These results identify functional substitutability as a reusable criterion for allocating shared capacity, complementary to the parameterization used to realize it.

1 INTRODUCTION

Depth pruning compresses a model by removing blocks. Cross-layer parameter sharing poses a different problem: how can a model retain its logical computation while reducing the distinct parameters that realize it? Several logical positions can reference one physical base, preserving the original sequence of transformations while lowering unique-weight storage. Recursive and factorized models show that this separation between logical depth and physical capacity can preserve substantial capability (Bae et al., 2025; Wang et al., 2025a). It also creates a structural decision that ordinary pruning does not face: under a fixed physical budget, which pretrained components should share, and which component should supply their common base?

The difficulty comes from positional specialization. Pretrained components adapt to the activations, residual pathways, and downstream computation at their logical positions. Two layers can be close in depth or weight space yet perform different roles, while a more distant layer may be the safer substitute. Topology, weight similarity, and curvature provide useful structural or geometric signals (Wang et al., 2025b; Bae et al., 2025; Zhang et al., 2026a); activation similarity additionally compares component representations (Yun et al., 2026). Substitutability asks a different question: what happens to the model's task loss when a donor actually takes over a target's role? This intervention evaluates the donor together with the target's input distribution and downstream computation. Role transfer is directed: placing component j at position i changes a different computation from placing i at j. A sharing plan must preserve this donor–target relation, especially in hybrid models whose operator families define hard compatibility boundaries.

The resulting challenge is local-to-global. Evaluating and recovering every complete sharing plan is prohibitively combinatorial, whereas atomic substitutions can be measured once and reused across many budgets. Yet a deployed plan composes many replacements, constructs shared bases, adds position-specific residuals, and undergoes recovery training. The scientific question is therefore not only whether an individual layer is replaceable, but when local evidence of functional role transfer remains informative for a global physical-basis assignment.

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.

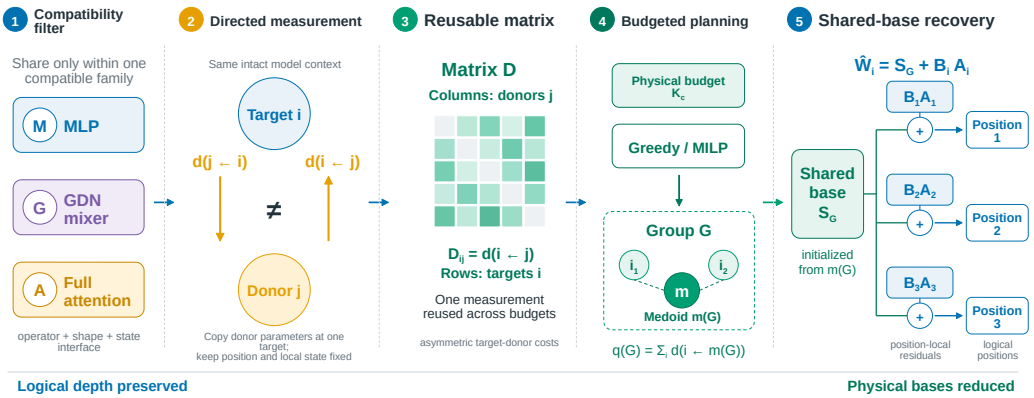

Figure 1: **SubShare from measurement to recovered model.** Compatibility filtering restricts substitutions to matched component families. Directed donor-to-target interventions (projection-matched scope shown) construct an asymmetric matrix that Greedy or MILP converts into a budgeted medoid partition. Each group uses a medoid-initialized shared base and position-specific low-rank residuals, preserving logical depth while reducing physical bases.

We address this question with SUBSHARE (Figure 1). For every compatible target i and donor j, we replace a specified parameter scope at i with its counterpart from j and measure the change in next-token loss in the intact network. The scope can cover shared projections or a complete hybrid mixer; both measure the downstream effect of the intervention. The resulting asymmetric matrix assigns each candidate group the sum of directed costs to its best medoid. Greedy and mixed-integer linear programming (MILP) use these costs to select budgeted partitions. The central object is the measured donor–target relation: it can guide different planners and different representations of shared weights. Our primary implementation uses medoid-initialized bases with position-specific low-rank residuals; Basis Sharing and exact-tying comparisons also benefit from functional grouping.

Our analysis identifies ranking as the bridge from local measurement to global recovery. A pathwise bound relates non-additive loss to replacement magnitude and cross-component coupling. A surrogate-regret result then shows that selection depends on the variation of interaction and recovery residuals between plans, relative to their proxy-cost gaps. Common loss offsets cancel. Thus local costs can guide assignment without reconstructing the absolute loss of every recovered model, and calibration and planner errors enter the same decision framework.

The experiments test both assignment quality and this local-to-global link. The main grid crosses two Qwen3.5 scales and two Gemma models with four physical budgets and three recovery seeds, comparing assignments under common parameterization and recovery. Across 32 complete Qwen3.5 plans, the whole-mixer functional score reaches Spearman correlation 0.731 with recovered loss and zero top-1 selection regret within this pool. Calibration and double-intervention tests examine sampling stability and non-additivity. Zero-shot evaluation connects recovery quality to task capability, while capacity ablations separate better assignment from larger residuals. Fivefold recovery tests whether the gains persist as the shared models adapt.

The contributions are:

- We identify cross-layer sharing as a directed functional assignment problem and measure donor suitability through its effect on task loss, yielding reusable costs for medoid selection and budgeted planning.
- We provide a local-to-global analysis that separates interaction, calibration, and solver errors, and identifies residual variation as the condition governing recovered-plan ordering.
- We establish cross-provider and cross-scale assignment gains on a four-budget grid, explain their relation to recovery through plan-level diagnostics, and demonstrate benefits for zero-shot capability and alternative sharing parameterizations.

2 Related Work

Choosing which positions share. Universal Transformers and ALBERT establish depth-wise reuse during pretraining (Dehghani et al., 2019; Lan et al., 2020); subsequent models study cyclic, narrow, and learned tying (Takase & Kiyono, 2023; Liu et al., 2024; David-Hay & Wolf, 2024). For pretrained networks, SHARP groups adjacent MLPs (Wang et al., 2025b), while Relaxed Recursive Transformers (RRT) use recurring blocks (Bae et al., 2025). SharVeT combines weight and activation similarity through spectral clustering (Yun et al., 2026); Geo-Sharing derives assignments from Hessian-weighted geometry (Zhang et al., 2026a). These methods motivate our Adjacent, Depth-spaced, SharVeT, and Geo-Sharing comparisons. Random and Weight similarity provide the assignment controls also studied by SharVeT. SubShare measures a different quantity: the end-to-end loss change when a specific donor serves a specific target, preserving the direction of functional role transfer.

Representing shared and local parameters. Assignment is distinct from its parameterization. RRT adds layer-wise LoRA to recurring bases (Bae et al., 2025; Hu et al., 2022); Basis Sharing and LiSA use shared subspaces (Wang et al., 2025a; Mu et al., 2026). SharVeT combines clustering with adaptive allocation and vector-based recovery (Yun et al., 2026). Conditional subspaces, routing, and input-conditioned residuals further expand the representations available to shared models (Li et al., 2026; Chen et al., 2026; Zhang et al., 2026b; Nouriborji et al., 2025; Jaber & Jaber, 2026). Our main comparison holds the shared-base-plus-residual implementation fixed while varying assignment. Additional Basis Sharing and exact-tying results examine the assignment contribution under different parameterizations.

Compression and functional compatibility. Pruning selects weights, dimensions, or blocks through magnitude, activation, gradient, and output criteria (Frantar & Alistarh, 2023; Sun et al., 2024; Ma et al., 2023; Ashkboos et al., 2024; Song et al., 2024; Men et al., 2025). Replacement-aware and global methods also admit noncontiguous structures (Shopkhoev et al., 2025; Jansen et al., 2026; Cunegatti et al., 2026). Sharing instead retains every logical position and allocates fewer physical bases across them. Representation similarity (Kornblith et al., 2019), directional stitching (Hernandez et al., 2023; Athanasiadis et al., 2026), and functional damage (Hinostroza et al., 2026; Khilar, 2026) offer complementary notions of compatibility. Component interventions and Möbius transforms expose higher-order effects (Kramár et al., 2024; Kang et al., 2024). We connect directed interventions to budgeted sharing through interaction bounds and selection regret, with disjoint calibration motivated by depth-pruning sensitivity (Kim et al., 2026).

3 Directed Functional Assignment

Let M_θ be a pretrained causal language model. A component instance is $v = (c, i)$, where c is a compatibility family (MLP, full attention, or linear-attention mixer) and i is its logical position. Search never crosses operator, tensor-shape, or state-interface boundaries.

3.1 Sharing plans and physical budget

For each family c, a plan P_c partitions its instances V_c into groups. Every non-singleton group G selects one physical base S_G used at all member positions. Position $i \in G$ retains a rank-r residual; a singleton remains unchanged. The model executes the original logical sequence, so sharing reduces unique stored tensors rather than logical depth.

If p_G is the size of base G, $a_i(r, G)$ is the residual size (zero for a singleton), and P_{fixed} contains embeddings and tensors outside the search, then

$$P_{\text{phys}}(P, r) = P_{\text{fixed}} + \sum_{G \in P} \left[p_G + \sum_{i \in G} a_i(r, G) \right]. \tag{1}$$

At fixed rank, matching each family's group-size histogram fixes the base and residual shapes and hence the physical parameter budget.

3.2 DIRECTED SUBSTITUTION MATRIX

An atomic intervention $a = (i \leftarrow j, c)$ copies a predefined compatible parameter scope from donor j into target i. The target's logical position, residual path, and state interface remain fixed. Let δ_a be the resulting parameter perturbation. On calibration sequences D_{proxy}, define

$$d(a) = \mathcal{L}(\theta + \delta_a; D_{\text{proxy}}) - \mathcal{L}(\theta; D_{\text{proxy}}). \tag{2}$$

Every intervention uses the intact model's sequences, yielding paired differences; negative effects are retained. The score is asymmetric because opposite directions perturb different coordinates under different activations and downstream Jacobians.

For hybrid GDN mixers, the intervention scope has two choices. Projection-matched scoring replaces the matrices that will be shared; whole-mixer scoring additionally replaces the donor's local dynamics during measurement. Both use Equation 2: the distinction is the intervened parameter set, not the loss being measured. Whole-mixer scoring tests the donor as a complete operator, while projection-matched scoring tests its projections with the target's dynamics. Ordinary attention has one scope.

For example, write a GDN component as $D(W_i, \phi_i)$, with shareable projections W_i and local dynamics ϕ_i. A donor j gives $D(W_j, \phi_i)$ under projection-matched measurement and $D(W_j, \phi_j)$ under whole-mixer measurement. The recovered target uses $D(S_G + B_i A_i, \phi_i)$ in both cases. Measurement therefore compares two notions of donor compatibility while deployment preserves the target's local dynamics.

For a legal group G, its best directed medoid and additive cost are

$$m(G) = \arg\min_{j \in G} \sum_{i \in G \setminus \{j\}} d(i \leftarrow j, c), \qquad q(G) = \sum_{i \in G \setminus \{m(G)\}} d(i \leftarrow m(G), c). \tag{3}$$

We enumerate groups of size one through three and reuse their costs across physical budgets.

For a family with n_c components, this cap permits base counts $\lceil n_c/3 \rceil \le K_c \le n_c$ before any additional restrictions. All main-grid budgets are feasible under this cap. It bounds catalog growth while retaining singleton, pair, and triple assignments; partitions containing larger groups lie outside this search space.

For family sizes n_c, constructing a full matrix requires $\sum_c n_c(n_c - 1)$ atomic evaluations. Candidate-group scoring then uses these stored entries without additional model evaluations. The cap of three gives $\sum_c \sum_{k=1}^{3} \binom{n_c}{k}$ candidate groups: for Qwen3.5, 888 directed measurements yield 3,352 candidates across its three families. Changing the base budget changes the partition constraints rather than the measurements. This amortizes functional evaluation over a frontier of shared models instead of recovering every possible partition.

3.3 GREEDY AND EXACT PLANNING

Let $x_G \in \{0, 1\}$ indicate whether candidate group G is selected. For each family, exact coverage and a budget of K_c physical bases require

$$\sum_{G \ni i} x_G = 1 \quad \forall i \in V_c, \qquad \sum_{G \in \mathcal{G}_c} x_G = K_c. \tag{4}$$

The planning objective is $Q(P) = \sum_G q(G) x_G$. SUBSHARE-MILP solves this set-partitioning problem exactly with HiGHS. Its realized per-family group-size histogram defines the physical template for the histogram-matched comparisons. SUBSHARE-GREEDY selects the lowest-cost feasible groups under that same histogram. Their difference is solver error for one functional objective.

Both planners generate candidates. For hybrid mixers, candidates also vary the intervention scope. At each model and budget, matched recovery selects the lowest-development-NLL plan before frozen-test evaluation. We call this procedure SUBSHARE and retain planner-specific names for analysis. Each scope's matrix supports approximate, exact, and multi-budget planning.

3.4 SHARED-BASE INITIALIZATION AND RECOVERY

For every shareable matrix W_i in a non-singleton group,

$$\widehat{W}_i = S_G + B_i A_i, \qquad S_G \leftarrow W_{m(G)}. \tag{5}$$

Truncated SVD of $W_i - W_{m(G)}$ initializes a non-medoid residual; the medoid residual starts at zero. Biases, normalizations, positional mechanisms, and GDN dynamics remain local unless stated otherwise. Matched recovery trains bases and residuals with next-token, teacher-KL, hidden-state, and available gate-matching losses while fixing teacher, token order, optimizer, rank, and token budget.

The deployed model keeps GDN dynamics local under either scoring scope. Comparing the two scores on the same recovered plans tests how the measurement scope affects assignment quality without changing the shared parameterization.

4 WHEN DO LOCAL COSTS GUIDE A GLOBAL PLAN?

A planner needs to order complete plans, whereas the score measures one replacement at a time. We analyze this gap in two steps: identify the non-additive terms, then relate their variation across plans to selection regret. Each plan specifies both its groups and their donors. Fix a distribution D and suppress its subscript on d and Q.

4.1 INTERACTIONS AND RECOVERY

For a non-overlapping intervention set S, let $\Delta(S) = \mathcal{L}(\theta + \sum_{a \in S} \delta_a) - \mathcal{L}(\theta)$. The additive error of plan P is $I(P) = \Delta(\mathcal{A}(P)) - Q(P)$, where $\mathcal{A}(P)$ contains its assignments. Möbius inversion expresses $I(P)$ as the sum of all higher-order interactions (Appendix A.2). Directional costs retain the distinct target roles before these interactions are composed; a scalar example in Appendix A.1 demonstrates why they are asymmetric.

Let $F_t(P)$ be the loss after t recovery tokens and define $R_t(P) = F_t(P) - \mathcal{L}(\theta) - \Delta(\mathcal{A}(P))$. Then

$$F_t(P) = \mathcal{L}(\theta) + Q(P) + I(P) + R_t(P). \tag{6}$$

Here $I(P)$ captures interactions among direct interventions. $R_t(P)$ captures the difference introduced by shared-base construction, low-rank approximation, and recovery. This separates the composition of substitutions from the construction and training of the deployed model.

The interaction term is controlled by cross-component curvature along the replacement path, rather than only at the pretrained weights.

Proposition 1 (Pathwise additive error)**.** *Suppose a plan has K disjoint parameter perturbations $\delta_1, \ldots, \delta_K$, each with norm at most ρ. Let $\mathcal{L}$ be twice continuously differentiable on the rectangles $\theta + s\sum_{j<i} \delta_j + t\delta_i$, $s, t \in [0, 1]$. If every corresponding cross-Hessian block has operator norm at most γ, then*

$$|I(P)| \le \binom{K}{2} \rho^2 \gamma. \tag{7}$$

Appendix A.3 derives the exact pathwise identity and a non-uniform bound. The result distinguishes the number of replacements, their magnitudes, and their cross-component coupling. A larger sharing group can alter all three. This bound identifies the sources of non-additivity; the paired-intervention diagnostic in Table 2(b) tests whether modeling their joint effect improves damage ordering.

4.2 ORDERING AND SELECTION REGRET

For feasible plans $\mathcal{F}_B$ at a fixed budget and recovery horizon, write

$$F_t(P) = \alpha + \beta Q(P) + e(P), \qquad \beta > 0, \tag{8}$$

where α absorbs a common offset and $e(P)$ is the plan-varying residual. Comparing two plans cancels that offset:

$$F_t(P) - F_t(P') = \beta[Q(P) - Q(P')] + e(P) - e(P'). \tag{9}$$

The decisive quantity is the difference $e(P) - e(P')$, not a common bias in absolute-loss prediction. A proxy gap preserves the ordering when it exceeds the opposing residual difference.

Theorem 1 (Surrogate regret)**.** *Let $\widehat{P}$ minimize Q and P^* minimize F_t over $\mathcal{F}_B$. If $|e(P)| \leq \epsilon$ for every feasible plan, then $F_t(\widehat{P}) - F_t(P^*) \leq 2\epsilon$. An η-suboptimal planner for Q has regret at most $\beta\eta + 2\epsilon$. The proxy and recovered optima coincide uniquely if $\beta[Q(P) - Q(\widehat{P})] > 2\epsilon$ for all $P \neq \widehat{P}$.*

The proof appears in Appendix A.5. The bound separates two routes to better selection: a measurement that better preserves recovered-plan ordering, and a solver with smaller proxy error η. Exact MILP addresses the latter; the complete-plan ranking test evaluates the former.

Appendix A.4 extends the result to finite calibration, separating sampling error, interaction, and construction/recovery mismatch. These terms organize the diagnostics: calibration-size comparisons probe sampling stability, paired interventions probe composition, and recovered-plan correlation and regret test the combined effect of composition and recovery. The fixed-score Greedy–MILP comparison then isolates the planner choice.

5 Experiments

The main grid evaluates Qwen3.5-0.8B/2B, Gemma 3 1B PT, and Gemma-2B at B50/B58/B67/B75, where Bx retains $x\%$ of shareable bases (Qwen Team, 2026; Gemma Team et al., 2024; Gemma Team, 2025). For Qwen3.5-2B at B67, this reduces the actual model from 1.882B to 1.491B parameters (20.8%) and its BF16 file from 3.51 to 2.78 GiB (Table 23). Qwen3.5-2B is the primary hybrid backbone. Its MLP, Gated DeltaNet, and full-attention projections form separate compatibility families (Yang et al., 2025). Full-attention Qwen2.5-1.5B and the RRT topology provide transfer controls (Yang et al., 2024; Bae et al., 2025). Recovery uses an 80/20 English–Chinese FineWeb-Edu/Ultra-FineWeb mixture with disjoint proxy, development, and frozen test shards (Penedo et al., 2024; Wang et al., 2025c).

Baselines and recovery. The main grid compares assignment pipelines using a common shared-base and recovery framework. We include Random (Yun et al., 2026), Adjacent (Wang et al., 2025b), Depth-spaced recurrence (Bae et al., 2025), Weight similarity and SharVeT (Yun et al., 2026), and Geo-Sharing (Zhang et al., 2026a). SharVeT combines weight and CKA activation similarity; our Geo-Sharing control uses diagonal curvature scores. Random, Adjacent, Depth-spaced, Weight similarity, Geo-Sharing, and SubShare use common group-size histograms and physical sizes. SharVeT matches the family-wise base counts with its native spectral partition. This preserves SharVeT's clustering decisions rather than reshaping its groups to a different template. Appendix B.2 reports the resulting physical parameter counts and assignment implementation. All methods use rank-128 residuals and matched teachers, token order, and optimization for 1M tokens per recovery seed. We evaluate three seeds on the frozen test after development NLL selects SubShare's planner and hybrid measurement scope.

Mechanism and evaluation. We rank 32 legal Qwen3.5 B67 plans, measure 96 two-intervention cases, and compare functional and curvature signals over 808 directed pairs. Frozen-test token-mean NLL is primary. Six LM Evaluation Harness tasks test downstream transfer (Biderman et al., 2024). Appendices B.1, C, D, E, and J specify the mechanism tests, implementations, data, optimization, and statistical protocol.

6 Results

6.1 Functional assignment wins across backbones and budgets

Table 1 compares assignments across 16 model–budget cells. SubShare beats SharVeT, Random, Adjacent, Depth-spaced, and Weight similarity in every cell, and Geo-Sharing in 15 of 16 cells. Geo-Sharing leads on Gemma 3 1B at B75 by 0.073 NLL.

The common-histogram controls hold the physical structure fixed, so their gaps identify the effect of assigning different pretrained components to the same storage budget. The comparison with SharVeT

Table 1: **Layer-sharing assignment across backbones and physical budgets.** Token-mean frozen-test NLL after matched 1M-token recovery; lower is better. Cells are means over three recovery seeds; mean ± sample standard deviations are reported in Appendix Table 6. Methods retain the same family-wise number of shared bases and use matched rank-128 residuals, teacher, data order, and optimizer. SharVeT retains its native spectral partition. For SUBSHARE, the solver and hybrid-mixer measurement scope are selected by development NLL before the frozen test is evaluated.

Qwen3.5-0.8B					**Qwen3.5-2B**				
Method	B50	B58	B67	B75	Method	B50	B58	B67	B75
Random (2026)	5.017	4.783	4.175	4.104	Random (2026)	4.158	4.359	3.818	3.691
Adjacent (2025b)	4.843	4.657	4.276	4.053	Adjacent (2025b)	4.460	4.350	3.937	3.533
Depth-spaced (2025)	5.055	4.886	4.474	4.226	Depth-spaced (2025)	4.863	4.552	4.252	3.949
Weight sim. (2026)	5.118	5.002	4.416	4.072	Weight sim. (2026)	5.225	4.875	4.523	4.133
SharVeT (2026)	5.455	5.110	4.438	4.179	SharVeT (2026)	4.842	4.360	4.059	3.809
Geo-Sharing (2026a)	4.360	4.106	3.690	3.355	Geo-Sharing (2026a)	3.983	3.720	3.373	3.107
SUBSHARE	**4.199**	**3.982**	**3.642**	**3.328**	SUBSHARE	**3.905**	**3.557**	**3.336**	**3.075**
Gemma 3 1B PT					**Gemma-2B**				
Method	B50	B58	B67	B75	Method	B50	B58	B67	B75
Random (2026)	5.862	4.930	4.548	4.225	Random (2026)	4.762	4.505	4.445	3.731
Adjacent (2025b)	5.366	4.870	4.853	4.076	Adjacent (2025b)	5.230	5.180	4.218	3.552
Depth-spaced (2025)	5.366	4.868	4.193	3.921	Depth-spaced (2025)	6.821	6.764	6.172	5.496
Weight sim. (2026)	5.560	5.208	4.848	4.400	Weight sim. (2026)	5.482	5.281	4.835	4.337
SharVeT (2026)	5.064	4.725	4.668	4.151	SharVeT (2026)	6.284	5.753	4.901	5.143
Geo-Sharing (2026a)	5.185	4.823	4.029	**3.549**	Geo-Sharing (2026a)	5.681	5.677	5.329	4.079
SUBSHARE	**4.807**	**4.491**	**3.970**	3.622	SUBSHARE	**4.641**	**4.379**	**3.995**	**3.539**

extends this test to fused weight–activation similarity and native spectral partitions. Functional assignment improves recovery across both Qwen3.5 scales, full-attention Gemma-2B, and Gemma 3. Its gains under matched residual rank, teacher, token order, and recovery horizon show that the selected groups are easier to repair, connecting local replacement measurements to trained model quality.

6.2 LOCAL COSTS PREDICT GLOBAL RECOVERY

The main grid tests the selected assignments; the 32-plan Qwen3.5 B67 pool tests whether the score orders complete alternatives correctly (Table 2(a)). Whole-mixer cost reaches $\rho = 0.731$ against 1M development NLL and zero top-1/top-3 regret within the pool, compared with $\rho = 0.339$ and 0.0117 top-1 regret for projection-matched scoring. The score transfers to Qwen2.5 with $\rho = 0.929/0.976$ at B50/B67 and zero regret. Across 196 frozen operations, increasing calibration from 8K to 32K tokens preserves rank at $\rho = 0.991$, agrees on 95.6% of pair orders, and selects the same donor for 86.7% of targets.

Both scoring scopes measure end-to-end loss. Whole-mixer scoring intervenes on the donor's projections and local dynamics together; projection-matched scoring retains the target's dynamics. Their different correlations on the same recovered plans show that the scope of a functional intervention matters even when the deployed parameterization is unchanged.

Across 96 frozen pairs, unary sums rank joint damage better than the tested interaction features ($\rho = 0.978$ versus 0.931) and retain higher pair-order accuracy in both parameterizations (Table 2(b)). Explicitly modeling these interactions does not improve the tested ranking. The useful distinction is between detecting non-additivity and improving selection: a more detailed damage model need not give a better ordering.

At fixed projection-matched scope, MILP and Greedy win all 72/72 planner–control comparisons and all 9/9 best-plan budget–seed cells, while differing by only +0.014 NLL on average. Each planner wins a budget. Their shared advantage lies in the directed costs; Greedy offers low-overhead selection and MILP optimizes the constrained partition exactly.

Table 2: **Why local functional costs support global assignment on Qwen3.5-2B.** (a) Projection-matched and whole-mixer costs rank the same 32 complete B67 plans after matched 1M-token recovery. (b) Pairwise interactions are measurable, but explicit interaction features reduce held-out ordering accuracy relative to unary sums. (c) With projection-matched scoring, both SUBSHARE planners beat Geo-Sharing at B50/B67, while Geo-Sharing is strongest at the light-compression B83 boundary. Higher is better for correlation and pair order; lower is better for regret and NLL.

(a) Plan-level proxy validity

Proxy	n	Spearman ρ [95% CI]	Pair order	Top-1 regret	Top-3 regret
Projection-matched	32	0.339 [−0.004, 0.625]	0.627	0.0117	0.0117
Whole-mixer	32	**0.731 [0.497, 0.861]**	**0.770**	**0.0000**	**0.0000**

(b) Direct interaction diagnostic

Param.	n	Rel. $\lvert\kappa\rvert$	Unary	+inter.
Whole	96	0.072	**0.978/0.944**	0.931/0.891
Rank-128	24	0.096	**0.954/0.928**	0.895/0.873

Whole module / rank-128 residual; cells report ρ/pair order.

(c) Planner and curvature boundary

Method	B50	B67	B83
Geo-Sharing (2026a)	3.983	3.373	**2.878**
SUBSHARE-MILP	3.965	**3.336**	2.930
SUBSHARE-GREEDY	**3.905**	3.352	2.931

Three-seed mean frozen-test NLL after 1M tokens.

The extended Qwen3.5-2B frontier includes a B83 crossover: Geo-Sharing leads projection-matched scoring by 0.0518 (Table 2(c)). Curvature–functional agreement remains high across perturbation quartiles ($\rho = 0.704$–0.769), but the smallest quartile trails the largest by 0.044 Spearman (95% target-cluster interval [-0.430,0.168]). Perturbation magnitude alone therefore does not explain the observed B50/B67/B83 crossover.

6.3 FUNCTIONAL PLANS RETAIN MORE ZERO-SHOT CAPABILITY

On six zero-shot tasks, both Qwen3.5-2B B67 planners beat Adjacent throughout and raise the mean from 38.5 to 43.0/44.0 (Table 3). Greedy leads four tasks; MILP leads PIQA and ARC-E. The 4.5/5.5-point gains connect improved recovery NLL to retained task capability, while the task-level differences distinguish the two plans.

6.4 MEASUREMENT IMPROVES THE CAPACITY TRADE-OFF

Table 4 asks whether improved measurement can substitute for additional local capacity. Whole-mixer scoring improves development NLL by 0.064 at the same 1.491B parameter count. Doubling the reference rank to 256 improves NLL by 0.048 but adds 67M parameters. At this budget, better measurement thus yields a larger gain than expanding the residual capacity. Symmetrizing the score costs 0.245 NLL, and weight scoring costs 1.230, highlighting the information carried by directed functional effects.

Keeping full attention local reaches 3.262 NLL, close to whole-mixer's 3.261, but uses 22M more parameters. Rank 64 saves 33M parameters at a 0.024-NLL cost. These alternatives locate the contribution of assignment within the size–quality trade-off: selecting a better shared base can preserve capacity that would otherwise require additional local parameters.

6.5 ASSIGNMENT GAINS PERSIST THROUGH RECOVERY AND REPARAMETERIZATION

Longer recovery tests whether a better initial assignment retains its advantage as the shared model adapts. In the matched Qwen3.5-2B B67 comparison, all six assignments are trained from scratch on a common 5M-token schedule. MILP and Greedy reach frozen-test NLLs of 3.238 and 3.255, compared with 3.652 for Random and 3.736 for Adjacent (Table 14). On Gemma-2B, the matched 1.672B two-loop comparison gives MILP 4.304 versus RRT's 5.655 after 5M tokens, with three-seed standard deviations of 0.003 and 0.006 (Table 17(b)). These results connect early assignment quality to the outcomes of fivefold recovery.

Table 3: **Zero-shot accuracy (%).** Qwen3.5-2B B67 after 1M tokens. Bold: best plan; mean: unweighted.

	Controls		SUBSHARE	
Task	Rand.	Adj.	MILP	Greedy
LAMBADA	13.7	23.0	23.6	**26.4**
HellaSwag	35.0	33.4	39.8	**40.8**
PIQA	60.1	57.5	**65.5**	64.4
ARC-E	39.2	40.4	**49.2**	48.6
ARC-C	23.0	25.4	27.1	**29.4**
WinoGrande	52.6	51.4	52.6	**54.2**
Mean	37.3	38.5	43.0	**44.0**

Table 4: **Measurement and residual capacity shape recovery.** Qwen3.5-2B B67 after 1M tokens. Development NLL; Δ uses projection-MILP; lower is better.

Variant	Params	Dev. 1M ↓	Δ
Projection-MILP ref.	1.491B	3.325	–
Whole-mixer score	1.491B	**3.261**	-0.064
Symmetric score	1.489B	3.570	+0.245
Weight score	1.491B	4.555	+1.230
Full attention local	1.513B	3.262	-0.063
Rank 64	1.458B	3.349	+0.024
Rank 256	1.558B	3.277	-0.048

Cross-parameterization comparisons test whether the assignment criterion depends on shared-base-plus-residual implementation. Keeping Basis Sharing's factorization and recovery fixed, SubShare groups improve 5M test NLL from 3.9444 to 3.5782 on Qwen3.5-2B and from 3.9845 to 3.7471 on Gemma-2B (Appendix B.3). Exact-tying results in Tables 21 and 22 preserve the functional advantage without low-rank residuals. Together, these comparisons identify group selection as a source of quality that complements both parameterization and recovery.

7 DISCUSSION

Sharing assigns functions to logical roles. A donor's suitability depends on where it is used: the target supplies its inputs, residual pathways, and downstream computation. Directed substitution measures this contextual property, whereas symmetrization discards which role must be served. The symmetric-score ablation supports retaining that distinction. At a fixed storage budget, assignment therefore allocates more than weights: it determines which positional differences the shared base already captures and which must be repaired by local residuals.

Assignment and parameterization are complementary. The same functional grouping principle improves both factorized Basis Sharing and exact tying, which offer different mechanisms for expressing positional differences. This transfer locates the value of the measurement before the choice of a compression backend. Within the primary implementation, changing the scoring scope improves quality without adding parameters and gives a larger gain than doubling the residual rank. Better role allocation and richer local capacity are thus distinct levers for the storage–quality trade-off.

Measuring once supports several storage budgets. The Qwen3.5-2B projection-matched matrix takes 26.3 wall-clock minutes and 1.64 GPU-hours, equivalent to 3.3 1M-token recovery runs at 0.50 GPU-hours each. MILP takes 0.228 seconds. Reusing stored costs across budgets amortizes measurement without further model evaluations for group scoring. Development search adds a separate cost: each Qwen grid cell compares up to four scope/planner candidates with three recovery seeds (Appendix B.2).

8 CONCLUSION

SUBSHARE identifies shareable pretrained components by measuring directed functional role transfer in the intact network. The resulting costs support medoid grouping and both Greedy and MILP planning, reusing one matrix across physical budgets. The local-to-global analysis connects this measurement to plan ordering, the quantity needed to select a sharing assignment.

Across four backbones and four budgets, functional assignment improves recovered model quality; on Qwen3.5-2B, these gains extend to zero-shot capability. Comparisons after 5M-token recovery and under Basis Sharing and exact tying show that group selection transfers across recovery horizons and parameterizations. These results motivate a general compression principle: allocate shared capacity to preserve functional roles, and use local residuals for positional specialization.

Reproducibility Statement

Appendix C specifies sharing, initialization, and baseline implementations. Appendices D and E document model and data sources, split construction, preprocessing, and recovery hyperparameters. Plan selection and statistical evaluation are detailed in Appendices B.2 and J, with per-seed results in Appendix H. Section 4 states the theoretical assumptions; Appendix A provides the complete proofs.

AI Use Statement

Generative AI tools were used in the preparation of this work to assist with language editing and polishing, the generation and debugging of experimental code, and the statistical summarization of experimental results. All AI-assisted outputs were carefully reviewed, verified, and, where necessary, revised by the authors. The authors take full responsibility for the content, methodology, experimental results, and conclusions presented in this paper.

References

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vXxardq6db.

Ioannis Athanasiadis, Anmar Karmush, and Michael Felsberg. Grounding functional similarity by invariance-aware model stitching. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yUEFOa32CG.

Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. Relaxed recursive transformers: Effective parameter sharing with layer-wise LoRA. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WwpYSOkkCt.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024. doi: 10.48550/arXiv.2405.14782. URL https://arxiv.org/abs/2405.14782.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Yilong Chen, Zitian Gao, Yihao Xiao, Jason Klein Liu, Xinyu Yang, Yifan Luo, Haoming Luo, Zhengmao Ye, Tingwen Liu, Ran Tao, and Bryan Dai. Polymorphic universal transformer. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39001–39013. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1809. URL https://aclanthology.org/2026.acl-long.1809/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. doi: 10.48550/arXiv.1803.05457. URL https://arxiv.org/abs/1803.05457.

Elia Cunegatti, Marcus Vukojevic, Erik Nielsen, and Giovanni Iacca. From layers to submodules: Rethinking granularity in replacement-based LLM compression. *arXiv preprint arXiv:2606.02559*, 2026. doi: 10.48550/arXiv.2606.02559. URL https://arxiv.org/abs/2606.02559. Accepted to Findings of the Association for Computational Linguistics: EMNLP 2026.

Tamir David-Hay and Lior Wolf. Dynamic layer tying for parameter-efficient transformers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d4uL2MSe0z.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337. PMLR, 2023. URL https://proceedings.mlr.press/v202/frantar23a.html.

Gemma Team. Gemma 3 technical report. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. doi: 10.48550/arXiv.2403.08295. URL https://arxiv.org/abs/2403.08295.

Adriano Hernandez, Rumen Dangovski, Peter Y. Lu, and Marin Soljačić. Model stitching: Looking for functional similarity between representations. *arXiv preprint arXiv:2303.11277*, 2023. doi: 10.48550/arXiv.2303.11277. URL https://arxiv.org/abs/2303.11277.

Cristian Hinostroza, Rodrigo Toro Icarte, Christ Devia, Andres Carvallo, Eugenio Herrera-Berg, Denis Parra, and Jorge F. Silva. Rethinking layer relevance in large language models beyond cosine similarity. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mRLnS8jQWt.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531. Presented at the NIPS 2014 Deep Learning Workshop.

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.

Jaber Jaber and Osama Jaber. Ouroboros: Dynamic weight generation for recursive transformers via input-conditioned LoRA modulation. *arXiv preprint arXiv:2604.02051*, 2026. doi: 10.48550/arXiv.2604.02051. URL https://arxiv.org/abs/2604.02051.

David Jansen, Roman Rausch, Ali Hashemi, David Montero, and Román Orús. LLM compression by block removal with constrained binary optimization. *arXiv preprint arXiv:2602.00161*, 2026. doi: 10.48550/arXiv.2602.00161. URL https://arxiv.org/abs/2602.00161.

Justin Singh Kang, Yigit Efe Erginbas, Landon Butler, Ramtin Pedarsani, and Kannan Ramchandran. Learning to understand: Identifying interactions via the möbius transform. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46160–46202, 2024. doi: 10.52202/079017-1467. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/520b379123d16e41f85472e766846486-Abstract-Conference.html.

Snigdha Chandan Khilar. Cross-layer subspace coupling for LLM compression: A unifying framework and its empirical limits. *arXiv preprint arXiv:2605.30836*, 2026. doi: 10.48550/arXiv.2605.30836. URL https://arxiv.org/abs/2605.30836.

Minkyu Kim, Vincent-Daniel Yun, Youngrae Kim, Suin Cho, Woosang Lim, and Sunwoo Lee. Rethinking layer redundancy: Calibration matters more than search in LLM depth pruning. *arXiv preprint arXiv:2604.24938*, 2026. doi: 10.48550/arXiv.2604.24938. URL https://arxiv.org/abs/2604.24938.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024. doi: 10.48550/arXiv.2403.00745. URL https://arxiv.org/abs/2403.00745.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Guanghao Li, Wenhao Jiang, Li Shen, Ming Tang, and Chun Yuan. Efficient transformer parameter reuse via zero-token mechanism. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14498–14515. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.711. URL https://aclanthology.org/2026.findings-acl.711/.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32431–32454. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24ce.html.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21702–21720, 2023. doi: 10.52202/075280-0950. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/44956951349095f74492a5471128a7e0-Abstract-Conference.html.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204, 2025. doi: 10.18653/v1/2025.findings-acl.1035. URL https://aclanthology.org/2025.findings-acl.1035/.

Yongyu Mu, Yuzhang Wu, Yuchun Fan, Chenglong Wang, Hengyu Li, Jiali Zeng, Qiaozhi He, Murun Yang, Fandong Meng, Jie Zhou, Tong Xiao, and Jingbo Zhu. Cross-layer attention sharing for pre-trained large language models. *Transactions of the Association for Computational Linguistics*, 14: 656–688, 2026. doi: 10.1162/tacl.a.616. URL https://aclanthology.org/2026.tacl-1.30/.

Mohammadmahdi Nouriborji, Morteza Rohanian, and Omid Rohanian. Improving recursive transformers with mixture of LoRAs. *arXiv preprint arXiv:2512.12880*, 2025. doi: 10.48550/arXiv.2512.12880. URL https://arxiv.org/abs/2512.12880.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, 2016. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144/.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Qwen Team. Qwen3.5-2B-Base model card. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-2B-Base.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Dmitriy Shopkhoev, Ammar Ali, Magauiya Zhussip, Valentin Malykh, Stamatios Lefkimmiatis, Nikos Komodakis, and Sergey Zagoruyko. ReplaceMe: Network simplification via depth pruning and transformer block linearization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19487–19517, 2025. doi: 10.52202/085713-0655. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1c10d0c087c14689628124bbc8fa69f6-Abstract-Conference.html.

Jiwon Song, Kyungseok Oh, Taesu Kim, Hyungjun Kim, Yulhwa Kim, and Jae-Joon Kim. SLEB: Streamlining LLMs through redundancy verification and elimination of transformer blocks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46136–46155. PMLR, 2024. URL https://proceedings.mlr.press/v235/song24f.html.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tz9JV2PRSv.

Sho Takase and Shun Kiyono. Lessons on parameter sharing across layers in transformers. In *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 78–90. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.sustainlp-1.5. URL https://aclanthology.org/2023.sustainlp-1.5/.

Jingcun Wang, Yu-Guang Chen, Ing-Chao Lin, Bing Li, and Grace Li Zhang. Basis sharing: Cross-layer parameter sharing for large language model compression. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=gp32jvUquq.

Yiping Wang, Hanxian Huang, Yifang Chen, Jishen Zhao, Simon Shaolei Du, and Yuandong Tian. SHARP: Accelerating language model inference by sharing adjacent layers with recovery parameters. *arXiv preprint arXiv:2502.07832*, 2025b. doi: 10.48550/arXiv.2502.07832. URL https://arxiv.org/abs/2502.07832.

Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, Xu Han, and Zhiyuan Liu. Ultra-FineWeb: Efficient data filtering and verification for high-quality LLM training data. *arXiv preprint arXiv:2505.05427*, 2025c. doi: 10.48550/arXiv.2505.05427. URL https://arxiv.org/abs/2505.05427.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r8H7xhYPwz.

Jeongin Yun, Jaeri Lee, Jongjin Kim, Minjun Kim, Jinho Song, and U Kang. SharVeT: Similarity-aware parameter sharing with vector-based tuning for efficient LLM compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47921–47937. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2213. URL https://aclanthology.org/2026.acl-long.2213/.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Boyang Zhang, Daning Cheng, Yunquan Zhang, and Fangming Liu. Rethinking parameter sharing as graph coloring for structured compression. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=vNCE03zff8.

Xiaojing Zhang, Haifeng Wu, Gang He, Jiyang Shen, Bochen Lyu, and Zhanxing Zhu. MoDr: Mixture-of-depth-recurrent transformers for test-time reasoning. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8a81180649011d1690e05c68bfeaa57c-Abstract-Conference.html.

A Proofs

A.1 Proof of directional asymmetry

Proposition 2 (Substitution is directional)**.** *There exist a model, a data distribution, and two shape-compatible layers i, j for which $d(i \leftarrow j) \neq d(j \leftarrow i)$.*

Consider a two-layer scalar linear model $f(x) = w_2 w_1 x$ with squared loss $\mathcal{L} = \frac{1}{2}(f(x) - y)^2$. Set $x = 1$, $y = 0$, $w_1 = 1$, and $w_2 = 2$. The intact loss is 2. Replacing the first layer by the second gives output 4, loss 8, and damage 6. Replacing the second layer by the first gives output 1, loss $1/2$, and damage $-3/2$. The two layers have identical parameter shape, but the directed damages differ, proving Proposition 2.

A.2 Proof of the Möbius decomposition

For a non-overlapping intervention set S, define $\mu(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} \Delta(U)$.

Proposition 3 (Exact intervention decomposition)**.** *For every finite S,*

$$\Delta(S) = \sum_{\varnothing \neq T \subseteq S} \mu(T). \tag{10}$$

The unary coefficient is $\mu(\{a\}) = d(a)$, and the pairwise interaction is

$$\kappa(a, b) = \Delta(\{a, b\}) - \Delta(\{a\}) - \Delta(\{b\}). \tag{11}$$

Möbius inversion on the Boolean lattice gives $\Delta(S) = \sum_{T \subseteq S} \mu(T)$. Since $\Delta(\varnothing) = 0$, the empty coefficient vanishes. Substituting $|T| = 1$ gives $\mu(\{a\}) = \Delta(\{a\}) = d(a)$. Substituting $|T| = 2$ gives Equation 11.

A.3 Proof of the pathwise bound

Writing H for the Hessian of $\mathcal{L}$, the pairwise interaction satisfies

$$\kappa(a,b) = \int_0^1 \int_0^1 \delta_a^\top H(\theta + s\delta_a + t\delta_b)\delta_b \, ds \, dt. \tag{12}$$

For a complete plan, let $v_{i-1} = \sum_{j<i} \delta_j$. The corresponding pathwise identity is

$$\begin{aligned} &\Delta(\mathcal{A}(P)) - Q(P) \\ &\quad = \sum_{i=2}^{K} \int_0^1 \int_0^1 v_{i-1}^\top H(\theta + sv_{i-1} + t\delta_i)\delta_i \, ds \, dt. \end{aligned} \tag{13}$$

For two directions u, v, define $h(s,t) = \mathcal{L}(\theta + su + tv)$. Applying the fundamental theorem of calculus in both coordinates gives

$$\begin{aligned} &\mathcal{L}(\theta + u + v) - \mathcal{L}(\theta + u) - \mathcal{L}(\theta + v) + \mathcal{L}(\theta) \\ &\quad = \int_0^1 \int_0^1 u^\top H(\theta + su + tv)v \, ds \, dt. \end{aligned} \tag{14}$$

Now telescope the K-intervention loss in the fixed order:

$$\begin{aligned} \Delta(\mathcal{A}(P)) = \sum_{i=1}^{K} \big[&\mathcal{L}(\theta + v_{i-1} + \delta_i) \\ &- \mathcal{L}(\theta + v_{i-1})\big]. \end{aligned} \tag{15}$$

Subtracting $Q(P) = \sum_i [\mathcal{L}(\theta + \delta_i) - \mathcal{L}(\theta)]$ leaves one rectangular difference per i. Applying the two-direction identity with $u = v_{i-1}$ and $v = \delta_i$ proves Equation 13. For disjoint coordinate blocks, $v_{i-1}^\top H \delta_i = \sum_{j<i} \delta_j^\top H_{ji} \delta_i$. Cauchy–Schwarz bounds each term by $\|\delta_j\|_2 \|\delta_i\|_2 \Gamma_{ji}(P)$, where $\Gamma_{ji}(P)$ bounds the relevant pathwise cross-Hessian block. Thus the non-uniform form is

$$|\Delta(\mathcal{A}(P)) - Q(P)| \le \sum_{j<i} \|\delta_j\|_2 \|\delta_i\|_2 \Gamma_{ji}(P). \tag{16}$$

Under the uniform assumptions, each term is at most $\rho^2\gamma$. There are $\binom{K}{2}$ ordered pairs with $j < i$, proving Equation 7.

A.4 Finite calibration

Let m interventions be estimated from n independent proxy sequences, with σ-sub-Gaussian paired damages for each intervention.

Corollary 1 (Mechanism-to-selection bound)**.** *Suppose every feasible plan contains at most K disjoint assignments and obeys the uniform ρ, γ conditions in Proposition 1. If $|R_t(P) - r_t| \le \zeta$ for some common r_t and every feasible P, the population-proxy optimum has regret at most $2[\binom{K}{2}\rho^2\gamma + \zeta]$. For a planner $\widetilde{P}$ that is η-suboptimal for the empirical proxy, with probability at least $1 - \delta$,*

$$\begin{aligned} F_t(\widetilde{P}) - F_t(P^*) \le \eta + 2K\sigma\sqrt{\frac{2\log(2m/\delta)}{n}} \\ + 2\left[\binom{K}{2}\rho^2\gamma + \zeta\right]. \end{aligned} \tag{17}$$

The bound separates solver error, calibration precision, interaction, and construction/recovery mismatch.

Table 5: Curvature–functional agreement by within-component normalized-Frobenius perturbation quartile. Rank statistics are computed within component families before pooling.

Perturbation quartile	n	Spearman ρ	Pair order	Negative functional fraction
Q1-small	203	0.704	0.804	0.000
Q2	201	0.751	0.823	0.000
Q3	203	0.769	0.841	0.000
Q4-large	201	0.748	0.823	0.000

A.5 Proof of surrogate regret

Let $\widehat{P}$ minimize Q and P^* minimize F_t. Then

$$\begin{aligned} &F_t(\widehat{P}) - F_t(P^*) \\ &\quad = \beta[Q(\widehat{P}) - Q(P^*)] + e(\widehat{P}) - e(P^*) \\ &\leq 2\epsilon. \end{aligned} \tag{18}$$

If $Q(\widetilde{P}) \leq Q(\widehat{P}) + \eta$, the same calculation adds $\beta\eta$. For every $P \neq \widehat{P}$,

$$F_t(P) - F_t(\widehat{P}) \geq \beta[Q(P) - Q(\widehat{P})] - 2\epsilon. \tag{19}$$

The stated proxy-gap condition makes this difference strictly positive, so $\widehat{P}$ is the unique recovered-loss minimizer.

For Corollary 1, Proposition 1 implies uniformly over feasible plans that $|\Delta(\mathcal{A}(P)) - Q(P)| \leq \binom{K}{2}\rho^2\gamma$. Writing $R_t(P) = r_t + [R_t(P) - r_t]$ therefore instantiates Equation 8 with $\beta = 1$ and $\epsilon = \binom{K}{2}\rho^2\gamma + \zeta$. Theorem 1 gives the population-proxy claim. For finite calibration, a sub-Gaussian tail bound and a union bound over m interventions yield $\sup_a |\widehat{d}_n(a) - d(a)| \leq \sigma\sqrt{2\log(2m/\delta)/n}$ with probability $1 - \delta$. Every plan contains at most K assignments, so its empirical proxy differs from its population proxy by at most K times this amount. Applying empirical planner optimality twice adds $\eta + 2K\sigma\sqrt{2\log(2m/\delta)/n}$; adding the mechanism residual proves Equation 17.

B Reproducibility and Extended Evidence

This appendix specifies the models, data, parameter accounting, assignments, recovery, and evaluation used in the experiments. Per-seed results and plans are retained in the accompanying machine-readable records.

Appendix guide. Planning and initialization are in Appendix C; data and training in Appendices D–E; extended results and ablations in Appendices H–I; and evaluation details in Appendix J.

Preregistered interaction strata. On the frozen direct pairs, median relative $|\kappa|$ is 0.162 for near versus 0.022 for far targets, and 0.038 in the low- versus 0.138 in the high-unary-severity stratum.The released analysis additionally reports all six component-family pairings with their stratified bootstrap intervals.

B.1 Proxy and mechanism test protocol

The complete-plan pool contains 32 Qwen3.5 B67 partitions with the same family-wise group-size histogram. The first 12 span evenly spaced projection-matched proxy ranks; the remaining 20 complete the pool under the same frozen protocol. Every plan is recovered to 1M tokens. We report Spearman and Kendall correlation, pair-order accuracy, top-1/top-3 selection regret, and plan-bootstrap intervals for projection-matched and whole-mixer scores on identical plans. No plan outcome is used to fit or retune either proxy or planner. Separately constructed whole-mixer B50/B83

plans test prospective scope transfer, Qwen2.5 supplies two frozen-test budget transfers, and 196 preselected atomic measurements repeated at 8K and 32K tokens quantify calibration stability.

Before observing any joint loss, the interaction test fixes 96 pairs of non-overlapping atomic interventions in 24 strata: six component-family pairings crossed with near/far logical distance and low/high unary severity, with four pairs per stratum. Each evaluation uses the same 16 sequences of length 2,048 and stores intact, two unary, and joint per-sequence losses. A residual-aware subset of 24 pairs repeats the intervention with the deployed donor base plus rank-128 truncated-SVD residual. We report paired bootstrap intervals and an eight-fold held-out diagnostic of joint damage. Interaction features pass only if they improve held-out pair order or Spearman by 0.05, or RMSE by 10%. A passing model proposes one B67 plan under the matched unary baseline's frozen per-family singleton/pair/triple histograms, expands from one development seed to three, and releases frozen-test evaluation only after a 0.020-NLL development gain.

The curvature test joins 808 eligible directed Qwen3.5 pairs by projection-matched functional damage, Geo-Sharing cost over the same shareable paths, normalized Frobenius displacement, component family, and depth distance. Displacement is partitioned into within-family quartiles before outcomes are analyzed. The preregistered mechanism requires the smallest-displacement quartile to exceed the largest by at least 0.10 Spearman, with a positive 95% target-cluster interval and the same direction in at least two of three components.

This test is motivated by the local expansion $d(a) = g^\top \delta_a + \frac{1}{2}\delta_a^\top H \delta_a + O(\|\delta_a\|_2^3)$. Curvature is the leading term when the directional gradient contribution is negligible relative to the quadratic term. The quartile analysis tests whether smaller pretrained replacements exhibit closer curvature–functional agreement.

B.2 Common-grid plan selection

The common-grid comparison uses the same development-selection rule in every cell: minimize mean development NLL after matched 1M-token recovery, then evaluate the selected plan on the frozen test for each recovery seed. For the hybrid Qwen checkpoints, the candidate set crosses Greedy and MILP with projection-matched and whole-mixer functional measurements. The selected Qwen3.5-0.8B candidates at B50/B58/B67/B75 are projection-MILP, projection-MILP, projection-MILP, and whole-mixer MILP, respectively. The selected Qwen3.5-2B candidates are projection-Greedy, whole-mixer MILP, projection-MILP, and whole-mixer Greedy. The Gemma checkpoints use their fixed functional-MILP plan because their full-attention blocks do not introduce a hybrid-mixer scope choice. Baselines are fixed before recovery and never enter the SubShare candidate selection. All reported values and per-seed records for Table 1 are retained in the accompanying result manifest.

Development selection uses four candidates per Qwen cell, each with three recovery seeds (12 runs), except Qwen3.5-2B B67, which compares the two projection-matched planners (six runs). Each fixed baseline and each Gemma functional plan uses three runs. The selected checkpoint in every comparison receives the same 1M-token recovery; these counts separately specify the cost of searching over plans.

SharVeT implementation. We implement SharVeT's complete assignment pipeline within each compatible component family (Yun et al., 2026). Structural similarity combines a transformed weight cosine with inverse Frobenius distance. Functional similarity is linear CKA over projected token embeddings from 1,024 calibration tokens, and probabilistic-OR fusion combines the two signals before spectral clustering. We use the public share parts: query, key, and value projections for attention, and up and gate projections for the MLP. For Gated DeltaNet blocks, the corresponding input QKV and gate projections define the shareable mixer operator. The similarity medoid becomes the donor of each cluster. The number of clusters matches the family-wise shared-base budget. We instantiate the resulting spectral partition with the common shared-base plus rank-128 position-specific residual parameterization and run the same recovery protocol as every other assignment. SharVeT retains its native group-size distribution, so its physical parameter count follows that partition rather than the common histogram used by the remaining methods.

At B50, B58, B67, and B75, respectively, SharVeT's physical parameter counts are as follows (in millions). Qwen3.5-0.8B: 573.480, 599.039, 634.297, and 665.099. Qwen3.5-2B: 1,297.113,

Table 6: **Full cross-backbone recovery statistics.** Token-mean frozen-test NLL after matched 1M-token recovery; lower is better. Values are mean ± sample standard deviation over three recovery seeds.

Method	B50	B58	B67	B75
Qwen3.5-0.8B				
Random (2026)	5.017±0.0683	4.783±0.0027	4.175±0.0007	4.104±0.0019
Adjacent (2025b)	4.843±0.0128	4.657±0.0123	4.276±0.0065	4.053±0.0022
Depth-spaced (2025)	5.055±0.0091	4.886±0.0038	4.474±0.0066	4.226±0.0064
Weight sim. (2026)	5.118±0.0080	5.002±0.0037	4.416±0.0046	4.072±0.0015
SharVeT (2026)	5.455±0.0405	5.110±0.0592	4.438±0.0272	4.179±0.0061
Geo-Sharing (2026a)	4.360±0.0016	4.106±0.0033	3.690±0.0015	3.355±0.0002
SUBSHARE	**4.199±0.0005**	**3.982±0.0030**	**3.642±0.0021**	**3.328±0.0003**
Qwen3.5-2B				
Random (2026)	4.158±0.0027	4.359±0.0020	3.818±0.0032	3.691±0.0034
Adjacent (2025b)	4.460±0.0024	4.350±0.0027	3.937±0.0032	3.533±0.0011
Depth-spaced (2025)	4.863±0.0084	4.552±0.0101	4.252±0.0030	3.949±0.0028
Weight sim. (2026)	5.225±0.0178	4.875±0.0124	4.523±0.0104	4.133±0.0050
SharVeT (2026)	4.842±0.0153	4.360±0.0092	4.059±0.0024	3.809±0.0020
Geo-Sharing (2026a)	3.983±0.0018	3.720±0.0017	3.373±0.0013	3.107±0.0014
SUBSHARE	**3.905±0.0027**	**3.557±0.0026**	**3.336±0.0003**	**3.075±0.0009**
Gemma 3 1B PT				
Random (2026)	5.862±0.0598	4.930±0.0090	4.548±0.0308	4.225±0.0044
Adjacent (2025b)	5.366±0.0186	4.870±0.0154	4.853±0.0285	4.076±0.0026
Depth-spaced (2025)	5.366±0.0354	4.868±0.0036	4.193±0.0003	3.921±0.0091
Weight sim. (2026)	5.560±0.0226	5.208±0.0095	4.848±0.0611	4.400±0.0357
SharVeT (2026)	5.064±0.0096	4.725±0.0159	4.668±0.0443	4.151±0.0375
Geo-Sharing (2026a)	5.185±0.0150	4.823±0.0118	4.029±0.0032	**3.549±0.0008**
SUBSHARE	**4.807±0.0182**	**4.491±0.0169**	**3.970±0.0034**	3.622±0.0045
Gemma-2B				
Random (2026)	4.762±0.0077	4.505±0.0455	4.445±0.0370	3.731±0.0168
Adjacent (2025b)	5.230±0.0161	5.180±0.0280	4.218±0.0018	3.552±0.0029
Depth-spaced (2025)	6.821±0.0498	6.764±0.0787	6.172±0.2336	5.496±0.1381
Weight sim. (2026)	5.482±0.0039	5.281±0.0062	4.835±0.0036	4.337±0.0005
SharVeT (2026)	6.284±0.0220	5.753±0.0233	4.901±0.0120	5.143±0.0051
Geo-Sharing (2026a)	5.681±0.0144	5.677±0.0132	5.329±0.0159	4.079±0.0005
SUBSHARE	**4.641±0.0046**	**4.379±0.0028**	**3.995±0.0035**	**3.539±0.0021**

1,383.096, 1,497.522, and 1,597.137. Gemma 3 1B PT: 733.122, 775.671, 818.417, and 877.891. Gemma-2B: 1,631.857, 1,722.886, 1,932.732, and 2,125.146.

B.3 SUBSHARE ASSIGNMENTS FOR BASIS SHARING

Basis Sharing factorizes compatible projection matrices into a shared basis and layer-specific coefficients (Wang et al., 2025a). We isolate the assignment decision by retaining this parameterization, its activation-aware initialization, and the complete recovery protocol while changing which layers share each basis. The structural control groups adjacent compatible layers. The SubShare variants instead use groups produced by directional functional substitutability at the same physical-parameter budget. We select between Greedy and MILP using 1M development NLL, choosing Greedy for Qwen3.5-2B and MILP for Gemma-2B, and keep that choice fixed for both test checkpoints.

Table 7 shows that functional assignment improves the same Basis Sharing parameterization on both backbones. On Qwen3.5-2B, it lowers test NLL by 0.4404 after 1M recovery tokens and by 0.3662 after 5M. On Gemma-2B, the corresponding reductions are 0.2914 and 0.2374. The gains persist as

Table 7: SubShare assignments strengthen the Basis Sharing parameterization on two 2B backbones. Planners are selected by 1M development NLL and evaluated on the frozen test split. Lower NLL is better; bold marks the better assignment at each recovery budget.

Backbone	Assignment	Params (M)	1M NLL	5M NLL
Qwen3.5-2B	Adjacent	1,292.1	4.1535	3.9444
	SubShare-Greedy	1,291.9	**3.7131**	**3.5782**
Gemma-2B	Adjacent	1,637.5	4.2089	3.9845
	SubShare-MILP	1,637.5	**3.9174**	**3.7471**

Table 8: Projection-matched Geo-Sharing comparison on Qwen3.5-2B at B50/B67/B83. Initialization uses recovery seed 20260809; 1M cells are mean ± sample standard deviation over three recovery seeds. The SubShare column selects the better functional planner at each budget. Lower is better.

Budget	Dev init.		Test 1M	
	Geo-Sharing	SubShare	Geo-Sharing	SubShare
B50	7.8876	8.3817	3.9825 ± 0.0018	**3.9048 ± 0.0027**
B67	5.4304	5.6825	3.3734 ± 0.0013	**3.3359 ± 0.0003**
B83	3.5227	3.6980	**2.8784 ± 0.0002**	2.9302 ± 0.0003

recovery increases fivefold, showing that directional substitutability improves the allocation of shared bases under both recovery horizons.

C Planner and Initialization Details

C.1 Geo-Sharing Implementation

Geo-Sharing derives sharing assignments from second-order geometry (Zhang et al., 2026a). We compute a diagonal empirical-Fisher/K-FAC curvature score and use it to select donors and groups under the same shared-base-plus-rank-128-residual parameterization as SubShare. This construction isolates the curvature and functional assignment signals while matching model size and recovery capacity.

We collect statistics on 10,240 development tokens, arranged as 20 non-overlapping 512-token sequences, using teacher next-token cross-entropy. For each shareable linear map at target position i, $A_{i,p}$ is the diagonal second moment of the input activation and $G_{i,p}$ is the diagonal second moment of the output gradient for projection path p. The directed curvature cost of replacing target i by donor j is

$$c^{\mathrm{Geo}}_{i\leftarrow j} = \sum_{p}\sum_{o,d} G_{i,p,o}A_{i,p,d} \times \left(W_{j,p,o,d} - W_{i,p,o,d}\right)^2 . \tag{20}$$

We divide costs by the non-zero median within each compatibility family for numerical conditioning. A MILP then chooses groups and medoids while matching SubShare's complete family-specific group-size histogram. Table 8 gives the projection-matched B50/B67/B83 diagnostic; Table 1 gives the B50/B58/B67/B75 main grid.

C.2 Gemma and RRT Implementation

The RRT-topology comparison uses Gemma-2B's 18 full-attention layers. We form separate attention and MLP compatibility families and retain nine physical bases in each, yielding binary groups throughout. SubShare-MILP and SubShare-Greedy use the Gemma substitutability matrix; Random uses one legal binary partition. Each occurrence receives a rank-128 residual initialized from the truncated pretrained difference. The four-budget main grid uses the group histograms of its corresponding functional-MILP plans.

Table 9: Pretrained base models. Counts refer to the instantiated text-only causal language path.

Checkpoint	Layers	Language-layer layout	Physical params	Role
Qwen3.5-2B-Base	24	$6 \times (3\ \mathrm{GDN} + 1\ \mathrm{Attn})$	1,881,825,088	Primary
Qwen3.5-0.8B-Base	24	$6 \times (3\ \mathrm{GDN} + 1\ \mathrm{Attn})$	752,393,024	Scale transfer
Qwen2.5-1.5B	28	$28 \times \mathrm{Attn}$	1,543,714,304	Architecture transfer
Gemma-2B	18	$18 \times \mathrm{Attn}$	2,506,172,416	Provider and RRT transfer
Gemma 3 1B PT	26	$22\ \mathrm{Sliding} + 4\ \mathrm{Attn}$	999,885,952	Provider and scale transfer

RRT uses the cyclic pairs $(i, i+9)$ from the 18-layer checkpoint. For every paired attention or MLP matrix, the physical base is initialized by the arithmetic mean of its two pretrained weights, and each logical occurrence receives its own rank-128 truncated-SVD residual. The two normalization modules in each repeated block are shared exactly. The instantiated model contains 1,672,124,416 physical parameters, compared with 1,672,161,280 for each SUBSHARE and Random model. All methods use the same Gemma tokenizer, training arrays, teacher losses, optimizer, token order, and independent test.

C.3 END-TO-END PROCEDURE

The complete procedure is:

1. Load the frozen pretrained model and partition component instances into compatibility families by operator type, tensor shapes, and state interface.
2. Evaluate the intact model on D_{proxy}. For every ordered compatible pair (i, j), apply the specified projection-matched or whole-mixer intervention, evaluate the same sequences, and store the paired loss differences from Equation 2.
3. Enumerate legal groups of size one through three. For each group, compute its directed medoid and cost with Equation 3.
4. Supply the functional group costs to either Greedy or MILP planning under the specified per-family base counts. Record the MILP plan's group-size histogram, require every controlled assignment to match it, instantiate each plan, and recount unique parameters from the module graph.
5. For every non-singleton group, initialize its shared base from the medoid, its non-medoid residuals by the truncated-SVD factorization in Equation 5, and the medoid residual to zero. Leave singleton components unchanged.
6. Train only shared bases and local residuals with the common recovery protocol in Section 5. Evaluate development data during protocol construction and the frozen test split after all choices are fixed.

Medoid ties are resolved by the smaller logical index, whereas MILP ties use lower physical size followed by lexicographic group order and Greedy ties use lexicographic group order. The released plan JSON records the planner, compatibility families, groups, medoids, proxy costs, tensor shapes, residual rank, and instantiated parameter counts.

C.4 COMPATIBILITY AND PARAMETER ACCOUNTING

For Qwen3.5, the 24 MLPs form one family. Each MLP group shares the gate, up, and down projection matrices. The 18 GDN mixers form a second family and share `in_proj_qkv`, `in_proj_z`, and `out_proj`. Convolution weights, decay parameters, biases, gates, and normalizations remain local. The six full-attention mixers form a third family and share query, key, value, and output projection matrices only with other full-attention positions. A main ablation forces this family to remain entirely local.

For Qwen2.5-1.5B, the 28 full-attention mixers form one family and the 28 MLPs form another. Attention groups share query, key, value, and output projection weights; MLP groups share gate,

up, and down projection weights. Projection biases and normalization parameters remain position-specific. The same medoid initialization, rank-128 residuals, and fixed-token recovery used for Qwen3.5 are applied without a gate-matching term.

Gemma 3 1B PT has 26 language layers: 22 sliding-window attention layers and four full-attention layers at zero-based positions 5, 11, 17, and 23 (Gemma Team, 2025). The 26 MLPs, 22 sliding-window mixers, and four full-attention mixers form separate compatibility families. Attention groups share query, key, value, and output projections; MLP groups share gate, up, and down projections. Attention-window types, query/key normalizations, and positional mechanisms remain local. Thus sharing preserves the alternating receptive fields as well as the original logical depth.

Embeddings, tied output heads, positional mechanisms, biases, and normalizations remain local in every model. Vision modules and auxiliary training heads are outside the evaluated causal-language path.

For a matrix of shape $d_o \times d_i$, one non-singleton shared group contributes a $d_o d_i$-parameter base and each of its occurrences contributes a rank-r residual with $r(d_o + d_i)$ parameters. A singleton contributes only its unchanged $d_o d_i$-parameter matrix. We count unique tensors by object identity in the instantiated module graph. The exported checkpoint stores each shared base once and reconstructs references from the plan. Serialized bytes are measured after this alias-aware export.

C.5 Optimization Forms

The SubShare-MILP program contains one exact-coverage constraint per logical component and one physical-base-count constraint per family. Histogram-matched assignment procedures, including SubShare-Greedy, additionally respect one count for every family–group-size pair using the MILP plan's realized histogram. SharVeT instead retains its native spectral group sizes at the same family-wise base counts. With group size capped at three, the Qwen3.5-2B functional search contains 3,352 candidate groups across the three family-specific MILPs and solves in 0.228 seconds on Intel(R) Xeon(R) Gold 6161 CPU @ 2.20GHz.

D Artifacts and Data

The artifact package records model and data digests, split manifests, plans, seeds, exact token counts, software versions, GPU assignments, start/end times, and status for every final run. Tables and figures are generated from frozen CSV exports.

D.1 Pinned Upstream Artifacts

Only the causal-language backbone and tokenizer are instantiated. No image examples or vision-module parameters enter the experiment.

D.2 Evaluation Slices and Packing

Recovery data use English shards 00000–00004. The proxy and development pool uses English shard 01822. For Chinese recovery and calibration, Ultra-FineWeb shard 001 is partitioned before tokenization by whether the first 4,096 UTF-8 bytes of a document have a 64-bit BLAKE2b digest modulo 100 below 2. The 98% side supplies recovery tokens and the 2% side supplies calibration tokens. EOS tokens separate source documents before fixed-length packing. Calibration-pool sequences 0–15 form the 32,768-token proxy slice, and sequences 64–95 form the 65,536-token development slice.

The confirmatory test is constructed separately from English shard 00005 and Chinese shard 002, neither of which appears in recovery or calibration. We concatenate 102 English and 26 Chinese 2,048-token sequences and shuffle them with seed 20260814. The resulting 262,144-token array has SHA-256 prefix `1eabf44dbc35`. Its manifest contains the complete digest. The manifest predates checkpoint evaluation. The released manifests record the source-file hashes, document counts consumed during packing, array shapes, token counts, and calibration sequence ranges.

Table 10: Model and data artifacts. ModelScope mirrors were used for transfer speed; repository identifiers and exact file digests in the released manifest define the evaluated content.

Artifact	Repository	Version evidence	License
Qwen3.5-2B-Base	`Qwen/Qwen3.5-2B-Base`	weights SHA-256 `928acbf11878. . .`	Apache-2.0
Qwen3.5-0.8B-Base	`Qwen/Qwen3.5-0.8B-Base`	weights SHA-256 `c2b1e5a17d9c. . .`	Apache-2.0
Qwen2.5-1.5B-Base	`Qwen/Qwen2.5-1.5B`	weights SHA-256 `a961db72e75d. . .`	Apache-2.0
Gemma-2B	`google/gemma-2b`	weights SHA-256 in manifest	Gemma Terms
Gemma 3 1B PT	`google/gemma-3-1b-pt`	pretrained text-only checkpoint	Gemma Terms
FineWeb-Edu mirror	`AI-ModelScope/fineweb-edu-100b-shuffle`	per-file SHA-256 in manifest	ODC-By-1.0
Ultra-FineWeb	`OpenBMB/Ultra-FineWeb`	per-file SHA-256 in manifest	Apache-2.0

Table 11: Frozen recovery configuration. Hyperparameters are selected on development data and then shared across assignment methods.

Category	Setting	Value	Scope
Data	Proxy / dev / test tokens	32,768 / 65,536 / 262,144	test uses independent shards
Data	Recovery pool	160M available tokens	source-disjoint from test
Packing	Train / evaluation length	1,024 / 2,048 tokens	all plans
Sharing	Maximum group size	3	each compatibility family
Sharing	Default residual rank	128	every shared occurrence
Recovery	Screening / extension	1M / 5M tokens	one schedule
Recovery	Batch / GPU count per run	1,024 tokens / 1	four runs in parallel
Optimizer	AdamW β_1, β_2 / decay	0.9, 0.95 / 0.01	shared base + residuals
Learning rate	Peak / minimum ratio	10^{-5} / 0.1	cosine schedule
Schedule	Warm-up	3% (150K at 5M)	token based
Loss	CE / KD / hidden / gate	1.0 / 0.5 / 0.5 / 0.02	all plans
Distillation	Temperature / top logits	2.0 / 32	renormalized top-k
Numerics	Precision / grad clip	BF16 / 1.0	no activation checkpointing

Qwen2.5 and Gemma use the same source shards and language mixture. Each backbone's native tokenizer independently produces a 128-sequence test array, preserving the source-domain comparison while respecting its vocabulary.

E Training Configuration

The teacher is evaluated without gradients. The student optimizer contains each unique shared base once and every position-specific LoRA factor (Hu et al., 2022). Singleton matrices and local GDN dynamics remain frozen. Shared bases use the peak learning rate in Table 11, as do the residual factors. Student hidden states are matched at every logical layer. GDN gate matching is applied only where the teacher and student expose the same gate. The top-32 teacher distribution is renormalized

Table 12: Extended Qwen3.5-2B projection-matched SUBSHARE-MILP frontier with recovery seed 20260809. Initialization uses development NLL; recovered checkpoints use frozen-test NLL.

Base retention	Physical params	Non-embed. ratio	Dev init.	Test 1M	Test 5M
42%	1.207B	0.509	9.692	4.277	4.030
50%	1.292B	0.570	7.990	3.965	3.772
58%	1.377B	0.632	7.451	3.548	3.423
67%	1.491B	0.716	5.682	3.336	3.238
75%	1.593B	0.790	4.578	3.116	3.043
83%	1.687B	0.858	3.698	2.930	2.884
100% teacher	1.882B	1.000	2.593	2.594	2.594

before KL computation (Hinton et al., 2015; Kim & Rush, 2016). All compared plans use the same optimizer steps and token order.

F ADDITIONAL ANALYSES

Direction and proxy residual. Mean $|\delta_{i\leftarrow j} - \delta_{j\leftarrow i}|$ is 0.183 for MLPs, 1.255 for GDN mixers, and 2.047 for full-attention mixers. The corresponding medians are 0.060, 0.381, and 0.176, indicating that a small number of direction-sensitive mixer substitutions account for much of the asymmetry. Symmetrizing the matrix changes development 1M NLL by +0.245. For a complete plan $\mathcal{P}$, the proxy residual is

$$\epsilon(\mathcal{P}) = \ell(M_{\mathcal{P}}, D_{\text{dev}}) - \ell(M, D_{\text{dev}}) - \sum_{G\in\mathcal{P}} q(G). \tag{21}$$

Its median absolute value is 1.914, and the largest value occurs for plan random_028. This residual includes the difference between the constructed shared model on development data and the additive proxy. Its variation, rather than a common offset, can change the ranking of complete plans.

Hybrid constraints and residual capacity. Keeping every full-attention mixer local changes development NLL by -0.063. This tests the effect of an additional architectural constraint. Ranks 64, 128, and 256 obtain NLLs 3.349, 3.325, and 3.277, respectively. Table 4 relates these gains to their physical parameter counts. Across all three Qwen3.5 budgets, exact tying preserves the functional ordering without position-specific residuals. Qwen2.5 exhibits the same result, demonstrating that functional assignment remains effective after removing low-rank residuals in both architectures.

G DEPLOYMENT BOUNDARY AND MEASUREMENT COST

Recovery and the native-model boundary. At the medium budget, extending recovery from 1M to 5M tokens improves frozen-test NLL by 0.098. The resulting shared model is worse than native Qwen3.5-0.8B by 0.419 NLL. The two models represent complementary deployment routes: post-training sharing preserves the teacher's logical computation, while the native model benefits from architecture and pretraining designed for its scale.

Cost of selecting a plan. For the Qwen3.5-2B projection-matched setup, the full proxy matrix costs 26.3 wall-clock minutes and 1.64 GPU-hours. Planning takes 0.228 seconds. One 1M recovery run costs 30.2 minutes and 0.50 GPU-hours. The measurement overhead is therefore equivalent to 3.3 short recovery runs and can be amortized across all budgets derived from one matrix. Appendix B.2 gives the candidate counts used for development selection in the main grid.

H FULL RESULTS

Table 12 reports the complete SUBSHARE-MILP budget frontier, while Table 13 compares the two SUBSHARE planners on the common 5M schedule. Table 14 gives the matched from-scratch 5M comparison of all six B67 assignments. Table 16 gives every Qwen2.5 functional-planner comparison

Table 13: Recovery at three representative Qwen3.5-2B budgets. Initialization uses development NLL; 1M and 5M use frozen-test NLL.

Budget	Method	Physical params	Dev init.	Test 1M	Test 5M
B50	SUBSHARE-MILP	1.292B	7.990	3.965	3.772
	SUBSHARE-GREEDY	1.292B	8.382	3.908	3.729
B67	SUBSHARE-MILP	1.491B	5.682	3.336	3.238
	SUBSHARE-GREEDY	1.491B	5.405	3.350	3.255
B83	SUBSHARE-MILP	1.687B	3.698	2.930	2.884
	SUBSHARE-GREEDY	1.687B	3.682	2.931	2.884

Table 14: Complete matched Qwen3.5-2B B67 recovery from scratch. Development NLL is measured at initialization, 1,000,448 tokens, and 5,001,216 tokens; the frozen test is evaluated only after the final checkpoint. All six methods share the same rank-128 parameterization, data order, and optimization protocol.

Method	Dev init.	Dev 1M	Dev 5M	Test 5M
SUBSHARE-MILP	5.682	3.325	3.229	3.238
SUBSHARE-GREEDY	5.405	3.349	3.255	3.255
Random	7.552	3.810	3.646	3.652
Adjacent	9.331	3.966	3.761	3.736
Depth-spaced	9.902	4.261	4.000	3.989
Weight	10.987	4.554	4.178	4.153

Table 15: Paired frozen-test differences (SUBSHARE-MILP minus comparator) with 95% sequence-bootstrap intervals. Negative values favor SUBSHARE-MILP. The SUBSHARE-GREEDY row is a solver ablation; Random mean averages the three fixed random plans sequence-wise.

Comparator	B50	B67	B83
SUBSHARE-GREEDY	+0.057 [+0.052, +0.063]	-0.014 [-0.019, -0.010]	-0.001 [-0.003, +0.001]
Random mean	-0.500 [-0.522, -0.480]	-0.710 [-0.745, -0.676]	-0.400 [-0.417, -0.383]
Adjacent	-0.494 [-0.512, -0.475]	-0.600 [-0.618, -0.583]	-0.292 [-0.304, -0.281]
Depth-spaced	-0.895 [-0.953, -0.838]	-0.918 [-0.968, -0.868]	-0.631 [-0.671, -0.591]
Weight	-1.280 [-1.337, -1.225]	-1.198 [-1.264, -1.134]	-0.664 [-0.696, -0.633]

on its frozen test for the first recovery seed. The main Qwen2.5 table aggregates a 36-run grid over three seeds; Random uses a different legal partition at each seed, while the other plans vary only the recovery seed. The primary Qwen3.5 robustness table aggregates 54 method–budget–seed cells. Each assignment and token order is fixed across its three recovery seeds, and the separate Random plan-seed analysis measures assignment variation. The Qwen3.5 rank-zero table contains 54 cells across B50, B67, and B83 and uses three disjoint recovery-token slices because nominal seeds do not change residual initialization when no residual exists. The Qwen2.5 rank-zero table contains 36 cells over two budgets and three disjoint recovery-token slices. All rows are also exported in machine-readable form.

I ABLATIONS AND SYSTEMS MEASUREMENTS

The full-attention-local ablation reports its instantiated physical parameters explicitly. It is interpreted as an architectural constraint test rather than an exactly matched global reallocation.

Table 16: Qwen2.5-1.5B paired frozen-test differences (functional planner minus control) with 95% sequence-bootstrap intervals. Negative values favor functional planning.

Budget	Functional planner	Control	Difference [paired 95% CI]
B50	SUBSHARE-MILP	Random	-0.828 [-0.847, -0.810]
	SUBSHARE-MILP	Adjacent	-1.815 [-1.846, -1.785]
	SUBSHARE-MILP	Depth-spaced	-1.341 [-1.380, -1.304]
	SUBSHARE-MILP	Weight	-3.278 [-3.337, -3.225]
	SUBSHARE-GREEDY	Random	-0.459 [-0.472, -0.445]
	SUBSHARE-GREEDY	Adjacent	-1.446 [-1.471, -1.421]
	SUBSHARE-GREEDY	Depth-spaced	-0.972 [-1.004, -0.942]
	SUBSHARE-GREEDY	Weight	-2.909 [-2.965, -2.859]
B67	SUBSHARE-MILP	Random	-0.649 [-0.665, -0.633]
	SUBSHARE-MILP	Adjacent	-1.367 [-1.400, -1.336]
	SUBSHARE-MILP	Depth-spaced	-0.979 [-1.012, -0.947]
	SUBSHARE-MILP	Weight	-4.030 [-4.097, -3.964]
	SUBSHARE-GREEDY	Random	-0.644 [-0.660, -0.629]
	SUBSHARE-GREEDY	Adjacent	-1.362 [-1.394, -1.331]
	SUBSHARE-GREEDY	Depth-spaced	-0.974 [-1.008, -0.942]
	SUBSHARE-GREEDY	Weight	-4.025 [-4.092, -3.962]

Table 17: **Architecture transfer and comparison with recursive sharing.** (a) On full-attention Qwen2.5-1.5B, both SUBSHARE planners beat all four controls at B50/B67 after 1M tokens. (b) In RRT's original two-loop Gemma-2B setting (2025), SUBSHARE-MILP outperforms RRT after both 1M and 5M tokens. Cells report frozen-test NLL as the three-seed mean ± sample standard deviation; lower is better. Physical sizes are 1.005B/1.198B in (a) and 1.672B in (b); – denotes an unextended run. MILP and Greedy denote SUBSHARE planners.

(a) Full-attention Qwen2.5-1.5B

Assignment	B50	B67
Random	4.962 ± 0.085	4.335 ± 0.592
Adjacent	6.028 ± 0.011	4.805 ± 0.033
Depth-spaced	5.550 ± 0.014	4.433 ± 0.008
Weight	7.490 ± 0.021	7.498 ± 0.043
MILP	**4.225 ± 0.002**	**3.462 ± 0.001**
Greedy	4.605 ± 0.010	3.468 ± 0.001

(b) RRT setting on Gemma-2B

Assignment	1M	5M
Random	5.328 ± 0.038	–
RRT	6.243 ± 0.028	5.655 ± 0.006
MILP	**4.558 ± 0.004**	**4.304 ± 0.003**
Greedy	4.790 ± 0.008	–

Table 18: Excess frozen-test NLL over the uncompressed Qwen3.5-0.8B teacher after 1M-token recovery. Cells are mean ± sample standard deviation over three recovery seeds. Plans, rank-128 residuals, physical parameters, and data are fixed within each cell; lower is better.

Method Physical params	B50 568.5M	B67 631.1M	B83 691.6M
SUBSHARE-MILP	1.379 ± 0.000	0.822 ± 0.002	0.382 ± 0.001
SUBSHARE-GREEDY	1.848 ± 0.004	0.930 ± 0.001	0.384 ± 0.000
Random	2.197 ± 0.068	1.355 ± 0.001	0.660 ± 0.001
Adjacent	2.023 ± 0.013	1.456 ± 0.007	0.704 ± 0.002
Depth-spaced	2.235 ± 0.009	1.655 ± 0.007	1.100 ± 0.004
Weight	2.298 ± 0.008	1.596 ± 0.005	0.955 ± 0.002

Table 19: Development-set mechanism ablations at Qwen3.5-2B B67. Rank variants report their actual parameter overhead.

Variant	Mechanism tested	Physical params	Dev init.	Dev 1M
SUBSHARE-MILP	Directed functional assignment	1.491B	5.682	3.325
Symmetric costs	Directionality	1.489B	6.086	3.570
SUBSHARE-GREEDY	Downstream planner	1.491B	5.405	3.349
Weight costs	Functional measurement	1.491B	10.987	4.555
Whole-mixer proxy	Intervention-scope alignment	1.491B	5.508	3.261
Full attention always local	Hybrid compatibility rule	1.513B	4.928	3.262
Rank 64	Residual capacity	1.458B	5.568	3.349
Rank 256	Residual capacity	1.558B	5.731	3.277

Table 20: Held-out-budget intervention-scope transfer on Qwen3.5-2B. Values are frozen-test NLL means over three matched recovery seeds; Δ is whole-mixer minus projection-matched planning, so negative is better. Both scopes use identical per-family group-size histograms and rank-128 recovery.

Budget	Planner	Matched scope	Whole mixer	Δ	Wins
B50	SUBSHARE-MILP	3.965	3.925	-0.041	3/3
B50	SUBSHARE-GREEDY	3.905	3.967	+0.063	0/3
B83	SUBSHARE-MILP	2.930	2.897	-0.033	3/3
B83	SUBSHARE-GREEDY	2.931	2.898	-0.032	3/3

Table 21: Qwen3.5-2B assignment under residual-free exact tying after 1M-token recovery. Cells are mean $\pm$ sample standard deviation over three disjoint recovery-token slices; all methods are parameter matched within each budget. Lower is better.

Method Physical params	B50 1.196B	B67 1.425B	B83 1.653B
SUBSHARE-MILP	4.129 ± 0.004	$\mathbf{3.420 \pm 0.003}$	2.961 ± 0.000
SUBSHARE-GREEDY	$\mathbf{3.988 \pm 0.005}$	3.421 ± 0.001	$\mathbf{2.960 \pm 0.000}$
Random	4.362 ± 0.009	3.998 ± 0.003	3.366 ± 0.002
Adjacent	4.608 ± 0.011	4.237 ± 0.010	3.327 ± 0.000
Depth-spaced	5.345 ± 0.013	4.765 ± 0.006	3.918 ± 0.003
Weight	6.094 ± 0.006	5.078 ± 0.003	4.022 ± 0.005

Table 22: Qwen2.5-1.5B assignment under residual-free exact tying after 1M-token recovery. Cells are frozen-test NLL, reported as mean $\pm$ sample standard deviation over three disjoint recovery-token slices; all methods are parameter matched within each budget. Lower is better.

Method Physical params	B50 888.6M	B67 1.123B
SUBSHARE-MILP	$\mathbf{4.515 \pm 0.008}$	$\mathbf{3.620 \pm 0.002}$
SUBSHARE-GREEDY	4.874 ± 0.010	3.627 ± 0.003
Random	5.609 ± 0.004	4.651 ± 0.004
Adjacent	6.898 ± 0.050	5.406 ± 0.017
Depth-spaced	6.193 ± 0.008	5.036 ± 0.006
Weight	7.664 ± 0.021	7.929 ± 0.039

Table 23: Deployment accounting in BF16. Throughput uses batch one, a 2,048-token prompt, and 256 generated tokens. Each card-level value is the median of five measured runs after two warm-up runs; the table reports the median across four RTX 3090 GPUs.

Model / plan	Physical params	Trainable params	BF16 file	Resident VRAM	Peak decode VRAM	Prefill tok/s	Decode tok/s
Qwen3.5-2B teacher	1.882B	–	3.51 GiB	3.51 GiB	3.76 GiB	5502.8	24.3
SUBSHARE-MILP, B67	1.491B	330.8M	2.78 GiB	2.80 GiB	3.04 GiB	5707.1	21.9
SUBSHARE-GREEDY, B67	1.491B	330.8M	2.78 GiB	2.80 GiB	3.04 GiB	5528.7	22.1
Adjacent, B67	1.491B	330.8M	2.78 GiB	2.80 GiB	3.05 GiB	5547.2	21.9
Qwen3.5-0.8B native	0.752B	–	1.40 GiB	1.42 GiB	1.67 GiB	5460.6	24.7

Table 24: Projection-matched MILP partitions for the extended frontier and scale-transfer diagnostic. Singleton groups are omitted and retained in the released plan files.

Model	Budget	Selected non-singleton groups (one-based positions; medoid in brackets)
Qwen3.5-2B	B50	gdn_mixer: ([1],11,15) ([2],21) ([5],19,23) (7,9,[22]) (10,14,[18]); attention_mixer: ([4],8,12) ([16],20); mlp: ([1],6,12) ([2],8,13) ([3],9,11) ([4],14,15) ([5],10,16) ([7],21,22)
Qwen3.5-2B	B67	gdn_mixer: ([1],11,15) ([5],23) (10,14,[18]) ([13],22); attention_mixer: ([4],12) ([16],20); mlp: ([1],6,12) ([2],8,10) ([3],9,11) ([4],13,14)
Qwen3.5-2B	B83	gdn_mixer: ([5],23) (10,11,[18]); attention_mixer: ([16],20); mlp: ([1],10,12) ([2],9,11)
Qwen3.5-0.8B	B67	gdn_mixer: (7,11,[22]) ([9],23) (10,14,[18]) (15,[21]); attention_mixer: (8,[12],16); mlp: ([1],12,22) ([2],9,10) ([4],13,14) ([6],8) ([7],11)

J EVALUATION AND STATISTICAL DETAILS

Token-mean NLL is computed from the sum of causal token losses divided by the number of predicted, non-padding tokens. Sequence-level bootstrap intervals resample complete 2,048-token sequences. Correlation intervals resample complete plans, and random-plan variation is reported separately from recovery-seed variation.

The six zero-shot tasks use LM Evaluation Harness 0.4.12 with no chat template (Biderman et al., 2024; Paperno et al., 2016; Zellers et al., 2019; Bisk et al., 2020; Clark et al., 2018; Sakaguchi et al., 2020). Table 3 reports `acc` for LAMBADA and WinoGrande, and `acc_norm` for HellaSwag, PIQA, ARC-Easy, and ARC-Challenge. Their evaluated sample counts are 5,153, 1,267, 10,042, 1,838, 2,376, and 1,172, respectively, in that order. LAMBADA measures target-word accuracy; multiple-choice tasks score options by conditional log-likelihood with each task's specified normalization. The reported plans use recovery initialization seed 20260809 and harness random seeds 20260813; all tasks have zero few-shot examples.

Throughput measurements use a warmed process, batch one, and identical prompt/output lengths. On each of four RTX 3090 GPUs, we take the median of five measured runs after two warm-up runs; Table 23 reports the median across GPUs. Peak memory is read after synchronization. Proxy and recovery costs include model loading and evaluation but exclude one-time model download.

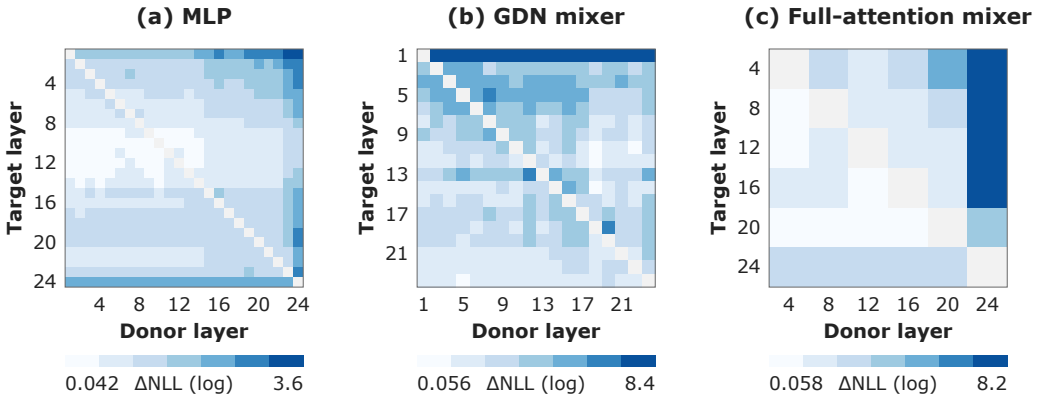

Figure 2: Qwen3.5-2B directed substitution matrices on 32,768 proxy tokens. Rows are targets and columns are donors. Each panel uses a logarithmic, family-specific color scale so both low-cost structure and large failures remain visible.

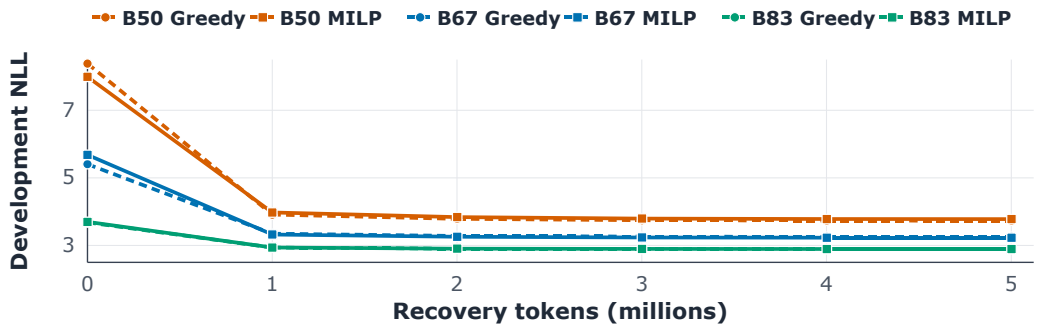

Figure 3: Recovery trajectories on the shared 5M-token schedule. Points denote logged development evaluations; 1M tokens was the screening checkpoint.

K LIMITATIONS

The main comparison evaluates assignment pipelines with a common shared-base and rank-128 residual parameterization. SharVeT retains its spectral partitions at matched family-wise base counts, so its group-size distribution and physical parameter count can differ from the histogram-matched methods. The additional Basis Sharing and exact-tying comparisons test other parameterizations.

The experiments cover the evaluated Qwen and Gemma checkpoints, an English–Chinese recovery mixture, and recovery horizons of 1M and 5M tokens. The six zero-shot tasks evaluate representative Qwen3.5-2B B67 plans. Larger backbones, longer recovery, and additional languages would test how far the observed assignment advantage transfers. Candidate groups contain at most three compatible components, restricting the partitions considered at a given budget. The theoretical bounds require uniform controls on interactions and construction/recovery mismatch; the empirical tests assess plan ordering rather than estimate these constants over every feasible partition.

Sharing reduces unique weights while retaining the original logical depth. Position-specific residuals introduce additional operations, so storage savings do not imply faster decoding. The systems measurements report this trade-off for the evaluated implementation and hardware.