WHEN IS COMPONENT SENSITIVITY COMPOSITIONAL? A CAUSAL STUDY OF NEURAL RESOURCE ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural resource-allocation methods commonly assign each component a scalar sensitivity and rank components independently. This raises two distinct questions: whether the score tracks singleton utility, and whether ranking by singleton utility remains useful under joint allocation. We separate these two questions and test them through direct interventions under a common, logged protocol across quantization, layer skipping, structured pruning, mixture-of-experts selection, and adapter placement.

The resulting behavior forms a spectrum. Layer skipping exhibits small median pairwise interactions, yet one-shot singleton selection can still incur substantial held-out NLL loss at the allocation budget; pruning rankings depend on background sparsity; GPTQ-based group refinement and adapter placement show strong non-additivity; and expert interactions vary across the tested depths. Large additive prediction errors need not imply large selection gaps: the six-layer adapter placement realizes only 15% of its singleton-sum NLL prediction yet trails the tested last-6 heuristic by only 0.011 nats. By contrast, on Qwen2.5-14B one-shot singleton selection at a 25% layer-skip budget incurs 1.013 nats of held-out damage versus 0.540 for sequential re-measurement, and under sparse pruning backgrounds the dense ranking transfers worse than background re-measurement in 16 of 18 conditions—separating additive prediction failure from realized allocation loss. The 14B layer-skip NLL advantage does not yield a statistically resolved accuracy advantage on ARC-Challenge or HellaSwag.

As an exploratory diagnostic, residual-stream perturbation alignment is positively associated with pairwise interaction magnitude in 11 of 12 runs. We conclude that component sensitivity need not behave as an intrinsic scalar: it is a configuration-dependent marginal effect in several tested regimes, and allocation methods should validate compositionality at the deployment granularity, background, budget, and metric.

1 INTRODUCTION

Component importance is the workhorse abstraction of one-shot neural efficiency methods. Pruning removes the "least important" channels, mixed-precision quantization upgrades the "most sensitive" layers, layer-skipping methods bypass the "least contributing" blocks, and expert selection trusts router scores as importance estimates. In this recipe a scalar is attached to a component and a subset is selected by ranking under a budget; the scalar may approximate the effect of changing one component at a reference configuration, but its usefulness for joint selection also depends on how marginal effects change as the configuration is modified.

Underlying this recipe is a quantity we call the *singleton utility* of a component: the causally measured effect of changing that component's resource decision alone—the loss change when one layer is skipped, one channel group is pruned, or one expert is disabled, with everything else held at the base configuration. Ranking by importance hides two distinct assumptions about this quantity. The first is familiar: the *proxy* (magnitude, gradient, Hessian, router probability) must track the singleton utility. The second is rarely stated: singleton utilities must remain useful under joint intervention.

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.

Additivity is one sufficient structure for the ranking rule, but it is not necessary—a ranking can remain optimal even when additive predictions are inaccurate (Section 2)—and when the ranking itself fails to transfer, no improvement to the proxy can in general rescue the recipe: even a perfect predictor of singleton utility can misallocate.

Aggressive W4A4 quantization shows the second failure in extreme form: below operator granularity, refinement utilities are non-additive and sign-unstable while coherent operator-wide interventions remain reliable (Section 3). Is this a quantization quirk or a broader property of resource allocation?

This paper measures the question causally across five allocation domains under a common intervention protocol (Section 2): exact singleton interventions, coherent joint interventions for pairs and budgeted subsets, additivity residuals, background dependence, and selector comparisons. Exact-singleton selection is evaluated in every domain, with quantization restricted to a fixed 12-block candidate space (Table A5). Alongside quantization group refinement, we measure fixed layer skipping and structured pruning of MLP channel groups on the same 1–2B dense models, sparse MoE expert selection as a designed-interaction domain, and training-dependent adapter placement.

The result is a *compositionality spectrum* rather than a universal law:

- **Layer skipping: small pairwise interactions, budget- and configuration-dependent selection quality.** Singleton damage is stable and interactions are modest, yet the one-shot ranking can fail at deployment budgets in larger configurations.
- **Pruning is conditionally compositional.** Channel-group utilities compose well in the dense model but reorder sharply under background sparsity, including sign-level failures.
- **Quantization group refinement is non-compositional.** Coherent joint refinements are stable; component attributions are seed- and configuration-dependent and can reverse sign.
- **MoE expert selection is depth-dependent.** Early/mid layers show large interactions and weak router-to-utility correlation; late layers are substantially more compositional.
- **Adapter placement: NLL saturation with protocol-dependent selection.** Seed-stable singleton utilities strongly overpredict joint NLL utility; at a matched six-layer budget the top-singleton placement cedes 0.011 nats to the tested last-layer heuristic under the packed protocol.

The distinction matters because improving the sensitivity estimator cannot repair a non-compositional allocation target: even a perfect predictor of every singleton utility can misallocate if those utilities change or interact once components are selected jointly. Proxy fidelity remains a separate axis in our experiments: proxy scores track measured singleton effects imperfectly, and exact singleton selection outperforms the tested magnitude-, activation-, gradient-, and router-based scores in nearly every matched comparison.

Contributions. Our contributions are threefold. First, we distinguish proxy fidelity, additive prediction error, and realized selection loss within a common intervention-based framework: large prediction errors can coexist with small selection gaps, and small pairwise interactions with substantial budget-scale loss. Second, controlled comparisons characterize configuration dependence across five domains and scales to 32B: matched-count layer skipping separates removal-count dependence from between-model variation, and a compensation-free quantization control shows the non-additivity does not require GPTQ's cross-column compensation. Third, we establish metric-specific limits of the measured selection gains: large NLL advantages need not produce resolved task-accuracy improvements, and adapter placement rankings can change across NLL evaluation protocols (Table A27). As an exploratory finding, we identify residual-stream coupling as a measurable correlate of pairwise interaction. Representative conditions from the four forward-only domains are re-evaluated on a disjoint C4 corpus at twice the primary sample size; the layer-skip, MoE, and quantization patterns persist while the pruning effects are substantially attenuated (Appendix D).

2 A Common Causal Framework

Utilities and residuals. For a base configuration C_0 and a set S of components whose resource decision is changed coherently, define the configuration utility

$$U(S) = \mathcal{L}(C_0) - \mathcal{L}(C_0 \oplus S), \tag{1}$$

with $\mathcal{L}$ the token-averaged NLL on fixed evaluation sequences. Write $u_i = U(\{i\})$; for a background configuration C, the background-conditioned singleton utility is $u_i(C) = \mathcal{L}(C) - \mathcal{L}(C \oplus \{i\})$.

For removal-type interventions we equivalently report *damage* $D(S) = -U(S)$, retaining its sign (negative damage = the removal improves the loss). The compositional assumption is $U(S) \approx \sum_{i \in S} u_i$; we quantify its failure by the raw residual $R(S) = U(S) - \sum_{i \in S} u_i$, by normalized pairwise interactions $|I_{ij}|/(|u_i| + |u_j| + \epsilon)$ with $I_{ij} = U(\{i, j\}) - u_i - u_j$ ($\epsilon = 10^{-9}$, identical in every domain), and by rank correlation of singleton utilities across backgrounds $\rho\big(u_i \mid C_a,\, u_i \mid C_b\big)$. For removal domains the damage interaction is $I^D_{ij} = -I_{ij}$, so positive damage residuals indicate super-additive damage; tables state which convention each statistic uses.

We further distinguish *local* from *budget-scale* compositionality, evaluating budgeted subsets directly and the gain from sequential re-measurement. Prediction error, rank instability, and realized selection loss are distinct failures. For example, if $U(S) = f\big(\sum_{i \in S} a_i\big)$ for a common strictly increasing saturating f, the singleton ranking can remain optimal at a fixed cardinality even when every additive prediction is inaccurate—so large residuals alone do not establish misallocation. Where candidate sets exist we also report the *selection gap* $\mathrm{Gap}_{\mathcal{C}}(\hat{S}) = \max_{S \in \mathcal{C}} U(S) - U(\hat{S})$ against the best set at the same budget (a candidate-space reference, not a global optimum).

Separating the two failure modes. Our diagnostic matrix compares four selector types at matched budgets: (i) a domain proxy, (ii) *exact singleton utility*, (iii) random selection, and (iv) re-measured sequential greedy. Proxy fidelity is assessed by comparing proxy scores with measured singleton effects in the same background; matched-budget selector comparisons assess the consequences for joint allocation (a proxy–singleton gap compares selection rules, a held-out advantage from sequential re-measurement demonstrates a limitation of the static singleton rule, and a singleton-sum/subset mismatch diagnoses additive prediction error). No selector comparison alone identifies proxy estimation error independently of component interactions. The comparisons instantiated in each domain are detailed in Table A5; for quantization, the exact-singleton arm operates on a pre-specified 12-block candidate space for that amended arm; the released decision log records the reconstructed protocol chronology (Table A23). Because evaluation is deterministic on fixed sequences, reported differences are numerically deterministic on that sample; inferential claims additionally use held-out subsets, repeated seeds, or bootstrap intervals where stated. "Exact" denotes direct intervention-based measurement on a fixed evaluation sample rather than a proxy approximation, not knowledge of population utility; generalization is assessed with held-out sequences, corpus transfer, and the per-condition replications, and the sample-size analysis (Appendix D) covers the 14B layer-skip findings, not every domain or effect size.

Intervention semantics. Each domain requires interventions that keep the evaluated configuration coherent (splicing pieces of solutions optimized under different configurations manufactures spurious interactions; we document this failure mode for quantization in Section 3 and apply the lesson everywhere). Here: layer skipping replaces a decoder layer with the identity; pruning zeroes a channel group of the MLP intermediate (structural removal, no retraining); expert selection masks router logits to $-\infty$ so routing renormalizes among available experts.

Protocol. Models, evaluation splits, replication level, and the selectors tested per domain are summarized in Table A5. The primary causal evaluation uses 8 sequences of 512 tokens from WikiText-2 (seed 314); held-out subsets (seeds 1315/1316) are used where stated; removal domains report damage, while quantization and adapter placement report utility (gain).

We use common operational gates across all production runs; the released decision log documents the protocol and its provenance. The gates are instantiated per domain as follows: layer skipping uses median normalized pairwise interaction > 0.20, or background rank correlation < 0.5 together with a held-out re-measurement advantage > 0.01 nats; pruning uses the median subset residual relative to joint damage > 0.25 or background rank correlation < 0.5, required on at least two target layers; MoE applies the generic interaction/availability thresholds as a case study backed by the frozen-assignment control; the adapter domain requires seed-stable singletons, pair residuals above both $0.20\times$ the singleton sum and twice the training-noise floor, and background correlation < 0.5. These are descriptive decision rules rather than statistical significance thresholds; we report the continuous statistics in full, and a threshold-sensitivity analysis (Table A16) shows the qualitative verdicts unchanged under perturbed thresholds, with the one threshold-sensitive case (the MoE availability criterion) flagged there. Experiments ran on a single NVIDIA H200 (141 GB) node under two production stacks (PyTorch 2.10 + Transformers 4.57; PyTorch 2.13 + Transformers 5.15); every

experiment fits on one GPU and evaluation is deterministic given the fixed sequences and seeds. The experiment scripts, raw JSON artifacts, and an AI-assisted measurement-code review (separate from the implementation pass) are released with the paper.

3 W4A4 Group Refinement: Utilities Do Not Compose

In a standard W4A4 pipeline (SmoothQuant balancing, GPTQ weights at group size 128, grouped 4-bit activations), refining the quantization groups of the MLP down-projection—the operator that dominates operator-level sensitivity on every family we tested—from 128- to 32-wide groups yields a stable quality gain of 0.080–0.117 nats on Llama-3.2-1B, Qwen2.5-1.5B, and SmolLM2-1.7B. The domain question is whether this gain can be attributed to individual weight groups. Components are 128-column blocks (each exactly one set of quantization cells for weights and activations), and coherence requires care: GPTQ propagates error compensation across columns, so splicing blocks from two independently quantized solutions manufactures spurious interactions—with spliced solutions, single-block sums disagree with joints by an order of magnitude and flip sign (Appendix A). We therefore quantize each tested group pattern inside a single GPTQ pass with cached per-layer Cholesky factors, making the coherent configuration the unit of measurement.

We find no stable, seed-independent compositional structure (full decomposition in Table A4). Across three calibration seeds per model the coherent full-refinement gain remains stable (0.080–0.117 nats) while component-level attribution does not: random quarters recover 9–35% of the full gain; the Hessian-weighted proxy—the natural local criterion for where finer quantization helps—recovers 0–42% and beats or loses to random by seed (at one Llama seed the bottom quartile recovers 30% while the top recovers essentially none); and the layer-sum/joint ratio ranges from $+0.03$ to $+2.8$, flipping sign on SmolLM2 (-2.4 to $+2.8$) while the joint total barely moves. In the most extreme example, 62 of 64 single-group refinements in SmolLM2's first layer are individually harmful while refining them jointly helps. These measurements characterize group refinement within the tested W4A4 pipeline. In a compensation-free control (group-wise round-to-nearest in the otherwise identical pipeline; same candidate space, three seeds), the 12-block singleton-sum prediction and joint gain have opposite signs in all three seeds—two joint gains near zero—while the pipeline-wide refinement gain stays stable ($+0.06$–$+0.09$ nats; Table A24): GPTQ's compensation is not required for the observed non-additivity; other quantizer designs remain untested. Within this pipeline the context dependence makes a context-independent additive per-group utility an inadequate allocation primitive, anchoring the non-compositional end of the spectrum. A targeted exact-singleton arm (12-block candidate space fixed before that arm's runs; decision log) tests whether direct singleton measurement yields effective joint selection within the restricted candidate space: exact top-3 selection lies at the 39th–49th percentile of 100 random draws on held-out data across three calibration seeds—a failure to select better than random, not consistently worse-than-random selection—and, separately, its singleton-sum prediction reverses sign against the realized joint gain in two seeds, establishing additive-prediction failure on the selection sample (Table A23). This result is limited to the tested Qwen-1.5B layer and candidate space. Full seed-wise proxy and random-selection results are reported in Appendix A.

4 Layer Skipping: Small Interactions, Budget-Scale Selection Loss

Fixed layer skipping replaces $h_{l+1} = F_l(h_l)$ by $h_{l+1} = h_l$; damage $D(\{l\})$ is measured for every decoder layer, all pairs, and budgeted subsets at 10/25/50% of layers.

At small scales, layer skipping has stable singleton rankings and median normalized pair interactions well below the operational threshold (per-model statistics in Table A6): singleton-damage rankings are highly stable across data subsets, the pairwise interaction median is 1.5–$4\times$ below the 0.20 gate, and after a quarter of the layers are skipped the remaining sensitivities correlate at 0.62–0.86 with their dense-model values. These local diagnostics nevertheless do not by themselves certify the quality of one-shot selection at the allocation budget.

Two refinements matter. First, the interactions that do exist are predominantly super-additive in damage (69–90% of pairs), consistent with redundant pathways absorbing single skips better than combined ones. Second, small median pairwise interactions coexist with substantial selection loss at

Table 1: Budget and metric dependence of layer-skip selection. Top: static−greedy held-out NLL gap (nats/token; paired bootstrap 95% CIs, sets fixed) at matched removal counts. Bottom: zero-shot accuracy (acc_norm, %) of the fixed 14B sets with the static−greedy difference (pp, paired 95% CI) and McNemar p; the NLL gap yields no resolved accuracy gap, and these tests do not establish selector equivalence. Positive NLL gaps favor greedy; positive accuracy differences favor static.

Static − greedy (NLL)	$k = 7$	$k = 12$
Qwen-7B	-0.024 [-0.046, -0.002]	+0.173 [+0.131, +0.217]
Qwen-14B	+0.043 [+0.021, +0.067]	+0.474 [+0.248, +0.749]

14B task	Dense	Static	Greedy	Static−greedy	p
ARC-C	62.3	43.4	43.7	-0.26 [-2.82, +2.39]	0.9
HellaSwag	84.3	70.3	69.6	+0.62 [-0.04, +1.26]	0.069

budget scale: sequential greedy with re-measurement improves held-out utility by 0.088–0.300 nats per token at the 25% budget on all three small models, at the cost of extra selection-time forward passes (Table A30).

Model size and allocation budget change the diagnosis. Repeating the full protocol on Qwen2.5-7B, 14B, and 32B shows that this domain's position is not fixed (Table A26). Pairwise interactions stay small at every scale (median 0.04–0.06), and at the 25% budget on 7B (k=7), static selection has slightly lower held-out damage than greedy (static−greedy -0.024 nats). At 14B (48 layers) the ranking abstraction breaks at budget scale despite the small pairwise terms: after skipping the bottom-25% set, the remaining singleton ranking collapses ($\rho = 0.12$), and re-measured greedy nearly halves the held-out damage of static selection ($+0.47$ nats; 95% CI $[+0.25, +0.74]$; Table A32). A background-variant grid sharpens this: the collapse reproduces on all three evaluation subsets under the bottom-k background but not under random backgrounds (Table A17)—the ranking fails precisely under the background its own deployment would create. At 32B (64 layers) the background gate is crossed again ($\rho = 0.41$, greedy $+0.08$), less severely than at 14B. At 14B, sequential selection costs 510 configuration evaluations versus 48 one-shot ($\approx 10.6\times$ by the measurement-count convention of Table A30). A fixed-k grid separates the two factors the 25% budget confounds (Tables 1 and A28): at matched k=7 the static−greedy gap is -0.024 $[-0.046, -0.002]$ nats on 7B and $+0.043$ $[+0.021, +0.067]$ on 14B; increasing k from 7 to 12 changes the gap from -0.024 to $+0.173$ on 7B and from $+0.043$ to $+0.474$ on 14B, demonstrating within-model budget dependence, while background correlation falls ($0.76/0.40$; $0.28/0.12$). For the fixed 14B skip sets, the NLL advantage is not accompanied by a statistically resolved accuracy advantage on ARC-Challenge or HellaSwag (-0.3 pp p=0.90; $+0.6$ pp p=0.07; Table 1), and both skipped configurations sit far below dense ($p < 10^{-32}$); as in the adapter domain, the deployed metric is part of the configuration.

5 Pruning Is Conditionally Compositional

Structured pruning zeroes one of 16 contiguous channel groups of a target layer's MLP intermediate. Three target layers (early/mid/late) are measured under three backgrounds: the dense model, and 25%/50% of groups pruned in every other layer using activation-magnitude selection, without retraining or weight compensation. These backgrounds test uncompensated structural removal and substantially degrade language modeling (base NLL 2.11–2.86 dense, 2.84–3.87 at 25%, and 4.37–6.78 at 50%; per-condition values in Table A10); the resulting marginal effects characterize these degraded backgrounds, not recovered deployment models. All activation-, gradient-, and product-based proxies are recomputed inside each background.

In the dense model pruning looks compositional: pairwise interactions are small and subset residuals modest. Under background sparsity the picture reverses (Table A1): on Qwen every tested layer falls below the $\rho < 0.5$ gate, twice with *negative* correlation (to $\rho = -0.32$)—the groups safest to prune in the dense model become among the most harmful once other layers are pruned—and on Llama two of three layers cross the gate. At 50% background, pruning the four groups judged least harmful by *background-conditioned* singleton damage improves the loss on one target layer of each small model (-0.071/-0.044 nats), the median subset residual reaching 0.42 of joint damage; with the set held fixed, the improvement is subset-robust on Llama but subset-sensitive on Qwen-1.5B (Table A32). The background-conditioned selector carries the correct sign and attains the lowest

damage among the tested selectors in all nine 50%-background conditions; what fails is transferring the *dense-model* ranking to the sparse configuration. Deploying the dense bottom-k under the sparse background (all sets evaluated held-out) costs a positive penalty over background re-measurement in 16 of 18 conditions, up to $+0.12$ nats at 1–1.5B where the dense set can fall below random; at 14B the penalty stays small despite collapsed rank correlations: rank instability alone does not determine the magnitude of selection loss (Table A20).

Proxy fidelity is a separate diagnostic dimension: Taylor scores correlate weakly or negatively with measured singleton effects in the same background (-0.45 to $+0.14$), and exact-singleton selection outperforms the tested magnitude-, activation-, Wanda-, and Taylor-style proxies in all but one matched condition (the exception by 0.006 nats)—distinguishing inaccurate singleton scoring from the failure of a dense-model ranking to transfer to a sparse background. These comparisons expose two separate risks: improving the proxy does not repair background transfer, while exact singleton measurement does not repair non-composition.

Held-out replications support the configuration dependence (Table A11); negative selected-set damage persists for the tested Llama condition on every subset but is not consistently resolved for the fixed Qwen-1.5B set (Table A32). At scale: on Qwen2.5-14B all three target layers cross the background gate, the middle target shows the residual blowup with sign reversal (0.46, -0.048 nats) and a strongly *anti*-informative Taylor proxy (-0.73 under both backgrounds), and the late target falls to $\rho = -0.42$ (Table A1). Under five different *random* 25% background masks the late layer's dense-model ranking retains only $\rho \approx 0.33$ on average, and attention-head and whole-MLP granularities show the same profile (Appendix A): the effect is specific to neither the activation-magnitude background nor the channel-group definition. The conclusion is conditional rather than global: dense-model importance is not portable to the sparse configuration where the policy will be deployed.

6 Expert Selection: Designed Interaction, Learned Composition

Sparse MoE routing provides a useful test because experts compete under top-k routing, so some interaction is designed into the mechanism. We measure exact singleton expert disables, sampled pairs, and availability-conditioned singletons at early, middle, and late layers of OLMoE-1B-7B, and repeat the depth comparison on Qwen1.5-MoE-A2.7B.

The main result is a consistent early-to-late depth gradient in the tested layers of both architectures (Table A2). In both architectures, early and middle layers show large pair interactions (medians 0.34–0.36 on OLMoE; 0.54 and 0.44 on Qwen1.5-MoE) and weak router-to-causal-utility correlation ($\rho = 0.21$–0.29 and -0.09–0.07 respectively), whereas late layers are much more compositional (interaction 0.08/0.13; router $\rho = 0.58$/0.56). Exact singleton selection also finds 8–30 experts per layer whose removal *improves* the loss. A frozen-assignment control preserves the interaction with rerouting and renormalization removed (medians $0.342/0.300/0.059$ frozen vs. $0.343/0.358/0.080$ rerouted on OLMoE; $0.367/0.478/0.221$ on Qwen1.5-MoE), so routing competition alone does not explain the non-additivity, and the singleton-sum prediction retains only 30–70% of its predicted benefit under joint evaluation. In the primary OLMoE availability conditions (Table A12), restriction reduces the early-layer ranking to $\rho = 0.44$–0.46, with 31% sign flips at 50% availability; the mid-layer correlations are 0.67 and 0.46, and the late-layer correlations are 0.90 and 0.78; the Qwen1.5-MoE late-layer range is 0.64–0.85 (further masks and evaluation subsets in the appendix). Across the tested layers, expert utility therefore becomes more compositional with depth under the same routing mechanism, implicating learned expert functions in addition to routing architecture.

7 Adapter Placement: NLL Saturation, Metric-Dependent Selection

Adapter placement is the training-dependent domain: the intervention changes the learned solution, not just inference. Following the coherent-training principle (each placement is trained from the same initialization protocol and data order; never train-everything-then-delete), we train rank-8 LoRA adapters on `q`/`v`/`down` projections of selected layers in Qwen2.5-1.5B/7B-Instruct for 250 steps on packed GSM8K text, measuring utility as held-out NLL improvement over the frozen base. The instrument is unusually clean: the training-seed noise floor, measured per scale, is $\sigma = 0.0009$ nats

(1.5B) and $\sigma = 0.0006$ nats (7B), with the singleton ranking perfectly seed-stable on the five three-seed-replicated layers at each scale ($\rho = 1.00$)—the replicated effects below are substantially larger than this seed variation. Residual noise multiples use each scale's own equal-variance, independent-noise reference $\sqrt{3}\sigma$, not a significance test; the estimate summarizes training-seed variation and does not quantify uncertainty over held-out examples.

Singleton utilities are large, positive, and concentrated in late layers ($+0.08$–0.11 nats, layers 21–27), but joint utility strongly saturates (Table A7): two layers that each contribute $+0.10$ alone contribute far less than $+0.20$ together, consistent with overlapping learned corrections. At 1.5B, the pairwise residual has median 0.73 of the singleton sum ($\approx 50\times$ the $\sqrt{3}\sigma$ reference scale; 60% negative), and the top-singleton budget set realizes only 15% of its summed prediction and loses to the naive last-k heuristic by 0.011 nats; adapting all 28 layers is worse than adapting none. Operating-point controls at half/double learning rate and doubled steps preserve the same ordering—the all-layer configuration degrades *further* with more training (-0.029 to -0.134)—arguing against simple undertraining.

At 7B the effect is stronger: a single adapted layer captures $+0.42$ nats—as much as adapting all 28 layers ($+0.39$) or any six ($+0.41$–0.43)—pair residuals are negative for 100% of sampled pairs ($\approx 333\times$ the same reference scale), the top-singleton set realizes 17% of its prediction, and background marginal utility is nearly uncorrelated with singleton utility ($\rho = 0.14$, 82% sign flips). The 7B singleton sum (2.50 nats) even exceeds the base NLL (1.56), so additivity is infeasible on boundedness grounds—saturation by construction. The measured singleton utility therefore largely reflects adaptation value shared across placements rather than a unique layer contribution; this establishes saturation of the measured NLL benefit without identifying whether the shared correction concerns reasoning, output format, or other token-level regularities, and the robust-EM evaluation below resolves no corresponding accuracy improvement. A token-segment decomposition on shared reference tokens (Table A25) locates the improvement: format tokens have the largest per-token gain (≈ 5.8–6.4 nats/token vs 0.15–0.55 on solution tokens), but solution tokens account for ≈ 59–65% of the total improvement because they are far more numerous—the NLL improvement is not confined to formatting. On the example-level evaluation, last-6 has the greater mean gain at 1.5B whereas top-singleton-6 has it at 7B (per-seed paired bootstrap CIs exclude zero; Table A27); for 7B this reverses the packed-protocol ordering of Table A7. The evaluations differ in example set and context handling, so the reversal does not isolate the effect of either change; the units also differ (nats/example vs nats/token), but a common rescaling of identical token losses alone would not reverse an ordering. Separately, the task audit below shows that legacy strict-EM gains are sensitive to output-format compliance and continuation parsing; they cannot be read directly as gains in arithmetic accuracy. The key point is therefore not estimation noise: the singleton quantity is directly measured and highly reproducible, yet it remains the wrong primitive for predicting joint placement.

Task-level evaluation: the metric is part of the configuration. Legacy strict-format GSM8K exact match—taking the last #### answer in the full completion—appears to induce different placement rankings at the two scales (all-28 at 1.5B, top-singleton-6 at 7B). Auditing the stored generations shows this metric mixes marker compliance with continuation behavior: many completions contain a hallucinated follow-up problem and more than one answer marker. We therefore report two boundary-aware scores using the audited, unit-tested parser (Table A21; parser, stored generations, and audit annotations released): segment-strict EM requires a #### answer before the first validated follow-up-problem boundary, whereas robust EM uses the same segment but falls back to its last number.

The unadapted bases score 0% under both marker-required metrics but 49.9% and 73.5% under robust EM; adapted placements span only 45.9–47.2% (1.5B) and 72.6–73.8% (7B). No robust-EM contrast among adapted placements is statistically resolved (exact paired McNemar $p \geq 0.17$; paired 95% bootstrap intervals include zero; Table A22). Against the base, no contrast is resolved at 7B (all $p \geq 0.41$); at 1.5B the base is 2.7–3.9 points higher, three of four contrasts surviving Holm correction (adjusted $p \leq 0.044$; top-singleton-6 $p = 0.060$).

Thus the large legacy-strict differences are not evidence of resolved arithmetic gains: they primarily reflect configuration-dependent output format and continuation behavior. NLL remains placement-dependent, but robust EM resolves neither a placement ordering nor an adaptation gain over the base on this test set; validation must use the metric that will actually be deployed.

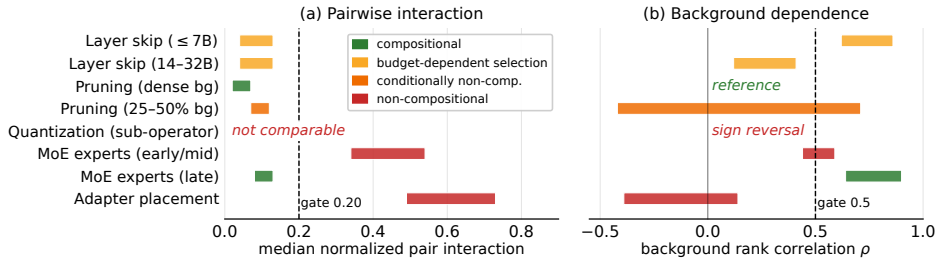

Figure 1: The spectrum across domains (measured ranges; source Table A3): (a) median normalized pair interaction, gate 0.20; (b) background rank correlation, gate 0.5; colors encode verdicts. Both layer-skip rows are budget-dependent: small pairwise interactions do not certify budget-scale selection quality (Table A26), the 7B/14B fixed-k comparisons show within-model budget dependence (Table 1), and the background correlations correspond to the 25% budget. The quantization row is text-only: its sign-flipping decomposition statistics are not directly comparable to the other domains' normalized pairwise interaction.

The NLL controls are complementary: marginal utility on an already-adapted background inverts relative to singleton utility ($\rho = -0.39$, 73% sign flips), and in the exhaustive late-layer triple search (all 20 three-of-six placements) the singleton sum overpredicts every joint utility by roughly $3\times$ while the sum-selected triple trails the best by only $+0.003$ nats across three paired seeds (range 0.001–0.005; Table A32)—large prediction error with a small candidate-space selection gap.

This domain sharpens the central finding: the singleton utility is a directly measured, seed-stable quantity—the best case any proxy could aspire to—yet its sum wildly mispredicts and its NLL ranking still cedes 0.011 nats to a structural heuristic, a pattern consistent with joint outcomes depending substantially on adaptation value shared across placements rather than on uniquely localized layer contributions. These are NLL selection gaps; under robust EM they do not correspond to a resolved ordering in arithmetic accuracy.

8 The Compositionality Spectrum

A working interpretation of what separates the domains. The ordering in Figure 1 is consistent with a working interpretation based on how strongly interventions share downstream state or capacity. Skipping a layer preserves computations upstream of that layer but changes the inputs to all downstream blocks, so small loss interactions need not imply additive representation changes: in the measured Qwen-1.5B run, representation and readout contributions have opposite signs for 97% of sampled pairs (Appendix F). Pruning in a dense model removes redundant capacity, but once background sparsity consumes the redundancy the marginal contribution of surviving groups changes, and dense-model singleton measurements describe a network that no longer exists. Quantization decisions can interact through the shared low-bit forward computation and, in GPTQ, the compensation path—consistent with the strong context dependence of sub-operator utilities. MoE adds designed competition—the same mechanism can be compositional or not depending on what the layer has learned. Adapter placement couples through a shared *learning target*: many layers recover much of the same benefit, and joint utilities saturate far below singleton sums. The working hypothesis: compositionality degrades as interventions share more downstream state, redundancy, routing competition, or adaptation capacity.

A measurable correlate. If non-compositionality arises partly from interventions perturbing shared downstream state, then singletons inducing similar representation changes should interact more strongly when applied together; we test this with the final residual stream. For intervention i, let Δh_i be its change to the final residual stream (input to the final normalization, flattened over evaluated tokens); for a pair (i, j), the *representation coupling* is $C_{ij} = \langle \Delta h_i, \Delta h_j \rangle / (\|\Delta h_i\|_2 \|\Delta h_j\|_2)$. Because interaction strength is measured in magnitude we compare it with $|C_{ij}|$, treating aligned and anti-aligned perturbations as equally strong coupling; the statistic needs only the singleton forward passes already used for utility, not the joint intervention. The final residual stream, just before the frozen readout, is a common measurement location across depths; the statistic thus includes all downstream propagation and is not local coupling at the intervention site.

We compare $|C_{ij}|$ with the observed normalized pairwise interaction across 2,856 pairs from layer skipping, pruning, and MoE (Fig. 2). Within-run correlations are positive in 11 of 12 runs, ranging -0.13 to 0.56 (the four MoE runs: 0.23–0.36; Table A35). The pooled correlation is 0.53 but mixes within- and between-run variation, so it is not within-configuration predictive accuracy.

The association concerns interaction magnitude (signed variants in Table A35); the representation–readout decomposition indicates that multiple mechanisms contribute. It is a measurable correlate of shared downstream representational effects; Appendix F gives the protocol and per-run statistics, including a collapsed-range pruning condition.

Practically: measure $O(n)$ singletons plus a diagnostic sample of pairs or subsets before trusting a ranking, validating gradient-based score signs against exact interventions. Where utilities compose at the deployment budget, exact singleton measurement is a strong first-order rule; where conditional, re-measure under that background; otherwise use scores only to propose candidates, validating them at the deployed metric.

9 Related Work

Magnitude, first-order Taylor (Molchanov et al., 2019), and Wanda (Sun et al., 2024) rank components under a sparsity budget; SparseGPT (Frantar & Alistarh, 2023) already captures some within-layer dependencies; HAWQ (Dong et al., 2019) allocates bit-width by sensitivity; GPTQ, SmoothQuant, and QuaRot (Frantar et al., 2022; Xiao et al., 2023; Ashkboos et al., 2024) are low-bit mechanisms coupling component decisions—we ask when a context-independent singleton ranking survives joint intervention. ShortGPT ranks blocks by input–output similarity (Men et al., 2024); SLEB evaluates block removal by language-model loss, re-scoring after each removal (Song et al., 2024). Our layer-skip comparison follows this static-versus-sequential distinction, measuring how pairwise diagnostics relate to budget-scale selection loss; early-exit and LayerDrop methods (Teerapittayanon et al., 2016; Xin et al., 2020; Fan et al., 2020) instead allocate or train per-input depth. Sparse-MoE routing and load-balancing work (Lepikhin et al., 2021; Fedus et al., 2022; Zhou et al., 2022) evaluates load balance, stability, efficiency, and downstream quality, while we ask whether router scores rank experts' causal utility and whether utilities compose. LoRA (Hu et al., 2022) restricts adaptation to a small trainable set and AdapterDrop (Rücklé et al., 2021) removes lower-layer adapters with limited loss, motivating placement as a resource-allocation problem; we measure every singleton and coherently trained joint placement. Shapley attribution (Shapley, 1953; Lundberg & Lee, 2017) and Shapley pruning (Adamczewski et al., 2024) handle coalition-dependent contributions but are combinatorial; COAR (Shah et al., 2024) *learns* coefficients from ablated configurations—a target our evidence does not rule out. On our released pruning subsets, a linear mask surrogate attains held-out R^2 0.78–0.96 but beats background-conditioned singletons in only one of three conditions despite $10\times$ more measurements (Table A31).

10 Limitations and Conclusion

The comparison is diagnostic, not factorial: definitions, semantics, and budgets differ across domains, so the spectrum is comparative, not causal. The primary evaluations use eight 512-token WikiText-2 sequences ($\approx$4,100 scored tokens; held-out subsets and C4 replication: Appendix D). Scale coverage is uneven (skip 1–32B, pruning 1–14B, adapters 1.5–7B, quantization/MoE $\leq$3B-active). Fixed-length perplexity is the primary metric; task-accuracy checks cover adapter placement and the 14B skip sets, where the measured NLL advantages do not yield statistically resolved accuracy improvements. One scale does not certify another—scale comparisons change width, depth, and pretraining together—and the sampled domains are a small subset of resource-allocation problems (Appendix E). Proxy fidelity, additive prediction accuracy, and joint selection quality are distinct; evaluate allocation rules under the deployed configuration and metric.

Reproducibility Statement

Section 2 specifies the intervention semantics, models, datasets, evaluation sequences, random seeds, operational diagnostic gates, and the released decision log used throughout the study. Appendix A reports the complete per-domain measurements and replications, while Appendix C specifies the

software and hardware environment. We provide anonymized supplementary code, raw JSON artifacts, and table-generation scripts. Detailed appendix and audit tables are generated mechanically, while the curated summary tables are traceable to the same released artifacts. The supplement also includes the AI-assisted code review of the measurement code and its dispositions. These materials support both verification of the reported measurements and independent rerunning of the interventions.

AI USE STATEMENT

Generative AI tools were used in an assistive capacity under the authors' direction: for drafting and editing portions of the manuscript, for literature discovery and reference formatting, and for assistance with code implementation, review, and debugging. The blinded parser-audit annotations released with the supplement were produced by an AI assistant under the stated blinding protocol and are labeled as such in the artifacts. The tools also provided feedback that the authors weighed when refining experimental designs and analyses. The research questions, hypotheses, experimental designs, and conclusions are the authors' own. All AI-assisted code and analyses were inspected by the authors. Model-performance measurements were computed by the experimental pipeline. Parser-audit agreement statistics compare parser outputs with the explicitly identified AI-generated annotations and do not constitute independent human validation. The authors take full responsibility for the final content, claims, and artifacts.

REFERENCES

Kamil Adamczewski, Yawei Li, and Luc van Gool. Shapley pruning for neural network compression. *arXiv preprint arXiv:2407.15875*, 2024.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

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.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

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.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11264–11272, 2019.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7930–7946. Association for Computational Linguistics, 2021.

Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and editing predictions by modeling model computation. In *International Conference on Machine Learning*, 2024.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

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*, 2024.

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.

Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. BranchyNet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition*, pp. 2464–2469. IEEE, 2016.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pp. 38087–38099, 2023.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2246–2251. Association for Computational Linguistics, 2020.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7103–7114, 2022.

A Full per-domain tables

Table A1: Pruning, singleton-rank stability across backgrounds: Spearman ρ between the target layer's 16 group damages in the dense model and under global background pruning (bold: below the $\rho < 0.5$ gate).

	Llama-1B (layers 2/8/14)			Qwen-1.5B (layers 3/14/24)			Qwen-14B (layers 6/24/42)		
Background	L2	L8	L14	L3	L14	L24	L6	L24	L42
25% pruned	.71	**.27**	**.40**	**.25**	**−.28**	.61	**.13**	**−.01**	**.21**
50% pruned	.70	**.42**	**.03**	**−.20**	**−.32**	**.02**	**.23**	**−.05**	**−.42**

The detailed tables imported below are generated mechanically from the released measurement artifacts; the curated summary tables above them transcribe the same artifacts, and the supplement includes the generation scripts and source JSON for auditing every reported value. Tables A8–A15 cover, in order: the quantization group decomposition (Table A8); layer skipping across all six models with bootstrap intervals and held-out budget evaluations (Table A9); all pruning conditions (Table A10) and the held-out late-layer replications (Table A11); all MoE runs including the frozen-assignment controls, evaluation-subset replications, and the second architecture (Table A12); adapter

Table A2: MoE expert selection, both architectures. Proxy ρ: Spearman of router mean probability vs. exact singleton disable damage. Availability ρ: singleton ranking under 25%/50% random expert removal vs. full availability. Selection: damage of disabling the 8 least-useful experts by each selector (negative = disabling improves the loss). Positive damage interactions indicate super-additive damage: $I_{ij}^D = D(\{i,j\}) - D(\{i\}) - D(\{j\}) = -I_{ij}$.

Layer	Proxy ρ	Pair interaction	Frac. $I_{ij}^D{>}0$	Avail. ρ (75/50)	Sign flips	Select exact / router / random
OLMoE-1B-7B (64 experts, top-8)						
L2 (early)	.21	.343	.45	.44 / .46	.29 / .31	$-.0096$ / $+.0061$ / $+.0087$
L8 (mid)	.29	.358	.60	.67 / .46	.25 / .38	$-.0036$ / $+.0092$ / $+.0060$
L14 (late)	.58	.080	.71	.90 / .78	.08 / .16	$-.0045$ / $+.0045$ / $+.0287$
Qwen1.5-MoE-A2.7B (60 routed experts + shared, top-4)						
L3 (early)	$-.09$	.543	.20	.49 / .54	.22 / .27	$-.0053$ / $+.0036$ / $+.0024$
L12 (mid)	.07	.439	.26	.59 / .55	.38 / .23	$-.0088$ / $-.0059$ / $-.0002$
L21 (late)	.56	.133	.36	.85 / .64	.11 / .17	$+.0001$ / $+.0010$ / $+.0124$

Table A3: The spectrum across domains (headline statistics; the source of Figure 1). Quantization entries are qualitative: their decomposition statistics are not directly comparable to the normalized pairwise interaction used for the other domains, and the background statistic is the layer-sum sign reversal rather than a rank correlation.

Domain	Singleton stability	Pair interaction	Background dependence	Verdict
Layer skipping (1–32B)	.97–.99 at $\leq$7B; .86–.96 at 14–32B	low (.04–.13)	mild at $\leq$7B; collapses at 14–32B (.12–.41)	small pair interactions; budget-dependent selection loss
Pruning, dense bg (1–14B)	high	low (.02–.07)	—	compositional
Pruning, sparse bg (1–14B)	—	grows (to .12)	severe ($-.42$–$.71$), sign flips	conditionally non-compositional
Quantization, sub-operator (1–2B)	low	not directly comparable	layer-sum sign reversal	non-compositional
MoE experts, early/mid (2 archs)	moderate–none	high (.34–.54)	strong (.44–.59), flips	non-compositional (learned)
MoE experts, late (2 archs)	high	low (.08–.13)	mild (.64–.90)	approx. compositional
Adapter placement (1.5–7B)	1.00 on the five seed-replicated configurations	extreme (.49–.73, 50–333$\times$ propagated noise)	weak or inverted ($-.39$ to .14)	non-compositional (redundancy $\rightarrow$ saturation)

Table A4: Quantization group refinement (consistent mixed-group GPTQ), ranges over three calibration seeds per model. Shares are fractions of the full-refinement gain; the last column is $\sum_l U(\text{layer } l) \,/\, U(\text{all layers})$, whose sign flips across seeds on SmolLM2.

Model (3 seeds)	Full gain (nats)	Random 25%	Scored 25%	Layer-sum/joint ratio
Llama-1B	0.091–0.104	14–35%	0–34%	$+0.03$ to $+1.8$
Qwen-1.5B	0.099–0.117	9–21%	20–28%	$+0.9$ to $+2.7$
SmolLM2	0.080–0.094	10–34%	26–42%	-2.4 to $+2.8$

Table A5: Per-domain protocol: models, replication, and selectors tested. All checkpoints are the instruction-tuned variants (`Qwen2.5-*-Instruct`, `Llama-3.2-1B-Instruct`, `SmolLM2-1.7B-Instruct`, `OLMoE-1B-7B-0924-Instruct`, `Qwen1.5-MoE-A2.7B-Chat`); short names are used throughout.

Domain	Models	Replication	Selectors tested
Quantization	Llama-1B, Qwen-1.5B, SmolLM2	3 calibration seeds each	exact singleton (candidate space), scored proxy, random
Layer skip	Llama-1B, Qwen-1.5B/7B/14B/32B, SmolLM2 (6 total)	singletons+selections on 3 subsets; 14B bg grid 3×3	singleton, cosine proxy, random, greedy
Pruning	Llama-1B, Qwen-1.5B, Qwen-14B	decisive layers on 3 subsets; 5 random backgrounds	singleton, 4 score proxies, random
MoE	OLMoE, Qwen1.5-MoE	3 subsets + frozen control + 3 masks (OLMoE)	singleton, router prob./freq., random
Adapter	Qwen-1.5B, Qwen-7B	3 training seeds (selections, noise floor); 3 operating-point variants	singleton, structural heuristics, random, joint search

Table A6: Layer skipping. Interactions are $|I_{ij}|/(|D_i| + |D_j| + \epsilon)$ over all pairs; background ρ compares singleton damage before and after skipping the bottom-25% set; greedy improvement is static-minus-greedy held-out damage at the 25% budget (positive = greedy better). Positive damage interactions indicate super-additive damage: $I_{ij}^D = D(\{i,j\}) - D(\{i\}) - D(\{j\}) = -I_{ij}$.

Model	Singleton stability	Interaction [CI95]	Frac. $I_{ij}^D{>}0$	Background ρ	Greedy improvement
Llama-1B	.985/.988	.133 [.080, .172]	.90	.62	$+0.300$
Qwen-1.5B	.984/.984	.051 [.043, .060]	.69	.86	$+0.261$
SmolLM2	.975/.970	.078 [.061, .101]	.79	.63	$+0.088$

Table A7: Adapter placement (GSM8K, 6-layer budget unless noted). Multi-seed rows: mean$\pm$std over 3 training seeds; single-best-layer and all-28 rows are single runs; sum-pred combines single-seed singleton measurements for the top-singleton set.

Model	Selector	Layers	Utility	Sum prediction
Qwen-1.5B	Top-singleton	21–25, 27	$+0.088 \pm .002$	(sum-pred $+0.596$)
	Last-6 heuristic	22–27	$\mathbf{+0.099 \pm .002}$	
	Evenly spaced	0, 5, 10, 16, 21, 27	$-0.009 \pm .002$	
	Random (3 draws)	—	-0.019 to $+0.020$	
	All 28 layers	—	-0.029	
Qwen-7B	Single best layer	18	$+0.420$	
	Top-singleton	3, 18–21, 23	$+0.415 \pm .001$	(sum-pred $+2.50$)
	Last-6 heuristic	22–27	$\mathbf{+0.434 \pm .001}$	
	Random (3 draws)	—	$+0.410$ to $+0.419$	
	All 28 layers	—	$+0.387$	

Table A8: Quantization group refinement, full decomposition statistics for all nine (model, calibration-seed) runs. Random-fraction gain shares are means over 3 draws; scored selects by the Hessian-weighted error-reduction proxy.

Model	Gain	Rnd 1/16	1/8	1/4	1/2	Scored 1/4	Bottom 1/4	Layer sum/joint
Llama-1B / s42	0.091	-0.06	0.19	0.35	0.35	0.19	0.14	0.032 / 0.091
Llama-1B / s43	0.104	0.04	0.21	0.14	0.53	0.34	0.12	0.182 / 0.104
Llama-1B / s44	0.096	-0.00	-0.02	0.14	0.36	0.00	0.30	0.003 / 0.096
Qwen-1.5B / s42	0.110	0.12	0.32	0.19	0.56	0.20	0.20	0.297 / 0.110
Qwen-1.5B / s43	0.099	0.00	0.14	0.09	0.49	0.28	0.09	0.120 / 0.099
Qwen-1.5B / s44	0.117	0.08	0.09	0.21	0.42	0.25	0.15	0.103 / 0.117
SmolLM2 / s42	0.080	-0.24	0.06	0.10	0.49	0.26	-0.05	-0.100 / 0.080
SmolLM2 / s43	0.092	0.00	0.02	0.19	0.55	0.30	0.19	-0.219 / 0.092
SmolLM2 / s44	0.094	0.30	0.29	0.34	0.54	0.42	0.22	0.268 / 0.094

Table A9: Layer skipping, full per-model statistics. Stability: min/max Spearman of singleton damage across the two held-out subsets. Interaction: median normalized pairwise residual with bootstrap CI95 and fraction of positive (damage-super-additive) residuals. Background: Spearman and sign-flip rate of re-measured singletons after skipping the bottom-25% set. Budget-25% damage (nats): static-singleton and re-measured greedy on held-out subsets, random on the selection subset (mean of 5 draws).

Positive damage interactions indicate super-additive damage, with $I_{ij}^{D} = D(\{i,j\}) - D_i - D_j = -I_{ij}$.

Model	L	Stability	Interaction [CI95]	Frac. $I_{ij}^{D}>0$	Bg. ρ	Flips	Static	Greedy	Random
Llama-3.2-1B	16	0.985/0.988	0.133 [0.080, 0.172]	0.90	0.62	0.00	1.849	1.548	7.724
Qwen2.5-1.5B	28	0.984/0.984	0.051 [0.043, 0.060]	0.69	0.86	0.00	0.776	0.514	5.085
SmolLM2-1.7B	24	0.970/0.975	0.078 [0.061, 0.101]	0.79	0.63	0.00	0.992	0.905	6.014
Qwen2.5-7B	28	0.967/0.969	0.049 [0.042, 0.060]	0.65	0.76	0.00	0.430	0.454	5.178
Qwen2.5-14B	48	0.891/0.920	0.063 [0.060, 0.067]	0.22	0.12	0.00	1.013	0.540	3.221
Qwen2.5-32B	64	0.856/0.955	0.044 [0.042, 0.047]	0.52	0.41	0.08	0.519	0.436	5.615

per-model summaries (Table A13); the exhaustive 20-triple joint search (Table A14); and the adapter operating-point control (Table A15); Tables A17, A18, and A19 report the background-variant, availability-mask, and granularity controls with their source artifacts.

B Measurement-code review

The intervention code underwent an AI-assisted code review, performed separately from the experiment implementation pass, before the production runs; all findings (a device-placement crash, a winner's-curse flaw in the greedy comparison, stale-proxy confounds, a Wanda-definition error, fp16 gradient underflow in Taylor scores, tie-unaware rank correlation) were fixed and the runs repeated. The review and dispositions ship with the artifacts; this supports code correctness and does not substitute for independent reproduction.

C Reproducibility

The supplement contains one measurement script per domain (layer skipping, pruning, MoE, adapter placement) plus the quantization group-decomposition code, all raw measurement artifacts, the generation scripts that reproduce all imported detailed and audit tables together with the released source artifacts, and the AI-assisted measurement-code review with its dispositions. Every experiment fits on a single 141 GB-class GPU (including the 32B layer skipping, the 14B pruning sweep, MoE interventions, and 7B adapter training). Two production software stacks were used (PyTorch 2.10 with Transformers 4.57 and PyTorch 2.13 with Transformers 5.15); artifacts record the per-run Transformers version where logged, and the released `requirements.txt` pins a tested reproduction environment. Evaluation is deterministic on fixed sequences and seeds, so reported differences are numerically reproducible on those samples; uncertainty across data subsets and training seeds is reported where inferential claims depend on it.

Table A10: Structured pruning, all (model, background, target-layer) conditions. Base NLL: the background configuration's own evaluation NLL (nats/token). Interaction: median normalized pair residual (damage convention). Rel. R^D: median subset damage residual ($R^D(S) = D(S) - \sum_i D(\{i\})$) relative to joint damage, and the fraction of subsets above 0.25. ρ: singleton-rank Spearman vs. the dense model. Selection at k=4 (damage, nats): exact singleton, best score proxy, and random mean.

Model	Bg.	Layer	Base NLL	Interaction	Rel. R^D	Frac.>.25	ρ	Singleton	Best proxy	Random
Llama-1B	00%	L2	2.86	0.040	0.127	0.16	–	0.0203	0.0309	0.0599
Llama-1B	00%	L8	2.86	0.022	0.033	0.00	–	0.0292	0.0316	0.0659
Llama-1B	00%	L14	2.86	0.027	0.078	0.12	–	0.0457	0.0571	0.0765
Llama-1B	25%	L2	3.85	0.047	0.135	0.19	0.71	0.0395	0.0612	0.1672
Llama-1B	25%	L8	3.87	0.023	0.053	0.00	0.27	0.0536	0.0756	0.0851
Llama-1B	25%	L14	3.86	0.036	0.119	0.19	0.40	0.0436	0.0642	0.1157
Llama-1B	50%	L2	6.54	0.058	0.053	0.02	0.70	0.0767	0.0967	0.3669
Llama-1B	50%	L8	6.49	0.049	0.091	0.06	0.42	0.0585	0.1631	0.2927
Llama-1B	50%	L14	6.78	0.120	0.424	0.66	0.03	-0.0705	-0.0155	0.0962
Qwen-1.5B	00%	L3	2.66	0.057	0.072	0.01	–	0.0195	0.0234	0.0346
Qwen-1.5B	00%	L14	2.66	0.042	0.082	0.02	–	0.0044	0.0105	0.0128
Qwen-1.5B	00%	L24	2.66	0.025	0.086	0.03	–	0.0215	0.0270	0.0412
Qwen-1.5B	25%	L3	3.69	0.054	0.087	0.01	0.25	0.0337	0.0407	0.0672
Qwen-1.5B	25%	L14	3.71	0.054	0.084	0.03	-0.28	0.0018	0.0282	0.0315
Qwen-1.5B	25%	L24	3.67	0.025	0.082	0.05	0.61	0.0539	0.0480	0.0738
Qwen-1.5B	50%	L3	5.61	0.119	0.165	0.20	-0.20	0.0673	0.1029	0.1854
Qwen-1.5B	50%	L14	5.82	0.110	0.422	0.65	-0.32	-0.0443	0.0349	0.0300
Qwen-1.5B	50%	L24	5.84	0.027	0.108	0.14	0.02	0.0180	0.0347	0.0846
Qwen-14B	00%	L6	2.11	0.074	0.084	0.07	–	0.0075	0.0076	0.0193
Qwen-14B	00%	L24	2.11	0.068	0.073	0.01	–	0.0078	0.0110	0.0168
Qwen-14B	00%	L42	2.11	0.029	0.056	0.00	–	0.0181	0.0242	0.0241
Qwen-14B	25%	L6	2.84	0.060	0.070	0.02	0.13	0.0020	0.0144	0.0206
Qwen-14B	25%	L24	2.86	0.106	0.173	0.36	-0.01	-0.0018	0.0049	0.0059
Qwen-14B	25%	L42	2.84	0.022	0.034	0.00	0.21	0.0131	0.0199	0.0225
Qwen-14B	50%	L6	4.40	0.090	0.084	0.10	0.23	0.0028	0.0282	0.0451
Qwen-14B	50%	L24	4.49	0.060	0.460	0.73	-0.05	-0.0478	-0.0214	-0.0058
Qwen-14B	50%	L42	4.37	0.013	0.038	0.00	-0.42	0.0162	0.0200	0.0413

Table A11: Pruning replication on held-out evaluation subsets (Llama layer 14 = late target; Qwen layer 14 = middle target) (seeds 1315/1316). Columns: background-rank Spearman at 25%/50% sparsity, median relative subset residual at 50%, and damage of pruning the four least-harmful groups (negative = the loss improves).

Model / eval seed	ρ (25%)	ρ (50%)	Rel. R (50%)	Singleton-k4
Llama-1B / 1315	0.20	-0.12	0.58	-0.0864
Llama-1B / 1316	0.33	0.33	0.45	-0.0700
Qwen-1.5B / 1315	0.11	0.19	0.22	-0.0919
Qwen-1.5B / 1316	0.22	0.19	0.11	-0.0579

D SECOND-CORPUS ROBUSTNESS

The primary causal evaluations use WikiText-2 sequences. To test whether the headline verdicts are corpus artifacts, we selected representative conditions from the four forward-only domains—Qwen2.5-14B layer skipping (the budget-scale collapse), Qwen2.5-14B pruning at the mid and late targets under 50% background sparsity, OLMoE expert selection at an early and a late layer, and the quantization decomposition on Qwen-1.5B and SmolLM2—and repeated them on the C4 validation split with the protocol otherwise fixed. The C4 evaluation uses 16×512 tokens (twice the primary size), drawn as two disjoint halves: selection statistics use the first half, held-out evaluations the second. For quantization, the GPTQ calibration is unchanged (WikiText-2, seed 42), so the identical quantized configurations are re-measured on the second corpus; this tests transfer across evaluation corpora at fixed calibration, not sensitivity to a change in calibration corpus. Adapter placement is evaluated separately on GSM8K and is not included in this replication.

Table A12: MoE expert selection, all runs. Proxy ρ: router mean probability vs. singleton disable damage. Neg.: experts whose removal improves the loss. Availability ρ at 75%/50% and sign-flip rate at 50%. Selection: damage of disabling the 8 least-useful experts by exact singleton vs. random mean.

Run	Layer	Proxy ρ	Interaction	Neg.	ρ_{75}	ρ_{50}	Flips	Sel. exact	Sel. random
OLMoE reroute	L2	0.21	0.343	17/64	0.44	0.46	0.31	-0.0096	0.0087
OLMoE reroute	L8	0.29	0.358	30/64	0.67	0.46	0.38	-0.0036	0.0060
OLMoE reroute	L14	0.58	0.080	8/64	0.90	0.78	0.16	-0.0045	0.0287
OLMoE frozen	L2	0.35	0.342	16/64	0.65	0.40	0.31	-0.0034	0.0093
OLMoE frozen	L8	0.30	0.300	19/64	0.51	0.73	0.19	0.0001	0.0072
OLMoE frozen	L14	0.53	0.059	12/64	0.92	0.89	0.09	-0.0046	0.0233
OLMoE seed 1315	L2	0.29	0.295	24/64	0.39	0.62	0.41	-0.0020	0.0160
OLMoE seed 1315	L8	0.43	0.245	6/64	0.67	0.63	0.06	-0.0084	0.0156
OLMoE seed 1315	L14	0.58	0.080	10/64	0.86	0.79	0.09	-0.0028	0.0261
OLMoE seed 1316	L2	0.22	0.494	13/64	0.44	0.24	0.28	-0.0120	0.0079
OLMoE seed 1316	L8	0.26	0.281	22/64	0.74	0.81	0.16	-0.0131	0.0160
OLMoE seed 1316	L14	0.66	0.068	8/64	0.93	0.71	0.09	-0.0064	0.0205
Qwen1.5-MoE reroute	L3	-0.09	0.543	12/60	0.49	0.54	0.27	-0.0053	0.0024
Qwen1.5-MoE reroute	L12	0.07	0.439	17/60	0.59	0.55	0.23	-0.0088	-0.0002
Qwen1.5-MoE reroute	L21	0.56	0.133	7/60	0.85	0.64	0.17	0.0001	0.0124
Qwen1.5-MoE frozen	L3	0.06	0.367	12/60	0.73	0.58	0.63	-0.0039	0.0031
Qwen1.5-MoE frozen	L12	0.04	0.478	20/60	0.66	0.77	0.50	-0.0074	0.0002
Qwen1.5-MoE frozen	L21	0.59	0.221	15/60	0.87	0.82	0.23	-0.0025	0.0081

Table A13: Adapter placement, per-model summary statistics. Singleton max/min over all layers; noise: training-seed σ on replicated layers; pair residual as fraction of singleton sum and in units of the noise floor; fraction of negative (redundant) residuals; background marginal-vs-singleton Spearman and sign-flip rate.

Model	Base NLL	All-layer U	Single max/min	Noise σ	Pair R^U/sum	R^U/σ	Frac. $R^U<0$	Bg. ρ	Flips
Qwen2.5-1.5B	1.2818	-0.0287	0.1052 / -0.0229	0.00095	0.73	86×	0.60	-0.39	0.73
Qwen2.5-7B	1.5551	0.3871	0.4205 / 0.3102	0.00064	0.49	576×	1.00	0.14	0.82

The headline statistics are compared in Table A33. The C4 evaluation preserves several headline patterns: small pairwise interaction with budget-scale ranking collapse and a large re-measured-greedy advantage (skip); the early-to-late compositionality gradient in expert interaction and availability-restricted rank stability (MoE); and positive, stable full-refinement gains whose layer-sum attribution mispredicts the joint gain (quantization). Pruning effects are substantially attenuated: the middle-layer Taylor correlation changes from -0.74 to -0.12, the selected-set damage from -0.048 to -0.001 nats, and the late-layer collapse is much shallower; these near-zero effects provide limited evidence for transferring the corresponding magnitude or improvement claims across corpora. The SmolLM2 layer-sum attribution flips sign between corpora (-0.10 on WikiText-2, $+0.36$ on C4) while the joint gain stays within 0.08–0.12 nats, strengthening the seed-level attribution instability of Section 3 to corpus-level instability. Cross-corpus singleton rankings correlate at $\rho = 0.77$ (14B layer skip), so the components themselves are moderately corpus-stable; it is their composition behavior that carries the domain verdicts.

A sample-size sweep on the tested 14B layer-skip condition is reported in Table A34: the background correlation and greedy advantage are recomputed on random sequence subsets of the joint 16-sequence pool (1,024 to 8,192 scored tokens) with the selected sets held fixed. The greedy advantage remains positive across all sampled subsets, while background-rank estimates vary more at smaller sample sizes, including draws above the operational threshold. Because this pool includes the selection half, this is a conditional resampling diagnostic, not an independent held-out validation of sample-size adequacy (the held-out comparison is Table A33).

E Coverage inventory (extended limitations)

The quantization decomposition spans three calibration seeds per model, which stabilizes the totals but reveals that per-layer attributions are seed-unstable in both magnitude and sign—a finding rather than a defect, but one that precludes per-model characterization. Pruning uses zeroing without retraining; fine-tuned pruning changes the object of measurement and would require the coherent-

Table A14: Exhaustive joint search over all 20 three-of-six late-layer adapter placements (Qwen-1.5B, layers 22–27), sorted by measured utility, with the singleton-sum prediction for each.

Layers	Utility	Singleton-sum pred.
24, 25, 27	0.1090	0.3064
23, 24, 27	0.1076	0.3076
24, 26, 27	0.1073	0.2898
25, 26, 27	0.1072	0.2861
23, 25, 27	0.1070	0.3039
24, 25, 26	0.1070	0.2911
23, 24, 25	0.1052	0.3089
23, 26, 27	0.1051	0.2873
23, 25, 26	0.1047	0.2886
23, 24, 26	0.1047	0.2923
22, 24, 27	0.1026	0.3014
22, 25, 27	0.1018	0.2976
22, 24, 25	0.1007	0.3026
22, 25, 26	0.1006	0.2823
22, 24, 26	0.1006	0.2861
22, 23, 27	0.1004	0.2989
22, 26, 27	0.0992	0.2811
22, 23, 25	0.0989	0.3001
22, 23, 24	0.0989	0.3039
22, 23, 26	0.0985	0.2836

Table A15: Adapter operating-point control (Qwen-1.5B, mean±std over 3 training seeds). The last-6 $>$ top-singleton ordering and the harm of spread/all-layer placement persist at every operating point.

Variant	Top-singleton-6	Last-6	Evenly-spaced-6	All-28
lr $\times 0.5$	0.0984±0.0017	0.1055±0.0008	-0.0057±0.0021	-0.0233±0.0011
lr $\times 2$	0.0725±0.0012	0.0856±0.0020	-0.0078±0.0022	-0.0419±0.0027
500 steps	0.0707±0.0014	0.0866±0.0023	-0.0322±0.0020	-0.1342±0.0063

training protocol we apply in the adapter domain. The pruning background construction is replicated across the activation-magnitude selector and five random masks (mean retained $\rho \approx 0.33$, per-mask range 0.02–0.59), but other structured selectors remain untested. Layer-pair sweeps use one selection subset with held-out verification of all selected sets, not full three-subset pair replication. The MoE domain covers two architectures with a frozen-assignment control isolating learned interaction from routing competition, but only three layers per model; availability masks are replicated (three per fraction) for the first architecture's early and mid layers and single elsewhere. The adapter domain uses two model scales, one adaptation task (GSM8K), and one background placement per scale; the selection-level findings were replicated at half/double learning rate and doubled steps, but the full singleton sweep uses a single operating point and task dependence of placement rankings is unmeasured.

F A MEASURABLE CORRELATE OF PAIRWISE INTERACTION

The coupling statistic C_{ij} is defined in Section 8. This appendix records the measurement protocol, exclusions, and per-run statistics. We measured C_{ij} jointly with the pairwise interaction $I_{ij} = U(\{i,j\}) - u_i - u_j$ in twelve runs across the three forward-only intervention domains, under the main protocol (same models, intervention semantics, and evaluation sequences): layer skipping at 1–14B, channel-group pruning at 1–14B, and expert masking on both MoE architectures at two depths each (2,856 pairs total). Singleton utilities, joint utilities, and Δh are all measured in the same run on the same evaluation set, so the coupling statistic and the residual are never drawn from different conditions. The quantization and adapter domains are excluded by design: there Δh_i requires a per-singleton GPTQ solve or training run, so the statistic loses its cost rationale.

Table A35 and Figure 2 summarize the result. The rank correlation is positive in eleven of twelve runs (within-run ρ up to 0.56; pooled $\rho = 0.53$), strongest in the MoE domain (pooled 0.67), intermediate for layer skipping (0.49), and weakest for pruning (0.14). The single negative run (14B pruning,

Table A16: Gate-sensitivity check: number of conditions crossing each gate when its threshold is perturbed (interaction gates at $\theta \in \{0.15, 0.20, 0.25\}$; correlation gates at $\theta \in \{0.4, 0.5, 0.6\}$). Qualitative domain verdicts are unchanged; the one threshold-sensitive statistic is the MoE availability criterion, whose crossers concentrate between $\rho = 0.4$ and 0.5—the MoE verdict in the main text accordingly rests on the (robust) interaction criterion and the frozen-assignment control rather than on this gate.

Criterion	Conditions	θ_1	θ_2	θ_3
Skip: interaction $> \theta$	6 models	0	0	0
Skip: background $\rho < \theta$	6 models	1	2	2
Prune: rel. residual $> \theta$	9 layers	4	3	3
Prune: background $\rho < \theta$	9 layers	8	8	8
MoE: interaction $> \theta$	6 layers	4	4	4
MoE: availability $\rho < \theta$	6 layers	0	3	4
Adapter: pair residual $> \theta$	2 models	2	2	2
Adapter: background $\rho < \theta$	2 models	2	2	2

Table A17: Qwen-14B layer-skip background-variant grid: singleton-rank Spearman vs. the dense model under bottom-k and random-k backgrounds crossed with three evaluation subsets.

Background	Eval subset	ρ vs. dense
bottomk	314	0.124
bottomk	1315	0.026
bottomk	1316	0.334
random0	314	0.723
random0	1315	0.723
random0	1316	0.779
random1	314	0.491
random1	1315	0.339
random1	1316	0.680

$\rho = -0.13$) coincides with a narrow coupling range $[-0.03, 0.08]$, leaving little dynamic range for reliable ranking—and that near-orthogonal regime is also where the domain is closest to additive. Coupling ranges also contract across the tested model scales (pruning at 1B spans $[-0.29, 0.61]$; at 7–14B, $[-0.05, 0.08]$); because width, depth, and pretraining differ jointly, we treat this as a descriptive pattern consistent with increasing orthogonality rather than an isolated effect of width. We treat this as supporting evidence for the shared-state account and as a candidate cheap screen, not as a validated allocation input: the correlation is a rank-level association measured at one evaluation granularity, and its failure mode (range collapse) must be checked before use.

Separating representation non-additivity from readout curvature. For the layer-skip run (Qwen2.5-1.5B-Instruct) we additionally store the final residual stream and decompose each pair's *loss* interaction $I_{ij}^{\text{total}} = \ell(h_{ij}) - \ell(h_i) - \ell(h_j) + \ell(h_0)$ —the negative of the utility interaction defined in Section 2—as $I_{ij}^{\text{total}} = I_{ij}^{\text{readout}} + [\ell(h_{ij}) - \ell(h_{ij}^{\text{add}})]$, where $h_{ij}^{\text{add}} = h_0 + \Delta h_i + \Delta h_j$ is the synthetic additive state and I^{readout} is the interaction an exactly additive representation would already exhibit through the frozen nonlinear readout. Over 100 pairs the two terms are individually far larger than their sum (median $|I^{\text{readout}}| = 0.124$ and $|\ell(h_{ij}) - \ell(h^{\text{add}})| = 0.086$, versus $|I^{\text{total}}| = 0.022$), and the cancellation is pairwise: the two terms carry opposite signs in 97% of pairs, and $|I^{\text{total}}|$ falls below the smaller term's magnitude in 80% (median signed values $+0.115$ and -0.083). In this measured layer-skip run, near-additivity of the loss thus arises from cancellation between readout curvature and genuine representation non-additivity (median normalized representation residual 0.30), not from additive representations. The representation residual also tracks the total interaction (Spearman 0.57). In the non-compositional MoE layer (OLMoE, layer 2, 100 pairs) the decomposition differs qualitatively: readout and representation terms are each of the same order as the total (median $|I^{\text{readout}}|/|I^{\text{total}}| \approx 0.9$), opposite signs occur in 53% of pairs, and the representation residual is larger (median 0.46) yet does not track the total interaction within the run ($\rho = -0.16$). The decomposition thus separates two regimes in the measured runs—cancellation-driven near-additivity in the layer-skip run versus much weaker cancellation in the MoE run—while cautioning against

Table A18: OLMoE availability-mask replication: singleton-rank Spearman vs. full availability under three random masks per fraction.

Layer	Fraction / mask	ρ
L2	avail75:m0	0.590
L2	avail75:m1	0.643
L2	avail75:m2	0.611
L2	avail50:m0	-0.055
L2	avail50:m1	0.538
L2	avail50:m2	0.481
L8	avail75:m0	0.595
L8	avail75:m1	0.588
L8	avail75:m2	0.503
L8	avail50:m0	0.502
L8	avail50:m1	0.574
L8	avail50:m2	0.538

Table A19: Pruning granularity and background-mask controls: pair interaction and subset relative residual for attention-head and whole-MLP granularities, and the range of singleton-rank Spearman under five random background masks for channel groups at the late target layer.

Model	Granularity / control	Interaction	Rel. R	Bg-mask ρ
Llama-1B	heads (L2)	0.032	0.091	–
Llama-1B	heads (L8)	0.131	0.213	–
Llama-1B	heads (L14)	0.028	0.038	–
Llama-1B	whole-MLP (all layers)	0.085	0.198	–
Llama-1B	ch. groups, 5 random bg (L14)	–	–	0.02–0.59 (mean 0.32)
Qwen-1.5B	heads (L3)	0.158	0.309	–
Qwen-1.5B	heads (L14)	0.087	0.248	–
Qwen-1.5B	heads (L24)	0.061	0.164	–
Qwen-1.5B	whole-MLP (all layers)	0.038	0.129	–
Qwen-1.5B	ch. groups, 5 random bg (L24)	–	–	0.05–0.59 (mean 0.35)

reading the coupling correlate as a single mechanism. The synthetic state need not correspond to a realizable configuration; this is a diagnostic decomposition, not a causal attribution.

Table A20: Deploying the dense-model pruning ranking in a sparse background (the transfer arm): held-out damage of the dense bottom-k set versus the background-re-measured bottom-k and random ($k = 4$; positive penalty = dense transfer worse). All sets evaluated under the same background on held-out sequences. Damage and penalty are reported in nats/token.

Model	Layer / bg	ρ	Overlap	Dense	Re-meas.	Penalty	Random
Llama-1B	L2 / bg25	0.71	0.50	0.083	0.061	0.022	0.115
Llama-1B	L2 / bg50	0.70	0.50	0.220	0.206	0.015	0.383
Llama-1B	L8 / bg25	0.27	0.50	0.071	0.051	0.021	0.104
Llama-1B	L8 / bg50	0.42	0.25	0.093	0.031	0.062	0.196
Llama-1B	L14 / bg25	0.40	0.25	0.106	0.054	0.052	0.111
Llama-1B	L14 / bg50	0.03	0.25	0.015	-0.069	0.083	0.022
Qwen-1.5B	L3 / bg25	0.25	0.50	0.058	0.044	0.014	0.074
Qwen-1.5B	L3 / bg50	-0.16	0.00	0.232	0.115	0.117	0.190
Qwen-1.5B	L14 / bg25	-0.26	0.25	0.030	0.007	0.023	0.024
Qwen-1.5B	L14 / bg50	-0.32	0.00	0.082	-0.016	0.098	0.068
Qwen-1.5B	L24 / bg25	0.66	0.50	0.074	0.066	0.008	0.069
Qwen-1.5B	L24 / bg50	-0.00	0.25	0.096	0.050	0.047	0.101
Qwen-14B	L6 / bg25	0.12	0.00	0.017	0.018	-0.001	0.019
Qwen-14B	L6 / bg50	0.23	0.25	0.036	0.049	-0.013	0.045
Qwen-14B	L24 / bg25	-0.02	0.00	0.011	0.008	0.002	0.012
Qwen-14B	L24 / bg50	-0.05	0.25	0.019	-0.016	0.035	0.017
Qwen-14B	L42 / bg25	0.28	0.25	0.016	0.015	0.000	0.017
Qwen-14B	L42 / bg50	-0.41	0.25	0.049	0.025	0.024	0.034

Table A21: GSM8K evaluation audit (full 1,319-problem test set, greedy decoding), three metrics from the released parser: *legacy strict* takes the last #### answer anywhere in the full completion (sensitive to hallucinated follow-up markers); *segment strict* requires a #### answer before the first validated follow-up-problem boundary; *robust* uses the same segment but falls back to its last number. All rules are applied identically to every configuration by the released, unit-tested parser v3; its exact frozen regex, boundary inventory, and audit protocol are provided in the supplement. After freezing v3, it agreed with blinded audit annotations on 195/200 completions (97.5%) from a disjoint stratified holdout sample; no rule was changed using that sample. Format: fraction of first segments containing a #### marker.

Model	Configuration	Legacy	Seg.-strict	Robust	Format	Unparsed
Qwen-1.5B	Base (no adapter)	0.0	0.0	49.9	0.0	0.0
Qwen-1.5B	Single best layer	27.9	43.4	46.4	89.1	0.0
Qwen-1.5B	Top-singleton-6	29.3	46.9	47.2	97.0	0.0
Qwen-1.5B	Last-6	28.5	45.5	45.9	95.2	0.0
Qwen-1.5B	All layers	35.6	46.3	46.4	97.3	0.0
Qwen-7B	Base (no adapter)	0.0	0.0	73.5	0.0	0.0
Qwen-7B	Single best layer	44.6	68.9	72.6	92.3	0.0
Qwen-7B	Top-singleton-6	63.5	72.4	72.7	96.8	0.0
Qwen-7B	Last-6	50.7	72.6	72.8	97.2	0.0
Qwen-7B	All layers	59.9	73.7	73.8	96.8	0.0

Table A22: Paired robust-EM comparisons using the audited extraction rule (parser v3), for all configuration pairs including the unadapted base. Differences are A minus B in percentage points; CIs are paired bootstrap 95% intervals obtained by resampling problem indices jointly across configurations (B=10,000, seed 0); p is an exact two-sided McNemar test on discordant pairs; Holm-adjusted p values are reported over the family of four base–adapter contrasts per scale. Among adapted configurations no pairwise difference is statistically resolved ($p \geq 0.17$, CIs spanning zero); at 1.5B three of the four base–adapter contrasts remain significant after Holm correction (adjusted $p \leq 0.044$; top-singleton-6 adjusted $p = 0.060$), and at 7B no contrast is resolved (all $p \geq 0.41$). Same artifacts as Table A21.

Scale	Configuration A	Configuration B	Diff. (pt)	95% CI	p	Holm p
Qwen-1.5B	Base (no adapter)	Single best layer	+3.5	[+0.8, +6.2]	0.01	0.04
Qwen-1.5B	Base (no adapter)	Top-singleton-6	+2.7	[+0.0, +5.4]	0.06	0.06
Qwen-1.5B	Base (no adapter)	Last-6	+3.9	[+1.2, +6.7]	0.01	0.02
Qwen-1.5B	Base (no adapter)	All layers	+3.5	[+0.6, +6.4]	0.02	0.04
Qwen-1.5B	Single best layer	Top-singleton-6	-0.8	[-3.0, +1.4]	0.51	–
Qwen-1.5B	Single best layer	Last-6	+0.5	[-1.8, +2.7]	0.74	–
Qwen-1.5B	Single best layer	All layers	+0.0	[-2.8, +2.9]	1.00	–
Qwen-1.5B	Top-singleton-6	Last-6	+1.3	[-0.5, +3.0]	0.17	–
Qwen-1.5B	Top-singleton-6	All layers	+0.8	[-2.0, +3.6]	0.60	–
Qwen-1.5B	Last-6	All layers	-0.5	[-3.3, +2.4]	0.79	–
Qwen-7B	Base (no adapter)	Single best layer	+0.8	[-1.9, +3.5]	0.59	1.00
Qwen-7B	Base (no adapter)	Top-singleton-6	+0.8	[-2.0, +3.4]	0.62	1.00
Qwen-7B	Base (no adapter)	Last-6	+0.7	[-1.9, +3.3]	0.65	1.00
Qwen-7B	Base (no adapter)	All layers	-0.3	[-3.0, +2.4]	0.87	1.00
Qwen-7B	Single best layer	Top-singleton-6	-0.1	[-2.5, +2.4]	1.00	–
Qwen-7B	Single best layer	Last-6	-0.2	[-2.6, +2.2]	0.95	–
Qwen-7B	Single best layer	All layers	-1.1	[-3.7, +1.4]	0.42	–
Qwen-7B	Top-singleton-6	Last-6	-0.1	[-2.6, +2.4]	1.00	–
Qwen-7B	Top-singleton-6	All layers	-1.1	[-3.4, +1.4]	0.41	–
Qwen-7B	Last-6	All layers	-1.0	[-3.6, +1.6]	0.48	–

Table A23: An exact-singleton selector arm for the quantization domain (Qwen-1.5B, highest-gain layer, 12 pre-specified candidate blocks, k=3): held-out refinement gain of exact-singleton top-k versus the Hessian-weighted proxy and 100 random draws (mean, range, and the exact selection's percentile within the random distribution), with the best evaluated k-set, per calibration seed. The final column gives the exact selection's singleton-sum prediction versus its realized joint gain on the selection subset.

Seed	Exact	Pctl.	Proxy	Random (n=100; mean [min, max])	Best	Sum pred / joint
42	0.0073	39%	0.0039	0.0111 [-0.009, 0.040]	0.0403	0.0201 / -0.0123
43	-0.0095	49%	-0.0100	-0.0104 [-0.034, 0.010]	0.0100	0.0288 / -0.0119
44	-0.0133	40%	-0.0043	-0.0110 [-0.036, 0.011]	0.0107	-0.0274 / -0.0117

Table A24: Compensation-free control for the quantization domain: group-wise round-to-nearest (RTN) replaces GPTQ for every weight while the rest of the W4A4 pipeline (SmoothQuant, activation quantization, calibration data, candidate space, target layer) is unchanged (Qwen-1.5B). Columns: RTN-pipeline baseline perplexity; the gain from refining every 128-column block of every layer's down-projection to 32-wide groups; number of individually harmful candidate blocks; the 12-block singleton-sum prediction versus the measured joint gain of refining exactly those 12 candidate blocks (selection sample); and the exact-top-3 percentile among five random draws (held-out; a coarse reference). The additive-prediction sign reversal and mostly harmful singletons persist without GPTQ's cross-column compensation.

Seed	Base ppl	Full gain	Neg. singles	Sum / joint	Pctl.
42	20.58	+0.072	10/12	-0.068 / +0.004	80%
43	20.47	+0.091	9/12	-0.122 / +0.003	40%
44	20.30	+0.058	6/12	+0.014 / -0.012	20%

Table A25: Token-segment decomposition of the adapter NLL improvement over the frozen base on shared GSM8K test reference tokens (400 examples, mean over three paired training seeds; identical tokens and masks for every model). This protocol differs from the packed evaluation of Table A7: examples are scored individually with the context reset per example (packing concatenates examples into 512-token windows), the example set is the first 400 test problems rather than the packed prefix, and the unit is nats/example rather than nats/token. Cells: per-example NLL improvement summed over the segment (per-token improvement in parentheses). Segments: prompt = question tokens; solution = answer body before the marker; final number = the numeric answer; format = the #### marker, separators, and EOS. This example-level scoring complements the packed-sequence NLL protocol of the main text.

Model	Config	Total	Prompt	Solution	Final no.	Format
Qwen-1.5B	Top-singleton-6	+28.6	-1.3 (-0.02)	+18.3 (+0.16)	+0.1 (+0.04)	+11.5 (+5.77)
Qwen-1.5B	Last-6	+30.0	+0.4 (+0.01)	+17.9 (+0.15)	+0.1 (+0.03)	+11.5 (+5.77)
Qwen-7B	Top-singleton-6	+99.4	+22.9 (+0.36)	+63.6 (+0.54)	+0.2 (+0.09)	+12.7 (+6.36)
Qwen-7B	Last-6	+96.2	+23.2 (+0.37)	+60.4 (+0.52)	+0.2 (+0.08)	+12.5 (+6.25)

Table A26: Layer skipping across scales. Background ρ: singleton-rank correlation after skipping the bottom-25% set; greedy advantage: static-minus-greedy held-out damage at the 25% budget (positive = greedy better).

Model	Layers	Median interaction	Background ρ	Greedy adv.
Llama-1B	16	.133	.62	$+0.300$
Qwen-1.5B	28	.051	.86	$+0.261$
SmolLM2-1.7B	24	.078	.63	$+0.088$
Qwen-7B	28	.049	.76	-0.024
Qwen-14B	48	.063	**.12**	$\mathbf{+0.473}$
Qwen-32B	64	.044	**.41**	$+0.083$

Table A27: Paired example-level NLL contrasts between the two six-layer placements on the shared 400 held-out examples (advantage in nats/example = the other placement's NLL minus the named placement's; positive = the named placement is better). Each row is one training seed: within a seed, the two placements are compared on identical examples and the CI is a paired bootstrap over examples (B=10,000, seed 0); training-seed variation is shown as the three per-seed rows, not pooled.

Model	Contrast	Seed	Diff.	Paired 95% CI
Qwen-1.5B	Last-6 advantage	1	+1.05	[+0.84, +1.26]
Qwen-1.5B	Last-6 advantage	2	+1.70	[+1.45, +1.96]
Qwen-1.5B	Last-6 advantage	3	+1.22	[+0.98, +1.45]
Qwen-7B	Top-singleton-6 advantage	1	+3.28	[+2.64, +3.90]
Qwen-7B	Top-singleton-6 advantage	2	+2.89	[+2.25, +3.53]
Qwen-7B	Top-singleton-6 advantage	3	+3.50	[+2.85, +4.14]

Table A28: Fixed-k layer skipping: the static-minus-greedy held-out damage gap (pooled over two held-out subsets, per-sequence paired bootstrap 95% CI, selected sets held fixed) and the background rank correlation after skipping the static set, at matched removal counts k on Qwen2.5-7B and 14B. Both selectors are re-derived per model and k from the same selection sequences. The 14B k=12 sets coincide with the released 25%-budget sets; values here are independent re-evaluations, and sub-0.002-nat differences in the mean gaps from Table A32 reflect fp16 evaluation on different hardware.

Model	k	Static−greedy	95% CI	Background ρ
Qwen-7B	7	-0.024	[-0.046, -0.002]	0.76
Qwen-7B	12	+0.173	[+0.131, +0.217]	0.40
Qwen-14B	7	+0.043	[+0.021, +0.067]	0.28
Qwen-14B	12	+0.474	[+0.248, +0.749]	0.12

Table A29: Downstream check of the strongest NLL selection result: Qwen2.5-14B zero-shot accuracy (acc_norm, full test sets) for the dense model and the released static-25% and greedy-25% skip sets (sets held fixed; no re-selection on task data). The final columns give the static-minus-greedy accuracy difference, its per-example paired bootstrap 95% CI (B=10,000, seed 0), and the exact two-sided McNemar p. These tests do not establish equivalence between the two selectors. Dense exceeds both skipped configurations on both tasks ($p < 10^{-32}$).

Task	Dense	Static	Greedy	Static−greedy (pp)	95% CI	p
ARC-Challenge	62.3	43.4	43.7	-0.26	[-2.82, +2.39]	0.9
HellaSwag	84.3	70.3	69.6	+0.62	[-0.04, +1.26]	0.069

Table A30: Selection-time measurement cost per domain (analytic counts from the protocol configuration). One evaluation is a forward pass over the 8×512-token evaluation set; greedy counts are $\sum_{s<k}(L-s)$ re-measurements for a k-of-L budget. Deployment cost is identical across selectors at matched budgets; these are search-time costs.

Domain	Exact singleton	Proxy	Sequential greedy	Random
Layer skip (28 layers)	28 evals	1 instrumented pass	175 evals (25% budget)	5 evals
Layer skip (48 layers, 14B)	48 evals	1 instrumented pass	510 evals	5 evals
Pruning (16 groups, per bg)	16 evals	1 pass (act./Wanda); 8 scaled bwd (Taylor)	not run	$\leq$5 evals
Quantization (12 blocks)	12 requant.+evals	free (from cached factors)	not run	requant.+eval per draw
MoE (64 experts, per layer)	64 evals	free (router capture)	not run	1 eval per draw
Adapter (28 layers)	28 training runs	—	not run	1 training run per draw
Pruning linear mask surrogate (per bg)	300+100 subset evals (150+50 at 14B)	one fitted model	not run	1 selected-set eval

Table A31: Configuration-trained linear surrogate versus exact background-conditioned singletons (pruning, k=4). A linear mask surrogate (COAR-inspired diagnostic, not the official implementation) is fit on the released random-subset measurements (small-model conditions: 300 fitting / 100 held-out subsets; the 14B condition: 150 / 50; permutation seed 0); its bottom-k selection is then evaluated under the same background on held-out sequences. R^2: held-out subset fit.

Condition	R^2	$\rho(w, \text{singl.})$	Surrogate $k4$	Singleton $k4$
Llama-1B, L14/50%	0.78	0.73	0.002	-0.069
Qwen-1.5B, L14/50%	0.93	0.95	0.013	-0.016
Qwen-14B, L24/50%	0.96	0.97	-0.021	-0.016

Table A32: Uncertainty reinforcement for the headline effects. Skip and pruning entries are per-sequence paired bootstrap 95% CIs with the selected sets held fixed (the deployment view; re-selecting per subset is reported in the replication tables). The adapter entry is the range over three paired training seeds, not a confidence interval. All damages and penalties are in nats per token.

Effect	Mean (nats)	95% CI / range	Basis
14B skip: static−greedy (held-out)	0.473	[0.253, 0.742]	n=16, 2 subsets
Llama-1B L14/50% bottom-4 (subset 314)	-0.070	[-0.103, -0.039]	fixed set, per-seq boot.
Llama-1B L14/50% bottom-4 (subset 1315)	-0.069	[-0.108, -0.029]	fixed set, per-seq boot.
Llama-1B L14/50% bottom-4 (subset 1316)	-0.066	[-0.100, -0.032]	fixed set, per-seq boot.
Qwen-1.5B L14/50% bottom-4 (subset 314)	-0.044	[-0.079, -0.010]	fixed set, per-seq boot.
Qwen-1.5B L14/50% bottom-4 (subset 1315)	-0.016	[-0.050, 0.030]	fixed set, per-seq boot.
Qwen-1.5B L14/50% bottom-4 (subset 1316)	0.008	[-0.054, 0.062]	fixed set, per-seq boot.
Qwen-14B L24/50% bottom-4 (subset 314)	-0.048	[-0.066, -0.028]	fixed set, per-seq boot.
Qwen-14B L24/50% bottom-4 (subset 1315)	-0.016	[-0.038, 0.007]	fixed set, per-seq boot.
Qwen-14B L24/50% bottom-4 (subset 1316)	-0.025	[-0.038, -0.011]	fixed set, per-seq boot.
Adapter triple gap (best−sum)	0.0029	[0.0010, 0.0046]	3 paired seeds (range)

Table A33: Second-corpus robustness: pre-selected headline statistics on the primary WikiText-2 evaluation versus the disjoint C4 replication (16×512 tokens, selection/evaluation split). ✓ = the qualitative pattern is preserved; verdicts for substantially attenuated effects are stated explicitly. Exact values are not expected to match.

Condition	Statistic	WT2 (primary)	C4	Verdict
Skip 14B	median pair interaction	0.063	0.038	small ✓
Skip 14B	background ρ	0.12	0.23	< 0.5 ✓
Skip 14B	greedy advantage (held-out)	+0.47	+0.40	✓
Skip 14B	cross-corpus singleton ρ	—	0.77	—
Prune 14B mid	background ρ (50%)	−0.05	0.38	< 0.5 ✓
Prune 14B mid	Taylor∼utility ρ	−0.74	−0.12	weak neg.; attenuated
Prune 14B mid	least-harmful-k4 (nats)	−0.048	−0.001	near zero (no CI)
Prune 14B mid	median rel. residual	0.46	0.36	> 0.25 ✓
Prune 14B late	background ρ (50%)	−0.42	0.47	< 0.5; much weaker
MoE OLMoE	interaction early / late	0.34 / 0.08	0.23 / 0.03	gradient ✓
MoE OLMoE	avail-50% ρ early / late	0.46 / 0.78	0.36 / 0.70	gradient ✓
Quant Qwen-1.5B	layer sum / joint	0.30 / 0.11	0.19 / 0.14	over-predicts ✓
Quant SmolLM2	layer sum / joint	−0.10 / 0.08	0.36 / 0.12	sign-unstable ✓

Table A34: Sample-size stability of the Qwen2.5-14B layer-skip headline statistics on C4: background rank correlation and greedy advantage recomputed on random sequence subsets of the joint 16-sequence pool (selection and evaluation halves combined; selected sets held fixed; 15 draws per size; mean [min, max]). The greedy advantage remains positive across all sampled subsets; background-rank estimates vary more at smaller sample sizes, including draws above the operational threshold.

Scored tokens	Background ρ	Greedy advantage (nats)
1,024	0.29 [-0.01, 0.58]	+0.48 [0.31, 0.66]
2,048	0.29 [0.05, 0.53]	+0.44 [0.37, 0.54]
4,096	0.19 [-0.05, 0.44]	+0.44 [0.38, 0.49]
8,192	0.23 [0.23, 0.23]	+0.45 [0.45, 0.45]

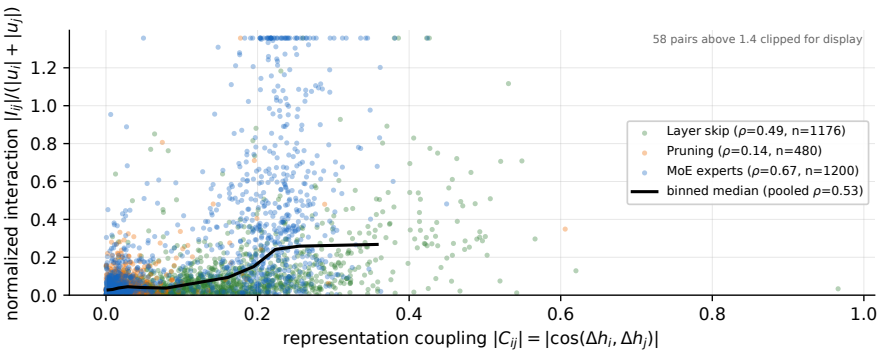

Figure 2: **Residual-stream coupling as a measurable correlate of pairwise interaction.** For each singleton pair (i, j), the x-axis shows $|C_{ij}|$, the absolute cosine similarity between singleton-induced changes to the final residual stream; the y-axis shows normalized interaction magnitude. Higher coupling tends to accompany larger interaction across the forward-only domains (11 of 12 runs positive)—supporting evidence for the shared-state hypothesis, not a validated allocation predictor.

Table A35: Rank correlation between the coupling statistic and the measured pairwise interaction, per run and pooled. $\rho(|C|, |I|)$: Spearman of $|C_{ij}|$ against $|I_{ij}|/(|u_i| + |u_j| + \epsilon)$; $\rho(C, I)$: signed variant, where I uses the damage convention of the underlying removal-domain measurements ($I^D = -I^U$).

Domain	Model / layer	n	$\rho(\|C\|, \|I\|)$	$\rho(C, I)$
MoE experts	OLMoE-1B-7B, layer 14	300	0.28	0.06
MoE experts	OLMoE-1B-7B, layer 2	300	0.26	0.22
MoE experts	Qwen1.5-MoE-A2.7B, layer 2	300	0.23	-0.13
MoE experts	Qwen1.5-MoE-A2.7B, layer 20	300	0.36	0.02
Pruning	Llama-3.2-1B, layer 14	120	0.34	0.31
Pruning	Qwen2.5-14B, layer 24	120	-0.13	0.19
Pruning	Qwen2.5-1.5B, layer 14	120	0.21	0.46
Pruning	Qwen2.5-7B, layer 14	120	0.07	0.32
Layer skip	Llama-3.2-1B	120	0.56	0.47
Layer skip	Qwen2.5-14B	300	0.28	-0.08
Layer skip	Qwen2.5-1.5B	378	0.51	0.50
Layer skip	Qwen2.5-7B	378	0.50	0.41
Pooled	all domains	2856	**0.53**	0.22