SHARED OR INDEPENDENT? A CONTROLLED STUDY OF TRANSFORM SEARCH IN LLM QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-bit quantization reduces model storage but introduces approximation error. Rotations can reduce this error by changing the coordinates in which weights are quantized. In attention layers, query, key, and value projections can share an input rotation or use independent rotations. Although independent rotations offer greater flexibility, we show that temporarily sharing Q and K during fitting improves the final independent models on OPT and Llama under matched controls. We study this effect in QKV-only ternary post-training quantization. Our construction copies shared rotations into independent variables without changing the quantized model at the transition, then continues fitting independently. This enables a comparison with independent fitting from the outset while holding the final transform family fixed. We evaluate OPT-1.3B, Llama-3.2-3B, Qwen3-4B-Base, and Ouro-1.4B with matched initializations and restart rules. Temporary sharing lowers negative log-likelihood (NLL) on C4 and WikiText-2 in 19 of 24 model–initialization–corpus comparisons, with 18 improvements remaining significant after correction for multiple comparisons. Additional fitting-budget controls retain the gains on OPT and Llama. Across HellaSwag, ARC-Challenge, and PIQA, normalized accuracy improves in 27 of 36 comparisons, including ten significant improvements after correction. On Llama-3.2-3B, HellaSwag accuracy increases by 1.20 and 0.89 percentage points at the two random initializations. Benefits depend on initialization, sharing duration, and evaluation setting. Lower reconstruction error can still accompany higher NLL even after accounting for model inputs. Interventions on the fitted models further identify the contributions of Q, K, and layer depth to these performance differences. These results establish temporary sharing as a useful fitting strategy in the studied setting and show why reconstruction error, language-model loss, and downstream accuracy must be distinguished when evaluating quantized models.

1 INTRODUCTION

Query, key, and value projections can share a quantization transform or use separate transforms. Independent transforms offer more choices, but the resulting model also depends on how those transforms are fitted. We study whether sharing Q and K temporarily can improve fitting even when the final model uses independent transforms.

Our setting is QKV-only ternary post-training quantization. An initial five-family comparison motivates the study: independent fitting reduces reconstruction error in every family, yet worsens negative log-likelihood (NLL) on Llama and Mistral while improving it on OPT. Separately fitted models alone cannot distinguish the effect of sharing from that of the fitting process.

Prior work studies sensitivity-aware quantization, learned rotations, and local preoptimization (Nagel et al., 2020; Frantar et al., 2023; Liu et al., 2025; van Breugel et al., 2026). We isolate temporary sharing while fixing the quantizer, fitting objective, and final transform family.

We compare two fitting routes. S begins with one rotation for Q and K and a separate rotation for V. It copies the shared rotation into separate Q and K variables, then continues fitting independently. Copying alone preserves the quantized model. C fits independently throughout, with matched starts

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.

and restarts. Both routes finish with independent transforms; the comparison tests whether sharing during the first phase helps find better quantized weights, not whether a shared final model has greater capacity (Table 1).

Matched restart and fitting-budget controls retain gains on OPT and Llama. Duration sweeps reveal an Ouro reversal with brief sharing and Qwen rankings that differ across evaluation settings. We then test input-aware reconstruction measures and locate Q/K and depth effects through fixed-weight replacements, without attributing them to particular optimization steps.

Our contributions are threefold:

- A model-preserving split that enables a controlled test of temporary sharing.
- A four-model evaluation of its benefits and ranking reversals under matched starts, restarts, fitting budgets, and varying sharing durations, including downstream tasks.
- Input-aware reconstruction measurements and fixed-weight interventions that locate Q/K and depth-dependent effects when reconstruction error and model quality disagree.

2 Related Work

Transforms and sharing. SmoothQuant and AWQ use channel scaling to improve quantization (Xiao et al., 2023; Lin et al., 2024). OmniQuant learns equivalent transforms and clipping through block reconstruction (Shao et al., 2024); QuIP couples orthogonal preprocessing with adaptive rounding (Chee et al., 2023). FlatQuant shares a Kronecker-factored affine input transform across QKV (Sun et al., 2025). FPTQuant couples pre-RoPE Q/K transforms and separately pairs per-head V transforms with inverse output transforms (van Breugel et al., 2026). Its Q/K transforms operate in head space and need not be identical; ours equate input-channel rotations. Q-Palette jointly selects fusible groups, including QKV subsets, and quantizers under resource constraints (Lee & Song, 2025). We fix the projector and transform family to study sharing during fitting.

Rotation search and initialization. SpinQuant optimizes rotations using task loss and reports variation across random rotations (Liu et al., 2025). FPTQuant improves training stability through local norm preoptimization (van Breugel et al., 2026). Our controls test temporary sharing while keeping the reconstruction objective fixed and the final transforms independent.

Reconstruction and sensitivity. AdaRound, BRECQ, and GPTQ use task-loss approximations or activation-dependent reconstruction to account for differences in error sensitivity (Nagel et al., 2020; Li et al., 2021; Frantar et al., 2023). GuidedQuant incorporates model-loss gradients and within-output-channel weight interactions (Kim et al., 2025). CAT analyzes concentration and weight–activation alignment through linear-layer signal-to-quantization-noise ratio (Federici et al., 2026). QEP uses the respective full-precision and quantized upstream activations to correct accumulated error (Arai & Ichikawa, 2025). We keep the reconstruction objective fixed to isolate fitting effects, then examine the fitted models through input-aware measurements and weight replacements.

Ternary quantization. TWLA combines asymmetric ternary calibration, Kronecker-structured orthogonal transforms, and activation bit allocation with adjacent-layer interactions (Zhao et al., 2026). Our symmetric row-scaled ternary projector applies only to QKV weights; activations and the KV cache retain backbone precision.

3 Quantization Setup and Search Strategies

3.1 Quantization and evaluation

At layer l, let $W_{l,a} \in \mathbb{R}^{m_{l,a} \times d_l}$ denote the nonzero pretrained weight matrix for branch $a \in \{q, k, v\}$. We count each physical layer once, even if a recurrent model reuses it. Output dimensions may differ, as in grouped-query attention. For each row, the deterministic ternary quantizer Q_3 chooses the nonnegative scale and ternary code that minimize squared reconstruction error (Appendix A.1).

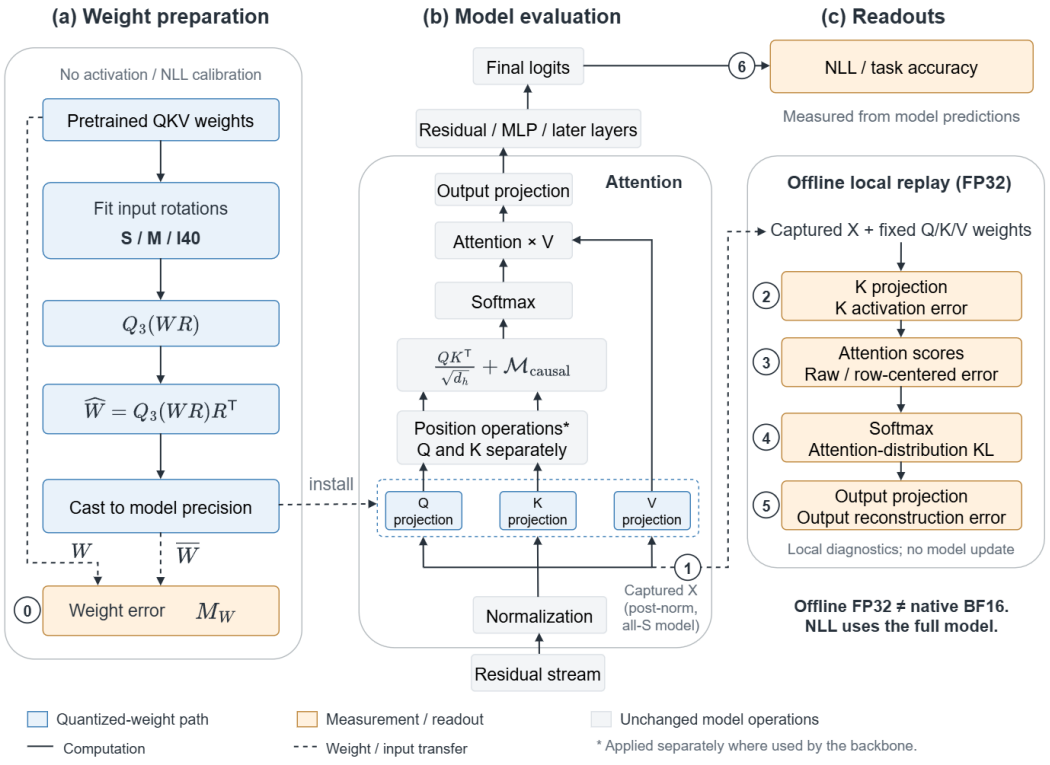

Figure 1: **Quantization and measurement locations.** (a) Ternary quantization and inverse rotation produce effective weights, converted to model precision. Readout 0 measures weight error. (b) Readout 6 evaluates predictions; d_h is head dimension and $\mathcal{M}_{\text{causal}}$ the causal mask. Position operations apply separately to Q and K where required. (c) Earlier S/M diagnostics reuse inputs captured from S at (1) to compare projection and attention outputs offline in FP32 at (2–5). These measurements are separate from the main BF16 S/C weight-replacement experiments. Dashed arrows show the transfer of weights or inputs, not optimization feedback.

For an orthogonal transform $R_{l,a} \in O(d_l)$, the effective quantized weight is

$$\widehat{W}_{l,a} = Q_3(W_{l,a}R_{l,a})R_{l,a}^\top, \qquad R_{l,a}^\top R_{l,a} = I. \tag{1}$$

Without Q_3, the rotation and its transpose cancel; with quantization, the weight depends on the rotation. Evaluation installs $\overline{W}_{l,a} = \text{cast}_{\text{backbone}}(\widehat{W}_{l,a})$. Other weights, activations, and the KV cache retain backbone precision: BF16 in the main study, but FP32 in the separately reported earlier Ouro partition experiment.

We measure reconstruction after this cast:

$$M_W = \frac{1}{3L}\sum_{l=1}^{L}\sum_{a\in\{q,k,v\}}\frac{\|\overline{W}_{l,a} - W_{l,a}\|_F^2}{\|W_{l,a}\|_F^2}. \tag{2}$$

Each of the L physical layers contributes three equally weighted relative errors. NLL and task accuracy are evaluated on the same installed weights (Figure 1).

3.2 Model-preserving splitting

Common uses one rotation for QKV; QK|V, QV|K, and Q|KV use one rotation for the indicated pair and another for the remaining branch. Independent assigns a separate rotation to each branch.

A shared QK solution becomes independent by copying its rotation to Q and K and retaining V. With the same deterministic quantizer, weight construction, and precision conversion, every installed

weight stays unchanged. Any model with shared transforms can therefore also be represented using independent transforms. The best possible independent reconstruction error cannot exceed the best shared error, but separate finite runs need not follow this ordering (Appendices A.2 and A.3). One can then continue from that same model either with or without sharing (Section 4.2). Any later weight repair or optimization is a separate operation.

3.3 Search routes and budget controls

Fitting alternates ternary quantization and orthogonal Procrustes updates, which fit the rotation to reduce weight error. It uses fixed pretrained weights, without activations or model-loss gradients. The stabilized solver stores states in FP64, corrects proposed FP32 updates, and evaluates their acceptance in FP64 (Appendices A.5 and F.2).

Table 1 summarizes the five routes. S and C match starts, phase boundaries, and restarts; restarting resets stopping state without drawing new rotations. A duration sweep retains 44 allocated rounds and varies the prefix $t \in \{4, 12, 32\}$: $S(t)$ shares QK for the first t rounds; $C(t)$ remains independent with the same phase boundary.

Table 1: Search routes. Rounds are allocated maxima; early stopping is recorded separately. All routes end with independent Q, K, and V transforms.

Route	Schedule and comparison role
S	QK|V for 12 rounds, then independent for 32: temporary sharing.
C	Independent for 12 rounds, restart, then 32: matches S's start and restart.
I44	Independent for 44 rounds without a restart: matches branch-round allocation.
C-CM	Same prefix and restart as C; continuation selected to target S's solver cost.
M	Two independent starts (20 and 19 rounds); branchwise reconstruction-error selection.

The cost control C-CM (I12R-CM-EXT) selects a saved model from the same second-phase run as C; C always uses its 32-round snapshot. M selects by reconstruction error, not NLL or task accuracy. Earlier I40 and I-CM controls match allocated group updates and target recorded cost, respectively. Appendix A.6 gives all schedules; Appendices B and F.5 specify the two cost controls.

3.4 Intervening on frozen endpoints

An endpoint is the model saved after fitting. We replace Q, K, or both in S with C's weights at every physical layer, without refitting (Table 2). V is bitwise equal between S and C in all twelve configurations.

Table 2: **Frozen-weight interventions.** Replace the indicated branches in S with C's weights at every physical layer, without refitting. Each effect is the configuration's NLL minus S's on identical prediction tokens. V and all other parameters stay fixed.

Configuration	Q weights	K weights	NLL effect relative to S
Reference S	Q_S	K_S	0
Replace Q	Q_C	K_S	a_Q
Replace K	Q_S	K_C	a_K
Replace Q and K (C)	Q_C	K_C	Δ

Relative to S, a_Q and a_K measure individual NLL changes; Δ measures the joint change. The interaction H is the total minus their sum. Positive Δ favors S. Exact definitions and depth-replacement identities are in Appendix A.7.

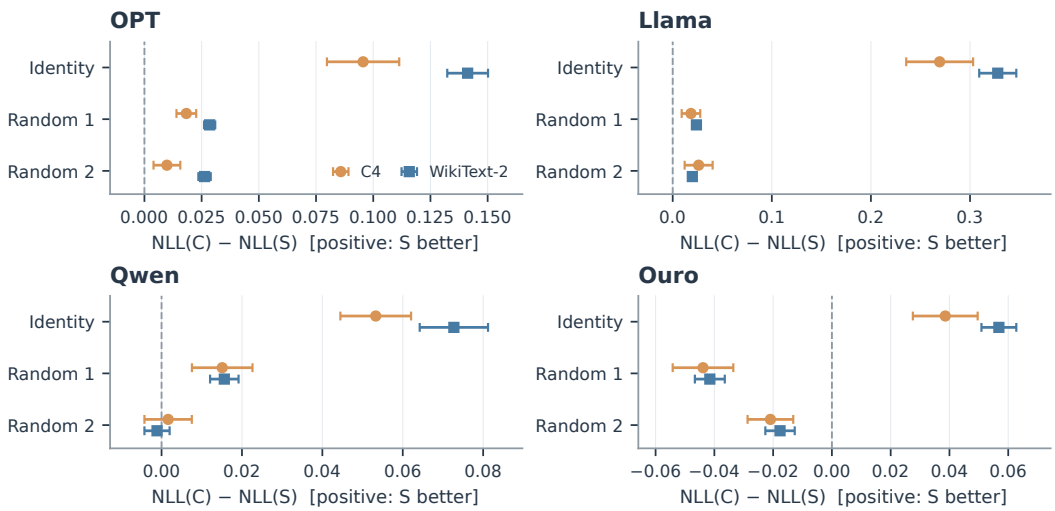

Figure 2: **Temporary sharing has different effects across models and starts.** NLL(C) minus NLL(S); positive values favor S. Both corpora evaluate the same fitted models at all three starts. Intervals use paired-bootstrap standard errors and Bonferroni correction, measuring evaluation uncertainty rather than fitting-seed variation. Horizontal scales differ across panels.

4 Experiments

4.1 Experimental Setup

The main comparison uses OPT-1.3B (Zhang et al., 2022), Llama-3.2-3B (Meta, 2024), Qwen3-4B-Base (Yang et al., 2025), and Ouro-1.4B (Zhu et al., 2025), with identity and two fixed random orthogonal initializations each. All physical QKV layers are quantized: 24, 28, 36, and 24, respectively. Ouro reuses its 24 layers for four recurrent steps; interventions change each physical weight once and NLL uses final-step logits. Effective weights are installed in BF16, with FP32 cross-entropy. Non-QKV weights, activations, and the KV cache are not quantized.

C is the primary control (Table 1); M also changes the starting set and branchwise selection. Initializations 0/1/2 denote identity and two fixed random starts, not evaluation resamples.

We evaluate NLL on a fixed 128-document C4 set and the full WikiText-2 test stream, split into non-overlapping 512-token chunks (Raffel et al., 2020; Merity et al., 2017). Each pair of models predicts the same target tokens. For the 24 primary NLL comparisons, we construct normal intervals from paired-bootstrap standard errors and apply Bonferroni correction; WikiText resampling uses groups of eight adjacent chunks. Downstream evaluation covers full HellaSwag, ARC-Challenge, and PIQA, with a separate zero-shot MMLU evaluation at initializations 0 and 1 (Zellers et al., 2019; Clark et al., 2018; Bisk et al., 2020; Hendrycks et al., 2021). Appendix B details the data, scoring, precision, and multiple-comparison corrections. Appendix C reports every candidate and task result.

4.2 Temporary Sharing under Matched Search Controls

In the initial five-family study, independent fitting lowers reconstruction error in every family but changes C4 NLL relative to QK|V by $+0.1107$ on Llama, $+0.0646$ on Mistral, and -0.3691 on OPT. Continuing independently from a shared model improves NLL on OPT and Llama, but these comparisons allocate different fitting budgets (Appendices F.1 and F.1).

Matching the restart allows a more direct test. S has lower NLL than C in 19 of 24 model–initialization–corpus comparisons (Figure 2 and Table 3). After multiple-comparison correction, eighteen significantly favor S, four favor C, and two are unresolved. All six OPT and all six Llama comparisons favor S with adjusted intervals excluding zero. The gain is therefore not explained by the second-phase restart alone. Ouro instead favors S at identity but C at both random initializations, on both corpora. Qwen's differences become small at initialization 2.

Table 3: NLL (↓), grouped by initialization. C matches S's start and restart; C-CM targets S's recorded solver cost. M uses two starts, the first matching the listed initialization. Bold marks the lowest values per row and corpus, including ties at four decimals. †: six Qwen identity layers miss the 5% cost window. S–C adjusted intervals are in Figure 2; C-CM marginal intervals, FP, and I44 are in Appendix C.

Model	Init.	WikiText-2				C4			
		S	C	C-CM	M	S	C	C-CM	M
Identity initialization									
OPT	–	3.3059	3.4471	3.4462	**3.2563**	3.0654	3.1610	3.1609	**3.0166**
Llama	–	2.7081	3.0361	3.0366	**2.4922**	2.7907	3.0600	3.0597	**2.6157**
Qwen	–	2.5541	2.6268	2.6295†	**2.4632**	2.8321	2.8854	2.8858†	**2.7631**
Ouro	–	2.8883	2.9451	2.9453	**2.8447**	2.9839	3.0224	3.0232	**2.9366**
Random orthogonal initialization									
OPT	R1	**3.1288**	3.1573	3.1574	3.1618	**2.9217**	2.9400	2.9401	2.9385
	R2	**3.1357**	3.1620	3.1623	3.1644	**2.9322**	2.9421	2.9422	2.9447
Llama	R1	**2.4660**	2.4899	2.4897	2.4925	**2.6093**	2.6277	2.6275	2.6319
	R2	**2.4592**	2.4790	2.4789	2.4810	**2.5892**	2.6153	2.6148	2.6241
Qwen	R1	**2.4558**	2.4714	2.4725	2.4693	**2.7565**	2.7716	2.7720	2.7673
	R2	2.4569	**2.4558**	2.4576	2.4713	2.7533	2.7549	2.7560	**2.7528**
Ouro	R1	2.9185	2.8769	**2.8765**	2.9040	3.0231	**2.9792**	**2.9792**	3.0115
	R2	2.8646	2.8469	2.8472	**2.8091**	2.9675	2.9466	2.9467	**2.9245**

The C-CM control matches the restart and targets S's recorded solver cost. Cost is within 5% of S's in 330 of 336 layer–initialization units; six Qwen identity layers miss this window. Aggregated across three starts, model-level cost ratios are 0.989–1.001. These measure recorded cost up to the selected solver endpoints, not end-to-end wall time (Appendix C).

S retains lower NLL than C-CM at every start and corpus for OPT and Llama, with all twelve marginal intervals excluding zero (Table 3). Ouro still favors C-CM at random starts; Qwen's Random 2 differences remain unresolved. These secondary intervals do not carry a simultaneous significance claim. The cost control preserves both the OPT/Llama benefit and the Ouro exceptions.

4.3 How Long Should Sharing Be Retained?

Using two additional random starts per model, we fit $S(t)/C(t)$ at $t = 4, 12, 32$ with matched starts and restarts and 44 allocated rounds. Evaluation uses 1,024 C4 documents (up to 2,048 tokens) and full WikiText-2 (512-token chunks). Within-setting targets match; cross-setting differences combine corpus, context-length, and segmentation effects (Appendix D).

OPT and Llama favor S at all six start–duration settings on both corpora (Figure 3). At Ouro's second start, four shared rounds worsen NLL on both corpora, but 12 and 32 improve it. Qwen's first start improves C4 but worsens WikiText at every duration; its second favors S throughout, with a C4 gain of only 0.00018 at $t = 4$. Rankings can therefore reverse with duration or differ across evaluation settings for the same weights.

4.4 Downstream Task Performance

S has higher normalized accuracy than C in 27 of 36 HellaSwag, ARC-Challenge, and PIQA comparisons; ten survive Holm correction. Table 4 reports HellaSwag; Appendix C.2 gives all candidates, both accuracy measures, and paired tests.

On Ouro at Random 1, S has worse C4 NLL than C (3.0231 versus 2.9792) but better HellaSwag accuracy (63.50% versus 62.72%); both differences survive their respective corrections. Random 2 has the same opposing rankings, although its accuracy difference is unresolved. Thus NLL alone misses a task-level benefit in the first case (Tables 3 and 4).

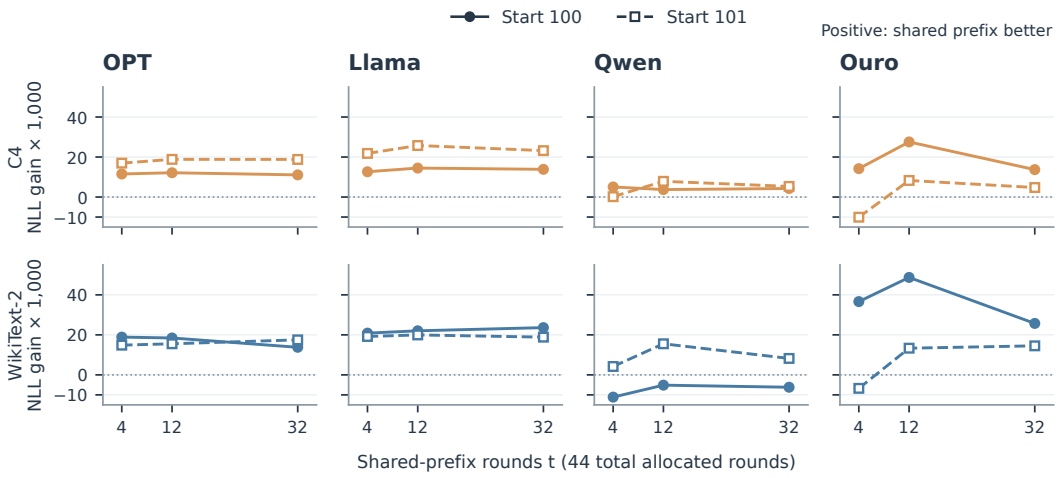

Figure 3: **Sharing duration and evaluation setting can change the ranking.** C-minus-S NLL, multiplied by 10^3, at three sharing durations. Starts 100 and 101 are additional random starts, distinct from the main comparison. Each pair matches the initial rotation and restart, with 44 allocated rounds. Positive values favor sharing. Lines connect measured durations, not optimization steps; points are descriptive. The same fitted models are evaluated on C4 documents (up to 2,048 tokens) and WikiText chunks (512 tokens).

Table 4: HellaSwag normalized accuracy (%, ↑) on all 10,042 examples. Initializations follow Table 3; repeated FP entries are the same unquantized reference. Bold marks the best of S/C/M. A star marks a Holm-significant S–C difference in the 36-test primary family. All tasks and controls are in Appendix C.2.

Model	Init.	FP	S	C	M
Identity initialization					
OPT	–	53.70	47.51*	46.30	**48.34**
Llama	–	73.64	62.16*	52.81	**69.89**
Qwen	–	73.75	66.02*	63.88	**68.96**
Ouro	–	71.71	61.80*	60.16	**64.12**
Random orthogonal initialization					
OPT	R1	53.70	**49.84**	49.35	49.21
	R2		**49.27**	49.13	48.99
Llama	R1	73.64	**70.20***	68.99	68.38
	R2		**69.80***	68.91	68.51
Qwen	R1	73.75	**68.70**	68.23	68.47
	R2		**69.31***	68.48	68.88
Ouro	R1	71.71	**63.50***	62.72	62.43
	R2		62.76	62.63	**63.58**

Zero-shot MMLU supplies a further test. Among 16 S–C and S–M comparisons over four models and two initializations, six significantly favor S and three favor the comparator after Holm correction; seven are unresolved. M can outperform S even when the restart-matched comparison favors S. The FP OPT reference is near chance, limiting the interpretation of small differences on that model. Appendix C.5 reports all 28 configurations and paired tests.

4.5 Can Input-Aware Error Resolve the Ranking Disagreement?

We test two input-aware alternatives to weight error: weighting channels by mean squared input and measuring full projection-output reconstruction. Calibration data supply these measurements; separate confirmation documents supply NLL, without model selection or refitting. Across 24 S–

Table 5: A restart-matched ranking reversal on Ouro. Each local error cell reports S / C (↓). All three calibration errors favor S, but confirmation NLL favors C at both random initializations. The NLL differences are S minus C, with adjusted intervals in Appendix C.3.

Init.	Weight error	Input-diagonal	Output error	NLL(S)−NLL(C)
R1	0.05828 / 0.06532	0.05730 / 0.06382	0.03662 / 0.03982	+0.02481
R2	0.05828 / 0.06531	0.05727 / 0.06378	0.03649 / 0.03966	+0.02098

C/S–M pairs, weight error agrees with NLL in 14 cases, diagonal weighting in 16, and output reconstruction in 17. Reused models prevent interpreting these counts alone as a statistically better selection rule (Appendix C.3).

On Ouro, all three reconstruction measures favor S over C at random starts 1 and 2, yet confirmation NLL increases by 0.02481 and 0.02098, respectively; both intervals exclude zero (Table 5). These disagreements occur under matched starts, not M's different starting set.

4.6 Effects of Q/K Weights and Layer Depth

On a separate 64-document confirmation set, we evaluate all Q/K/V replacement combinations and depth replacements in early-to-late and reverse order, without refitting. Each branch replacement affects every physical layer. V is bitwise identical in all twelve S/C configurations, leaving a_Q, a_K, and H (Table 2). NLL uses native BF16 forward passes and FP32 cross-entropy, not the offline FP32 diagnostics in Figure 1.

Simultaneous intervals for 36 prespecified contrasts significantly favor S in eight S/C comparisons and C in two; two remain unresolved. In all twelve configurations, $a_Q + a_K$ has the total's point-estimate sign: adding H never reverses it. Only three interaction intervals exclude zero (Figure 4; Appendix C.6).

Llama's largest difference is primarily associated with K: at identity, $a_K = 0.27814$, $a_Q = 0.04031$, and $H = -0.01987$, yielding $\Delta = 0.29859$. Ouro's initialization-1 reversal has $a_Q = -0.03343$, $a_K = 0.00108$, and $H = -0.00019$, yielding $\Delta = -0.03254$; its Q-only effect survives separate branch-family correction. Initialization 2 also reverses, but neither branch term is resolved under that correction.

Depth effects can oppose each other. At Ouro identity, replacing the first quarter with C weights improves NLL along both paths (-0.02013, -0.03148), whereas replacing the third worsens it ($+0.04308$, $+0.04381$); all four adjusted intervals exclude zero. Layers 6–11 favor C along both paths at the random starts.

The two replacement orders clarify what these depth effects measure. Each change is evaluated after the preceding groups have already been replaced, so the surrounding weights differ between the two paths. Both paths start at S and end at C, making their totals identical, but a group's contribution can depend on that background. These are conditional replacement effects, not context-free scores for ranking layers (Appendix A.7).

A separate S–M diagnostic tests whether K-error magnitude alone determines its effect on NLL. We change K's error direction, matching its Frobenius norm to S's within the checked numerical tolerance and keeping Q at S. Averaged over eight separate single-layer interventions, M's direction lowers NLL by 0.00571 on Llama but raises it by 0.00169 on Ouro at identity initialization; both adjusted intervals exclude zero. Changing Q to M reduces Llama's benefit to 0.00194, with a significant change in the direction effect. Norm-matched K errors can therefore have different consequences depending on their direction and the accompanying Q weights (Appendix C.4).

5 Discussion and Limitations

S and C finish with the same independent transform family, and copying leaves the installed weights unchanged. The OPT/Llama gains reflect the fitting procedure, without a Q/K constraint at inference (Sections 3.2 and 4.2).

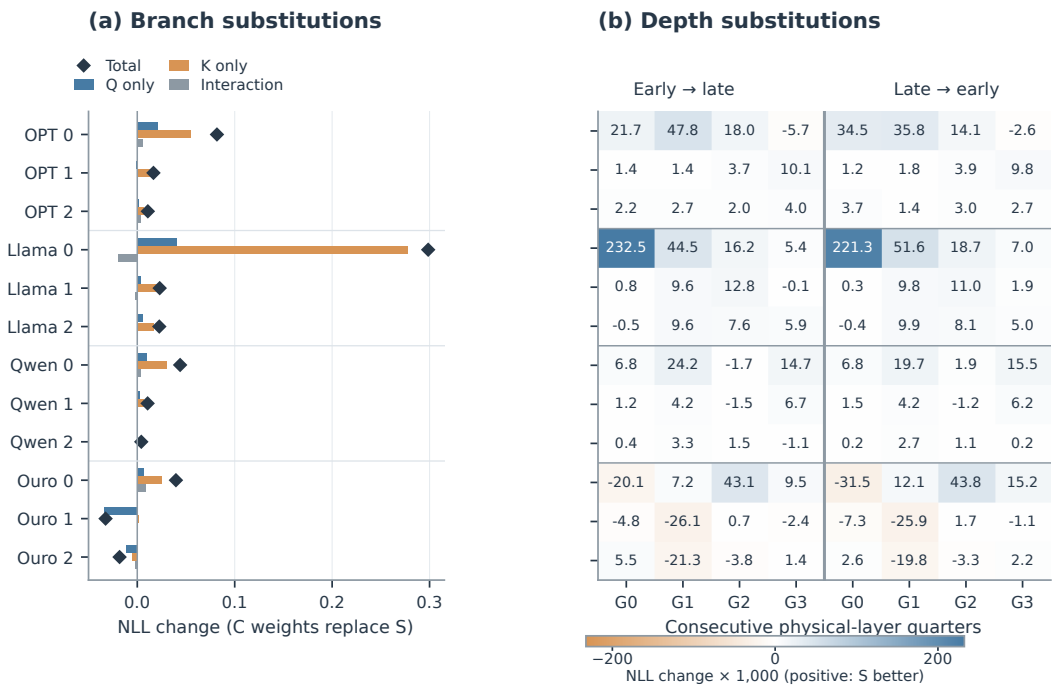

Figure 4: **Replacing fitted weights reveals branch and depth effects.** A separate 64-document confirmation set is used; 0/1/2 denote identity and the two random starts. Left (nats/token): Q-only (blue) and K-only (orange) effects, interaction (gray), and total C-minus-S NLL (diamond). Components sum to the total; intervals are in Appendix C.6. Right ($\times 10^3$): NLL changes from replacing depth groups in early-to-late (E) and reverse (R) order. Columns retain depth order; each four-group path sums to the same total. Blue denotes positive and orange negative changes. Branch and depth replacements are separate tests, not joint branch-by-layer attribution.

Matched S/C comparisons do not identify the best search strategy: at identity, M has lower NLL than S on both corpora for all four models. Sharing merits comparison with random initialization and multistart search. Ouro's opposing C4 and HellaSwag rankings also show why evaluation should include the intended tasks (Section 4.4).

The interventions locate branch and depth effects, but not the optimization steps that produced them, the specific Q layers responsible, or recurrence as the cause. Predicting the best sharing duration remains open.

Transfers to TWLA did not establish gains over its reference pipeline (Appendix E). MLP/full-network quantization, other bit widths, and learned sensitivity-weighted objectives remain untested. Intervals describe evaluation uncertainty for fixed fitted models, not variation across models or fitting seeds. Duration comparisons are descriptive. Deployment evidence is summarized in Appendix F.5.

6 Conclusion

Temporary QK sharing can improve independent fitting under restart and fitting-budget controls, with gains that depend on initialization, duration, and evaluation setting. Input-aware measurements and weight replacements distinguish reconstruction error from model behavior. Comparing shared and independent transforms requires controlling the fitting process, not just the allowed transforms.

Reproducibility Statement

Appendix A provides the projector and supporting derivations. Appendix B specifies the four-model fitting, scoring, diagnostics, and intervention protocols; Appendix C reports all outcomes. Appendices D and E give the duration sweep and endpoint-selection study. Appendix F consolidates earlier validation and its distinct protocols. The supplementary archive contains frozen source tables, plotting scripts, detailed historical records, and a standalone audit of the endpoint-selection NLLs and paired intervals. Compilation sources and numerical audit materials are provided separately.

Use of Large Language Models

The authors led the research and manuscript preparation. Large language models provided auxiliary support for research planning, code preparation and checking, analysis and mathematical reasoning, literature retrieval, figures, and manuscript drafting, revision, and translation. The authors reviewed the AI-assisted material and take responsibility for all scientific decisions, reported evidence, and final content.

References

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 151916–151951, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/df2034a516cbd617a96492cc476276c9-Abstract-Conference.html.

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

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M. De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4396–4429, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0df38cd13520747e1e64e5b123a78ef8-Abstract-Conference.html.

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

Marco Federici, Boris van Breugel, Paul Whatmough, and Markus Nagel. Dissecting quantization error: A concentration-alignment perspective. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.04359.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS. Also known as GPTQ in the author preprint and implementation.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation. Zenodo, 2023. URL https://zenodo.org/records/10256836. Framework reference; experiments use lm-eval 0.4.9.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Jinuk Kim, Marwa El Halabi, Wonpyo Park, Clemens J. S. Schaefer, Deokjae Lee, Yeonhong Park, Jae W. Lee, and Hyun Oh Song. GuidedQuant: Large language model quantization via exploiting end loss guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30011–30037, 2025. URL https://proceedings.mlr.press/v267/kim25d.html.

Deokjae Lee and Hyun Oh Song. Q-Palette: Fractional-bit quantizers toward optimal bit allocation for efficient LLM deployment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 11525–11558, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/10df8641edc837ca49360d5fb1d182f0-Abstract-Conference.html.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e5b1c0d4866f72393c522c8a00eed4eb-Abstract-Conference.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Meta. Llama 3.2 model card. Model documentation, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B/blob/main/README.md.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206, 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c6483c8a68083af3383f91ee0dc6db95-Abstract-Conference.html.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness matters for LLM quantization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume

267 of *Proceedings of Machine Learning Research*, pp. 57587–57613, 2025. URL https://proceedings.mlr.press/v267/sun25l.html.

Boris van Breugel, Yelysei Bondarenko, Paul Whatmough, and Markus Nagel. FPTQuant: Function-preserving transforms for LLM quantization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.04985v2. Author version v2, July 2026; first arXiv release June 2025.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.

Zhixiong Zhao, Zukang Xu, Zhixuan Chen, Xing Hu, Zhe Jiang, and Dawei Yang. TWLA: Achieving ternary weights and low-bit activations for LLMs via post-training quantization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.13054v2. Author version v2.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025. URL https://arxiv.org/abs/2510.25741.

A Mathematical Details

The derivations below support three parts of the experimental design: exact row projection, copying shared transforms into independent variables without changing the model, and comparing fixed model endpoints. None assumes that lower reconstruction error implies lower downstream loss.

A.1 Row-wise scaled ternary projection

For a row $w \in \mathbb{R}^d$, $d \geq 1$, the reconstruction uses one nonnegative scale and a ternary code:

$$(s^*, z^*) \in \operatorname*{arg\,min}_{s \geq 0,\ z \in \{-1,0,1\}^d} \|w - sz\|_2^2, \qquad Q_3(w) = s^* z^*. \tag{3}$$

The operator acts independently on each output row. Let $a_1 \geq \cdots \geq a_d \geq 0$ be the sorted magnitudes, with ties ordered by the original indices. For a fixed support S of size k, matching the signs of w gives error $\|w\|_2^2 - 2s \sum_{i \in S} |w_i| + ks^2$. Its minimizing scale is the mean magnitude on S. Among supports of size k, keeping the k largest magnitudes therefore minimizes the error. The exact projector chooses

$$k^* \in \operatorname*{arg\,max}_{1 \leq k \leq d} \frac{(\sum_{i=1}^k a_i)^2}{k}, \qquad s^* = \frac{1}{k^*} \sum_{i=1}^{k^*} a_i, \tag{4}$$

keeps those k^* signed entries, and sets the others to zero. Its squared error is $\|w\|_2^2 - (\sum_{i=1}^{k^*} a_i)^2 / k^*$. The implementation uses the first maximizing k and returns zero for an all-zero row. Floating-point summation may change the selected support near a tie.

A.2 Preserving the model when sharing is removed

At one layer, fix nonzero matrices $W_a \in \mathbb{R}^{m_a \times d}$, $a \in \{q, k, v\}$, the projector, and all other model parameters. A partition $\mathcal{P}$ specifies which branches share a rotation in $O(d) = \{R : R^\top R = I\}$. A refinement splits groups without merging any. Let $\mathcal{F}(\mathcal{P})$ be the set of effective-weight tuples obtainable under this constraint using Equation (1).

Proposition 1 (Model-preserving copying)**.** *If $\mathcal{P}'$ refines $\mathcal{P}$, copying each group's rotation to its new subgroups preserves every effective weight. Consequently $\mathcal{F}(\mathcal{P}) \subseteq \mathcal{F}(\mathcal{P}')$, and for any fixed model loss J,*

$$\inf_{\widehat{W} \in \mathcal{F}(\mathcal{P}')} J(\widehat{W}) \leq \inf_{\widehat{W} \in \mathcal{F}(\mathcal{P})} J(\widehat{W}). \tag{5}$$

This statement compares the lowest attainable losses, or their infima when no minimum is attained. It does not say which endpoint a finite search will return.

Proof. For each subgroup in $\mathcal{P}'$, use its parent group's rotation. Every branch then receives the same argument to Q_3 and the same inverse rotation in Equation (1). The effective weights are unchanged, proving inclusion; taking the infimum of the same loss over a superset proves Equation (5). Apply this construction at every physical layer, keeping all tied uses consistent. The same deterministic cast preserves the installed weights. □

For a nonzero branch matrix, define the relative local distortion by

$$f_a(R) = \frac{\|W_a R - Q_3(W_a R)\|_F^2}{\|W_a\|_F^2} = \frac{\|W_a - \widehat{W}_a(R)\|_F^2}{\|W_a\|_F^2}, \qquad R \in O(d). \tag{6}$$

The second equality follows by right-multiplying $W_a - \widehat{W}_a(R)$ by R and using Frobenius-norm invariance. Unlike M_W, f_a is defined before the backbone cast. Copying preserves installed weights when both representations use the same deterministic construction and cast; subsequent repair or optimization is a separate operation.

A.3 Minimum reconstruction error under sharing

Keep $\alpha_a = 1/3$ for every subset comparison. For a nonempty group $C \subseteq \{q, k, v\}$, define its ideal shared distortion and excess over separate rotations as

$$\ell(C) = \min_{R \in O(d)} \sum_{a \in C} \alpha_a f_a(R), \qquad \Delta(C) = \ell(C) - \sum_{a \in C} \ell(\{a\}). \tag{7}$$

For $D^*(\mathcal{P}) = \sum_{C \in \mathcal{P}} \ell(C)$ and the independent singleton partition Ind, the partition excess is

$$D^*(\mathcal{P}) - D^*(\mathrm{Ind}) = \sum_{C \in \mathcal{P}} \Delta(C). \tag{8}$$

These quantities compare ideal minimum errors; M_W in Equation (2) measures the error of the candidate returned by a search and installed in the model.

The minimum in Equation (7) exists. The row reconstruction set $\{sz : s \geq 0, z \in \{-1, 0, 1\}^d\}$ is a finite union of closed rays, and squared distance to this set is continuous. Hence $f_a(R)$ is continuous on the compact set $O(d)$. The tie-broken projected weight itself can be discontinuous, so the corresponding downstream loss need not attain its infimum in Equation (5).

For every R, $\alpha_a f_a(R) \geq \ell(\{a\})$. Summing and minimizing proves $\Delta(C) \geq 0$; adding and subtracting singleton minima gives Equation (8). Copying also shows that refining a partition cannot increase D^*. The same argument holds if every minimum is restricted to one common, nonempty finite set of rotations. It does not apply to separately searched candidate sets. Thus an error difference measured between search endpoints need not equal the nonnegative excess error imposed by a sharing constraint.

A.4 Lower local error need not give lower end loss

The following abstract example shows why minimizing reconstruction error need not minimize NLL. Take $\mathcal{E}_c = \{(0, 2)\} \subset \mathcal{E}_f = \{(0, 2), (1, 0)\}$ and local error $D(e) = e_1^2 + e_2^2$. Let $J(e) = \log(1 + \exp(e_1))$ be the binary label-one NLL for logit $-e_1$. Minimizing D selects distortion 4 in the smaller set and 1 in the larger, while the selected NLL rises from $\log 2$ to $\log(1 + \mathrm{e})$. The NLL infimum remains $\log 2$ in both sets. This example shows the logical distinction; it does not construct such a case within the ternary rotation family. Finite search introduces the additional possibility of missing the best reconstruction candidate.

A.5 Fixed-code update

Hold the complete scaled ternary reconstructions $B_a = Q_3(W_a R)$ fixed for a group C. Form

$$G_C = \sum_{a \in C} \frac{\alpha_a W_a^\top B_a}{\|W_a\|_F^2}, \qquad G_C = U \Sigma V^\top, \qquad R^+ = U V^\top. \tag{9}$$

Because the rotation is orthogonal, the quadratic terms in $W_a R$ are constant, giving

$$F_B(R) = \sum_{a \in C} \frac{\alpha_a \|W_a R - B_a\|_F^2}{\|W_a\|_F^2} = \text{constant} - 2\,\mathrm{tr}(R^\top G_C). \tag{10}$$

For $Z = U^\top R V \in O(d)$, $\mathrm{tr}(R^\top G_C) = \sum_i \sigma_i Z_{ii} \leq \sum_i \sigma_i$. Equality holds at $R^+ = U V^\top$, proving Equation (9). Rank deficiency can make the optimizer nonunique. The domain is $O(d)$, without a determinant-one constraint.

If $B_a = Q_3(W_a R)$ at the current strictly orthogonal state, re-projecting at the new state can only lower the fixed-code error. Exact arithmetic therefore yields

$$\sum_{a \in C} \alpha_a f_a(R^+) \leq F_B(R^+) \leq F_B(R) = \sum_{a \in C} \alpha_a f_a(R). \tag{11}$$

The exact alternating step is therefore non-increasing. The implementation uses the subset average $|C|^{-1} \sum_{a \in C} f_a$, which multiplies the fixed-$1/3$ objective by $3/|C|$ and leaves each group's minimizers unchanged. Comparisons across subsets retain the global weights.

The baseline solver uses FP32 matrix products, row projection, and proposals with FP64 reductions (Appendix F.2), so rounding can violate the exact descent chain. The stabilized solver evaluates its reconstruction objective in FP64 and checks fast proposals again in FP64 when their acceptance is ambiguous (Appendix F.2). Both perform finite searches; tolerance-based acceptance and tracking the best state found so far record numerical progress but do not certify stationarity or global optimality. Final repair, FP32 reconstruction, and native-dtype installation can further change the error; we therefore measure the pre-cast selection error and installed M_W separately.

A.6 Search schedules and initialization

The schedules below all return independent rotations. A round is an allocated sweep over the active groups; early stopping can reduce the updates actually performed.

Route	Search schedule	Comparison role
S	12 QK|V rounds, then 32 independent rounds	Shared prefix
C (I12R32)	12 independent rounds, restart, then 32 independent rounds	Matched phase boundary
I12R-CM-EXT	Same independent prefix and continuous second phase; cost-selected endpoint	Recorded-cost control
I44	44 independent rounds from the same start	Allocated branch rounds
M	20 and 19 independent rounds from two starts	Alternative candidate search
I40	40 independent rounds from S's first start	Allocated group updates
I-CM	Independent trajectory with cost-selected endpoint	Earlier cost control

Initialization pair 0 contains the identity and a fixed random rotation; the additional pairs each contain two random rotations. S, C, and the direct independent controls use the first rotation, while M fits from both. M selects each branch by lower reconstruction error after repair and before the backbone cast. Shared and independent rounds update two and three groups, respectively, so S and I40 allocate 120 group updates, whereas S, C, and I44 allocate 44 rounds per branch. These counts do not imply equal fitting time.

The I12R-CM-EXT run saves a separate copy of its state at second-phase round 32 as C. Its extension does not re-enter the phase or continue from repaired or cast endpoint weights; the cap is 128 second-phase rounds. Selection uses recorded solver-prefix cost plus one final repair, targeting S within 5%, with unmatched and early-stopped cases retained separately. This cost measure excludes the work needed to construct the full set of comparison endpoints and evaluate them.

A.7 Frozen-endpoint branch and depth contrasts

For the four hybrid states in Table 2, let Q_S, K_S, V_S and Q_C, K_C, V_C collect the installed branch weights across physical layers. With common V and all other parameters fixed, write $f(Q, K)$ for token-weighted NLL on identical documents and prediction tokens. The main-text quantities are

$$\begin{aligned} \Delta &= f(Q_C, K_C) - f(Q_S, K_S) = f(C) - f(S), \\ a_Q &= f(Q_C, K_S) - f(Q_S, K_S), \qquad a_K = f(Q_S, K_C) - f(Q_S, K_S), \\ H &= \Delta - a_Q - a_K, \qquad A = a_Q + a_K. \end{aligned} \tag{12}$$

Positive Δ favors S. Both single-branch effects use S as background. The remainder H measures the joint effect beyond their sum; it is computed from these four losses, not from an additional fitted model.

Let $B = \{q, k, v\}$. For $T \subseteq B$, construct a hybrid by installing C's branches in T at every physical layer and S's remaining branches. All other parameters, including projection biases, remain fixed. No hybrid is refitted or quantized again. Write $f(T)$ for its token-weighted NLL on a common set of documents and valid prediction tokens, so $f(\varnothing) = f(S)$ and $f(B) = f(C)$. The eight-mask protocol uses the standard finite-set contrasts

$$d(T) = f(T) - f(\varnothing), \qquad a(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} d(U), \qquad \Delta = d(B) = \sum_{\varnothing \neq T \subseteq B} a(T). \tag{13}$$

To see the last equality, substitute the definition of $a(T)$. The coefficient of a fixed nonempty $d(U)$ is $\sum_{U \subseteq T \subseteq B} (-1)^{|T|-|U|}$, equal to zero unless $U = B$, when it is one; $d(\varnothing) = 0$. This identity holds for any finite collection of loss values. It assumes neither a linear model nor small perturbations.

Let $A = a(\{q\}) + a(\{k\}) + a(\{v\})$ denote the singleton sum. Then $H = \Delta - A$ sums all higher-order contrasts. In the 12 configurations used for the main frozen-endpoint intervention, the installed V tensors are bitwise equal between S and C. Consequently $f(T \cup \{v\}) = f(T)$ for $v \notin T$, and all contrasts containing V vanish. The decomposition reduces to Equation (12), with $H = a(\{q, k\})$. These terms vanish because the V tensors are identical; this does not show that the model is insensitive to V errors. The descriptive Shapley allocation $\phi_b = \sum_{T \ni b} a(T)/|T|$ divides each interaction equally among its participating branches and sums to Δ.

For reverse replacements, the contrast is $r(T) = f(B \setminus T) - f(B)$. Its total is $-\Delta$, but its single-branch effects need not negate those measured on S: the background differs. In particular, with common V, replacing Q on the C background has effect $-(a_Q + H)$, whereas its forward effect is a_Q.

For depth replacements, index physical layers by $l = 0, \ldots, L-1$ and form four consecutive groups $G_j = \{l : \lfloor jL/4 \rfloor \le l < \lfloor (j+1)L/4 \rfloor\}$, $j = 0, \ldots, 3$. Let Z_t^π install C's QKV weights in the first t groups of ordering π and S's weights elsewhere. For early-to-late and late-to-early orders, respectively, $\pi = (0, 1, 2, 3)$ and $(3, 2, 1, 0)$; $Z_0^\pi = S$ and $Z_4^\pi = C$. Define

$$b_t^\pi = f(Z_t^\pi) - f(Z_{t-1}^\pi), \qquad \sum_{t=1}^{4} b_t^\pi = f(C) - f(S) = \Delta. \tag{14}$$

Cancellation of consecutive losses gives the identity. A group's marginal effect may change with the order because the other groups supply different backgrounds. For tied or recurrent layers, a physical-weight replacement affects every use of that parameter, not a single recurrent step. These contrasts show how replacing branches or layer groups changes the loss between two fixed endpoints. They do not describe the optimization trajectory that produced those endpoints.

B PROTOCOLS FOR THE FOUR-MODEL EXTENSION

This section describes the main four-model comparison and the diagnostic and intervention evaluations that followed it. Later appendices report earlier experiments under their original protocols. Those experiments used different evaluation samples and, in some cases, different numerical precision or task-scoring implementations, so we do not pool their absolute scores with the current results.

Models and fitting. OPT-1.3B, Llama-3.2-3B, Qwen3-4B-Base, and Ouro-1.4B have respectively 24, 28, 36, and 24 physical layers. The backbone for this extension is BF16. Ouro uses four recurrent applications of its tied layers, and only the final-step logits enter cross-entropy; the model's weighted intermediate-exit training loss is not used. An intervention on a tied weight applies at every recurrent use, but the weight is counted once in reconstruction and cost accounting. The shared and independent objectives normalize each Q/K/V branch by its own weight norm, rather than weighting branches by their different output dimensions.

All fitting uses pretrained weights only. Identity and two fixed random orthogonal initializations are retained for every model. S has a 12-round shared-QK prefix followed by 32 independent rounds; C has a 12-round independent prefix followed by the same phase-entry and 32-round independent continuation. Restarting resets stopping state but draws no new rotation. Copying preserves the installed model before subsequent numerical repair or optimization. The route table in Appendix A.6 distinguishes the other controls. No NLL or task outcome selects a checkpoint.

Expanded recorded-cost control. C-CM extends the independent run with the same restart as S. It differs from the earlier direct I-CM control. Its snapshot at second-phase round 32 is C. It continues the same unmodified solver state up to a maximum of 128 second-phase rounds, without re-entering the phase or continuing from repaired/cast evaluation weights. We select a checkpoint by comparing its recorded solver cost with S's fixed recorded cost. This cost includes phase entry, the iterations through the selected checkpoint, and one final repair. The $\pm 5\%$ tolerance applies to each

physical layer and initialization, rather than only to the model total. The six unmatched Qwen units are retained and explicitly reported. The numerator excludes evaluations, model loading, checkpoint I/O, and continuation work beyond the selected endpoint; therefore, these ratios do not measure total experimental cost.

Initialization pairs. In the current-suite tables, I denotes identity (index 0), while R1 and R2 denote the two fixed random initializations (indices 1 and 2). For M, the label identifies the first member of its two-start pair, not both members. Scores are not averaged across these initializations. A dash in an FP initialization field means that the same unquantized reference applies to every start. Historical experiments retain their original pair identifiers.

Index OPT, Llama, Qwen, and Ouro by $m = 0, 1, 2, 3$, respectively, and physical layers by $l = 0, \ldots, L-1$. Pair 0 combines the identity with a random rotation generated using seed $20260913 + 10000m + 10l + 1$. For pairs $b = 1, 2$, start $s = 1, 2$ uses seed $2026104000 + 100000(m+1) + 1000b + 10l + s$. Random rotations use FP64 QR decomposition of a Gaussian matrix drawn by NumPy `default_rng`; columns of Q are multiplied by the signs of the diagonal of R, with a zero diagonal assigned positive sign. The random member of pair 0 is rounded to FP32 before conversion to FP64; the other pairs retain FP64. S and C use the first member of each pair, while M uses both. These pairs are distinct from the duration study's starts 100 and 101.

Primary NLL evaluation. C4 uses the frozen 128-document validation cohort, with each document individually tokenized and truncated to at most 512 tokens. WikiText-2 uses the full raw test stream and the frozen tokenizer/join convention, in disjoint 512-token chunks; the final shorter chunk is kept and the first token of each chunk is unscored. For either corpus,

$$\mathrm{NLL} = \frac{\sum_i \mathrm{loss_sum}_i}{\sum_i \mathrm{token_count}_i}. \tag{15}$$

Each paired comparison uses the same evaluation items and the same numbers of scored target tokens. C4 bootstrap units are documents; WikiText units are clusters of eight consecutive chunks, with any final partial cluster retained. The 24 primary S–C contrasts form one Bonferroni-adjusted family. These simultaneous intervals use a normal approximation, $\widehat{\Delta} \pm z_{1-0.05/(2\cdot 24)} \widehat{\mathrm{SE}}$, with paired-bootstrap standard errors; the ordinary 95% intervals use bootstrap percentiles. Secondary S–M, S–I44, and S–C-CM contrasts are reported separately. The bootstrap intervals describe uncertainty for these fixed fitted endpoints; they do not include variability from refitting.

Tasks. HellaSwag uses all 10,042 validation examples, ARC-Challenge all 1,172 test examples, and PIQA all 1,838 validation examples. Candidate choices are scored with the same prompts, tokenizer, and answer boundaries within each task. Both raw and length-normalized accuracy are retained; normalized accuracy is used for the main three-task comparisons. Exact two-sided paired McNemar tests receive Holm correction over the 36 primary S–C contrasts, with the 108 secondary comparisons forming a separate family. MMLU covers all 14,042 test questions in 57 subjects at initializations 0 and 1 only, using zero-shot, unnormalized A/B/C/D log-likelihood. Its 16 S–C and S–M tests form a separate Holm family. Micro accuracy weights questions equally and macro accuracy weights subjects equally. These protocols differ from evaluations that use five demonstrations or score generated answers.

Input-aware diagnostics. Write $E_{l,a} = \overline{W}_{l,a} - W_{l,a}$. A site $s = (l, r, a)$ identifies a physical layer, recurrent step, and QKV branch, and $X_s \in \mathbb{R}^{d_l \times n_s}$ contains its n_s calibration input columns. Let $\mathcal{S}$ be the set of sites: $|\mathcal{S}| = 3LR$, with $R = 4$ for Ouro and $R = 1$ otherwise. In addition to M_W, we evaluate relative diagonal-input-weighted error and output reconstruction error:

$$M_{\mathrm{diag}} = \frac{1}{|\mathcal{S}|} \sum_{s=(l,r,a)\in\mathcal{S}} \frac{\mathrm{tr}(E_{l,a} D_s E_{l,a}^\top)}{\mathrm{tr}(W_{l,a} D_s W_{l,a}^\top)}, \qquad D_s = \mathrm{diag}\left(\frac{X_s X_s^\top}{n_s}\right), \tag{16}$$

$$M_{\mathrm{out}} = \frac{1}{|\mathcal{S}|} \sum_{s=(l,r,a)\in\mathcal{S}} \frac{\|E_{l,a} X_s\|_F^2}{\|W_{l,a} X_s\|_F^2}. \tag{17}$$

Here $\mathrm{diag}(A)$ retains the diagonal of A and sets its off-diagonal entries to zero. Inputs are captured from the original unquantized BF16 model at each actual QKV invocation, restricted to positions

with a scored next token. The same reference activations and masks are used for both candidates. Every recorded denominator in these ratios was finite and strictly positive, so no additive epsilon is used. Each site is normalized before taking the equal-weight average; Ouro's four recurrent steps are not pooled into a single ratio. Diagnostic matrix products use FP32 with TF32 disabled and FP64 sums. Ranking uses calibration measurements only; the confirmation NLL cohort is separate. These measurements diagnose frozen candidates; they are not used to fit candidates with sensitivity weights. Complete metrics and confirmation scores appear in Appendix C.3; Appendix C.4 specifies the direction/context interventions.

Frozen-endpoint interventions. The final S/C comparison evaluates all eight branch-substitution masks and both four-group depth paths on a separate cohort of 64 confirmation documents; 16 discovery documents are not pooled into confirmation. All substitutions use the same installed endpoint weights, target masks, and native BF16 forward computation with FP32 cross-entropy. The physical layers are divided into four consecutive groups. Ouro's physical-layer interventions affect all recurrent uses. No transform is refitted, and no favorable layer or replacement direction is selected on confirmation data. All states share 100,000 paired document-bootstrap indices. The 36 total/singleton-sum/interaction contrasts, 84 branch terms, and 96 depth marginals receive separate simultaneous corrections. The 48 differences between depth paths are exploratory. The corrections do not jointly control the error rate across these families. Complete intervals and the distinct cohort scores are in Appendix C.6.

C Complete four-model results and frozen-endpoint interventions

This section reports the expanded evaluation for every model and initialization bank. The short model labels OPT, Llama, Qwen, and Ouro denote OPT-1.3B, Llama-3.2-3B, Qwen3-4B-Base, and Ouro-1.4B, respectively. Each of the three banks specifies a fixed initialization configuration; they are not independent samples of model replications. We write C for the fixed I12R32 schedule and C-CM for I12R-CM-EXT. All displayed values are generated from the accompanying frozen CSV tables by `scripts/build_v10_tables.py`; the portable source manifest records their hashes. NLL differences are in nats per scored token unless otherwise stated. The main suite, diagnostic confirmation suite, and causal confirmation suite have distinct evaluation cohorts; their absolute NLL values refer to different samples and cannot be used interchangeably.

C.1 All endpoints, task outcomes, and fitting-cost controls

Table 6 reports all five quantized arms in each of the 12 cells and one FP reference per model. Table 7 reports all 24 primary S–C contrasts across the two corpora: 18 adjusted intervals favor S, four favor C, and two contain zero. The four C-favoring contrasts are the two random-bank Ouro cells on both corpora. The intervals describe paired evaluation uncertainty conditional on the frozen endpoints. The secondary comparisons in Table 8 include cases where M or a resource control performs better than S.

Table 6: All main-suite endpoints. C denotes I12R32; C-CM denotes I12R-CM-EXT. NLL and perplexity are lower-is-better. Time is recorded solver-endpoint seconds, not total calibration or inference time. FP is one reference per model.

Model	Init.	Route	WikiText-2		C4		Solver time
			NLL	PPL	NLL	PPL	(s)
OPT	–	FP	2.9822	19.73	2.8284	16.92	–
OPT	I	S	3.3059	27.27	3.0654	21.44	2979.2
OPT	I	C	3.4471	31.41	3.1610	23.59	2977.0
OPT	I	M	3.2563	25.95	3.0166	20.42	2679.8
OPT	I	I44	3.4471	31.41	3.1610	23.59	2375.9
OPT	I	C-CM	3.4462	31.38	3.1609	23.59	2933.9
OPT	R1	S	3.1288	22.85	2.9217	18.57	3539.8
OPT	R1	C	3.1573	23.51	2.9400	18.92	3728.7

Table 6 (continued)

Model	Init.	Route	WikiText-2		C4		Solver time
			NLL	PPL	NLL	PPL	(s)
OPT	R1	M	3.1618	23.61	2.9385	18.89	3478.1
OPT	R1	I44	3.1573	23.51	2.9400	18.92	3115.7
OPT	R1	C-CM	3.1574	23.51	2.9401	18.92	3500.5
OPT	R2	S	3.1357	23.01	2.9322	18.77	3510.5
OPT	R2	C	3.1620	23.62	2.9421	18.95	3706.6
OPT	R2	M	3.1644	23.67	2.9447	19.00	3483.6
OPT	R2	I44	3.1620	23.62	2.9421	18.95	3054.8
OPT	R2	C-CM	3.1623	23.63	2.9422	18.96	3486.5
Llama	–	FP	2.3240	10.22	2.4586	11.69	–
Llama	I	S	2.7081	15.00	2.7907	16.29	8960.9
Llama	I	C	3.0361	20.82	3.0600	21.33	9349.0
Llama	I	M	2.4922	12.09	2.6157	13.68	11179.9
Llama	I	I44	3.0356	20.81	3.0593	21.31	7240.7
Llama	I	C-CM	3.0366	20.83	3.0597	21.32	8951.4
Llama	R1	S	2.4660	11.78	2.6093	13.59	10001.3
Llama	R1	C	2.4899	12.06	2.6277	13.84	10964.5
Llama	R1	M	2.4925	12.09	2.6319	13.90	13052.0
Llama	R1	I44	2.4900	12.06	2.6273	13.84	8786.3
Llama	R1	C-CM	2.4897	12.06	2.6275	13.84	10001.8
Llama	R2	S	2.4592	11.70	2.5892	13.32	9854.1
Llama	R2	C	2.4790	11.93	2.6153	13.67	10776.5
Llama	R2	M	2.4810	11.95	2.6241	13.79	12893.3
Llama	R2	I44	2.4789	11.93	2.6150	13.67	8610.2
Llama	R2	C-CM	2.4789	11.93	2.6148	13.66	9856.3
Qwen	–	FP	2.3305	10.28	2.6364	13.96	–
Qwen	I	S	2.5541	12.86	2.8321	16.98	19236.5
Qwen	I	C	2.6268	13.83	2.8854	17.91	21510.5
Qwen	I	M	2.4632	11.74	2.7631	15.85	21608.0
Qwen	I	I44	2.6266	13.83	2.8851	17.91	12516.7
Qwen	I	C-CM	2.6295	13.87	2.8858	17.92	19389.1
Qwen	R1	S	2.4558	11.66	2.7565	15.74	14433.4
Qwen	R1	C	2.4714	11.84	2.7716	15.98	17614.8
Qwen	R1	M	2.4693	11.81	2.7673	15.91	19871.7
Qwen	R1	I44	2.4714	11.84	2.7717	15.98	10659.8
Qwen	R1	C-CM	2.4725	11.85	2.7720	15.99	14487.5
Qwen	R2	S	2.4569	11.67	2.7533	15.69	14703.9
Qwen	R2	C	2.4558	11.66	2.7549	15.72	17959.0
Qwen	R2	M	2.4713	11.84	2.7528	15.69	20187.2
Qwen	R2	I44	2.4556	11.65	2.7549	15.72	10840.1
Qwen	R2	C-CM	2.4576	11.68	2.7560	15.74	14561.7
Ouro	–	FP	2.4478	11.56	2.6333	13.92	–
Ouro	I	S	2.8883	17.96	2.9839	19.76	2644.6
Ouro	I	C	2.9451	19.01	3.0224	20.54	2817.7
Ouro	I	M	2.8447	17.20	2.9366	18.85	2687.2
Ouro	I	I44	2.9453	19.02	3.0226	20.55	2213.1
Ouro	I	C-CM	2.9453	19.02	3.0232	20.56	2620.3
Ouro	R1	S	2.9185	18.51	3.0231	20.55	3208.0
Ouro	R1	C	2.8769	17.76	2.9792	19.67	3598.9
Ouro	R1	M	2.9040	18.25	3.0115	20.32	3495.4
Ouro	R1	I44	2.8769	17.76	2.9792	19.67	2981.9
Ouro	R1	C-CM	2.8765	17.75	2.9792	19.67	3200.5
Ouro	R2	S	2.8646	17.54	2.9675	19.44	3146.4
Ouro	R2	C	2.8469	17.23	2.9466	19.04	3554.3
Ouro	R2	M	2.8091	16.60	2.9245	18.62	3468.6
Ouro	R2	I44	2.8469	17.23	2.9466	19.04	2944.8

Ouro	R2	C-CM	2.8472	17.24	2.9467	19.04	3127.5

Table 7: All 24 main-suite S minus C NLL contrasts (nats/token). Negative favors S. Fav. identifies the route favored by the 24-contrast Bonferroni-adjusted interval; ? denotes an interval containing zero. WT: WikiText-2; C4: frozen C4 cohort.

Model	Init.	Data	$S - C$	95% CI	Adjusted CI	Fav.
OPT	I	C4	-0.0956	[-0.1057, -0.0855]	[-0.1114, -0.0798]	S
OPT	I	WT	-0.1413	[-0.1469, -0.1356]	[-0.1502, -0.1323]	S
OPT	R1	C4	-0.0183	[-0.0210, -0.0156]	[-0.0226, -0.0140]	S
OPT	R1	WT	-0.0285	[-0.0301, -0.0270]	[-0.0309, -0.0261]	S
OPT	R2	C4	-0.0098	[-0.0135, -0.0061]	[-0.0157, -0.0040]	S
OPT	R2	WT	-0.0263	[-0.0280, -0.0246]	[-0.0290, -0.0236]	S
Llama	I	C4	-0.2693	[-0.2913, -0.2488]	[-0.3031, -0.2356]	S
Llama	I	WT	-0.3280	[-0.3398, -0.3162]	[-0.3467, -0.3092]	S
Llama	R1	C4	-0.0184	[-0.0247, -0.0128]	[-0.0278, -0.0090]	S
Llama	R1	WT	-0.0239	[-0.0255, -0.0224]	[-0.0263, -0.0215]	S
Llama	R2	C4	-0.0262	[-0.0360, -0.0179]	[-0.0403, -0.0120]	S
Llama	R2	WT	-0.0198	[-0.0212, -0.0185]	[-0.0219, -0.0177]	S
Qwen	I	C4	-0.0533	[-0.0590, -0.0478]	[-0.0621, -0.0445]	S
Qwen	I	WT	-0.0727	[-0.0783, -0.0675]	[-0.0812, -0.0642]	S
Qwen	R1	C4	-0.0151	[-0.0202, -0.0106]	[-0.0226, -0.0076]	S
Qwen	R1	WT	-0.0156	[-0.0178, -0.0133]	[-0.0192, -0.0121]	S
Qwen	R2	C4	-0.0016	[-0.0054, +0.0021]	[-0.0075, +0.0043]	?
Qwen	R2	WT	+0.0011	[-0.0008, +0.0032]	[-0.0020, +0.0043]	?
Ouro	I	C4	-0.0386	[-0.0455, -0.0316]	[-0.0496, -0.0275]	S
Ouro	I	WT	-0.0568	[-0.0607, -0.0532]	[-0.0627, -0.0509]	S
Ouro	R1	C4	+0.0439	[+0.0376, +0.0504]	[+0.0336, +0.0542]	C
Ouro	R1	WT	+0.0415	[+0.0383, +0.0447]	[+0.0365, +0.0466]	C
Ouro	R2	C4	+0.0209	[+0.0161, +0.0261]	[+0.0132, +0.0287]	C
Ouro	R2	WT	+0.0177	[+0.0145, +0.0209]	[+0.0127, +0.0227]	C

Table 8: All 72 secondary main-suite NLL comparisons against M, I44, and the recorded-cost control. Negative differences favor S. Intervals are marginal paired 95% intervals only, without a simultaneous significance claim. The cost interpretation remains subject to the unmatched Qwen bank-0 layers.

Model	Init.	Data	Pair	Difference	95% CI
OPT	I	C4	S-C-CM	-0.0954	[-0.1055, -0.0853]
OPT	I	WT	S-C-CM	-0.1404	[-0.1459, -0.1347]
OPT	I	C4	S-I44	-0.0956	[-0.1057, -0.0855]
OPT	I	WT	S-I44	-0.1413	[-0.1469, -0.1356]
OPT	I	C4	S-M	+0.0488	[+0.0408, +0.0570]
OPT	I	WT	S-M	+0.0495	[+0.0431, +0.0560]
OPT	R1	C4	S-C-CM	-0.0184	[-0.0212, -0.0157]
OPT	R1	WT	S-C-CM	-0.0287	[-0.0302, -0.0271]
OPT	R1	C4	S-I44	-0.0183	[-0.0210, -0.0156]
OPT	R1	WT	S-I44	-0.0285	[-0.0301, -0.0270]
OPT	R1	C4	S-M	-0.0168	[-0.0204, -0.0132]
OPT	R1	WT	S-M	-0.0330	[-0.0351, -0.0309]
OPT	R2	C4	S-C-CM	-0.0100	[-0.0138, -0.0062]
OPT	R2	WT	S-C-CM	-0.0266	[-0.0284, -0.0249]
OPT	R2	C4	S-I44	-0.0098	[-0.0135, -0.0061]
OPT	R2	WT	S-I44	-0.0263	[-0.0280, -0.0246]
OPT	R2	C4	S-M	-0.0124	[-0.0166, -0.0083]

Table 8 (continued)

Model	Init.	Data	Pair	Difference	95% CI
OPT	R2	WT	S-M	-0.0287	[-0.0314, -0.0261]
Llama	I	C4	S-C-CM	-0.2690	[-0.2910, -0.2485]
Llama	I	WT	S-C-CM	-0.3284	[-0.3403, -0.3167]
Llama	I	C4	S-I44	-0.2687	[-0.2906, -0.2484]
Llama	I	WT	S-I44	-0.3275	[-0.3394, -0.3157]
Llama	I	C4	S-M	+0.1749	[+0.1473, +0.2084]
Llama	I	WT	S-M	+0.2159	[+0.2064, +0.2256]
Llama	R1	C4	S-C-CM	-0.0182	[-0.0245, -0.0127]
Llama	R1	WT	S-C-CM	-0.0237	[-0.0252, -0.0221]
Llama	R1	C4	S-I44	-0.0180	[-0.0242, -0.0124]
Llama	R1	WT	S-I44	-0.0239	[-0.0255, -0.0224]
Llama	R1	C4	S-M	-0.0226	[-0.0306, -0.0151]
Llama	R1	WT	S-M	-0.0265	[-0.0285, -0.0244]
Llama	R2	C4	S-C-CM	-0.0256	[-0.0354, -0.0174]
Llama	R2	WT	S-C-CM	-0.0197	[-0.0211, -0.0185]
Llama	R2	C4	S-I44	-0.0258	[-0.0355, -0.0176]
Llama	R2	WT	S-I44	-0.0197	[-0.0210, -0.0184]
Llama	R2	C4	S-M	-0.0349	[-0.0456, -0.0257]
Llama	R2	WT	S-M	-0.0219	[-0.0237, -0.0200]
Qwen	I	C4	S-C-CM	-0.0537	[-0.0594, -0.0483]
Qwen	I	WT	S-C-CM	-0.0754	[-0.0812, -0.0700]
Qwen	I	C4	S-I44	-0.0530	[-0.0589, -0.0475]
Qwen	I	WT	S-I44	-0.0725	[-0.0780, -0.0673]
Qwen	I	C4	S-M	+0.0691	[+0.0610, +0.0774]
Qwen	I	WT	S-M	+0.0909	[+0.0870, +0.0949]
Qwen	R1	C4	S-C-CM	-0.0156	[-0.0208, -0.0110]
Qwen	R1	WT	S-C-CM	-0.0167	[-0.0189, -0.0145]
Qwen	R1	C4	S-I44	-0.0152	[-0.0203, -0.0106]
Qwen	R1	WT	S-I44	-0.0156	[-0.0178, -0.0133]
Qwen	R1	C4	S-M	-0.0108	[-0.0165, -0.0057]
Qwen	R1	WT	S-M	-0.0134	[-0.0170, -0.0101]
Qwen	R2	C4	S-C-CM	-0.0027	[-0.0065, +0.0014]
Qwen	R2	WT	S-C-CM	-0.0006	[-0.0026, +0.0014]
Qwen	R2	C4	S-I44	-0.0016	[-0.0053, +0.0023]
Qwen	R2	WT	S-I44	+0.0013	[-0.0007, +0.0033]
Qwen	R2	C4	S-M	+0.0005	[-0.0037, +0.0047]
Qwen	R2	WT	S-M	-0.0144	[-0.0171, -0.0117]
Ouro	I	C4	S-C-CM	-0.0393	[-0.0463, -0.0324]
Ouro	I	WT	S-C-CM	-0.0569	[-0.0608, -0.0533]
Ouro	I	C4	S-I44	-0.0388	[-0.0457, -0.0319]
Ouro	I	WT	S-I44	-0.0570	[-0.0608, -0.0533]
Ouro	I	C4	S-M	+0.0473	[+0.0349, +0.0601]
Ouro	I	WT	S-M	+0.0436	[+0.0347, +0.0525]
Ouro	R1	C4	S-C-CM	+0.0439	[+0.0377, +0.0504]
Ouro	R1	WT	S-C-CM	+0.0420	[+0.0387, +0.0452]
Ouro	R1	C4	S-I44	+0.0439	[+0.0376, +0.0504]
Ouro	R1	WT	S-I44	+0.0415	[+0.0383, +0.0447]
Ouro	R1	C4	S-M	+0.0116	[+0.0026, +0.0209]
Ouro	R1	WT	S-M	+0.0145	[+0.0085, +0.0201]
Ouro	R2	C4	S-C-CM	+0.0208	[+0.0160, +0.0260]
Ouro	R2	WT	S-C-CM	+0.0174	[+0.0142, +0.0206]
Ouro	R2	C4	S-I44	+0.0209	[+0.0161, +0.0261]
Ouro	R2	WT	S-I44	+0.0177	[+0.0145, +0.0209]
Ouro	R2	C4	S-M	+0.0430	[+0.0329, +0.0531]
Ouro	R2	WT	S-M	+0.0555	[+0.0488, +0.0623]

C.2 TASK-LEVEL OUTCOMES AND COST COVERAGE

Tables 9 and 10 separate raw task accuracy, length-normalized accuracy, and paired tests. These outcomes are not averaged with NLL or treated as interchangeable evidence. Table 11 compares the recorded fitting time needed to reach the selected solver endpoints. It does not measure total calibration wall-clock time, GPU-hours for producing all candidate endpoints, or inference latency. In particular, Qwen bank 0 has only 30 of 36 layers inside the prescribed five-percent matching window. We report this endpoint but do not claim that its cost is matched for the whole model.

Table 9: All task results (percent). For each task, raw and length-normalized accuracy are both retained. HellaSwag has 10,042 items, ARC-Challenge 1,172, and PIQA 1,838. FP is one reference per model, shared across initializations.

Model	Init.	Route	HellaSwag		ARC-Challenge		PIQA	
			Raw	Norm.	Raw	Norm.	Raw	Norm.
OPT	–	FP	41.51	53.70	23.46	29.44	71.65	72.20
OPT	I	S	37.00	47.51	21.93	26.79	68.88	69.48
OPT	I	C	36.62	46.30	22.61	27.30	69.21	70.24
OPT	I	M	37.76	48.34	22.87	26.28	69.21	70.40
OPT	I	I44	36.62	46.30	22.61	27.30	69.21	70.24
OPT	I	C-CM	36.40	46.18	22.87	26.88	68.88	70.08
OPT	R1	S	38.76	49.84	23.29	26.45	70.67	71.11
OPT	R1	C	38.40	49.35	23.98	27.39	70.40	71.00
OPT	R1	M	38.63	49.21	23.63	27.47	70.40	71.11
OPT	R1	I44	38.40	49.35	23.98	27.39	70.40	71.00
OPT	R1	C-CM	38.40	49.36	23.81	26.62	70.57	71.11
OPT	R2	S	38.85	49.27	24.23	28.24	70.46	71.00
OPT	R2	C	38.61	49.13	23.63	27.82	70.62	70.46
OPT	R2	M	38.20	48.99	22.87	26.96	70.29	71.00
OPT	R2	I44	38.61	49.13	23.63	27.82	70.62	70.46
OPT	R2	C-CM	38.69	49.11	23.63	27.73	70.62	70.40
Llama	–	FP	55.30	73.64	42.15	45.73	76.55	77.20
Llama	I	S	46.53	62.16	33.36	36.95	73.67	73.23
Llama	I	C	40.29	52.81	27.13	30.38	68.39	69.31
Llama	I	M	51.78	69.89	37.46	42.49	75.84	76.50
Llama	I	I44	40.44	52.78	27.39	30.89	68.72	68.61
Llama	I	C-CM	40.31	52.86	27.47	30.80	68.44	69.10
Llama	R1	S	52.07	70.20	38.31	42.58	75.68	75.63
Llama	R1	C	51.10	68.99	37.97	41.38	75.41	75.57
Llama	R1	M	50.98	68.38	37.63	40.78	74.92	76.17
Llama	R1	I44	51.07	68.92	37.71	41.21	75.46	75.63
Llama	R1	C-CM	51.12	69.02	37.54	41.38	75.08	75.57
Llama	R2	S	51.77	69.80	38.82	43.34	75.90	76.44
Llama	R2	C	51.20	68.91	38.14	42.24	75.30	76.12
Llama	R2	M	51.06	68.51	37.03	41.04	74.81	76.01
Llama	R2	I44	51.29	68.83	38.57	42.41	75.46	76.50
Llama	R2	C-CM	51.26	68.87	38.23	42.49	75.46	76.44
Qwen	–	FP	54.54	73.75	47.95	51.62	77.80	77.91
Qwen	I	S	49.53	66.02	44.28	47.95	74.48	75.90
Qwen	I	C	47.53	63.88	44.28	48.55	73.01	74.70
Qwen	I	M	51.26	68.96	46.84	51.37	76.22	77.64
Qwen	I	I44	47.44	63.90	44.20	48.29	73.07	74.43
Qwen	I	C-CM	47.69	63.86	44.20	48.55	72.91	74.86
Qwen	R1	S	51.18	68.70	47.18	50.68	75.63	76.71
Qwen	R1	C	50.91	68.23	48.12	51.45	75.63	76.06
Qwen	R1	M	51.17	68.47	43.17	46.59	75.73	76.50
Qwen	R1	I44	50.90	68.09	47.95	51.88	75.46	76.12
Qwen	R1	C-CM	50.90	68.21	47.18	50.94	75.95	76.28
Qwen	R2	S	51.74	69.31	44.88	47.70	76.55	76.99

Table 9 (continued)

Model	Init.	Route	HellaSwag		ARC-Challenge		PIQA	
			Raw	Norm.	Raw	Norm.	Raw	Norm.
Qwen	R2	C	51.11	68.48	45.73	49.49	76.77	76.93
Qwen	R2	M	51.18	68.88	46.93	49.74	76.50	77.20
Qwen	R2	I44	51.00	68.50	46.25	49.06	77.09	77.04
Qwen	R2	C-CM	51.07	68.43	46.33	49.66	76.93	76.61
Ouro	–	FP	53.74	71.71	50.26	52.30	78.13	78.51
Ouro	I	S	45.71	61.80	42.49	46.59	75.24	75.63
Ouro	I	C	44.64	60.16	42.49	45.39	74.86	75.63
Ouro	I	M	47.38	64.12	42.49	46.59	75.30	76.39
Ouro	I	I44	44.58	60.24	41.89	45.22	75.14	75.30
Ouro	I	C-CM	44.48	60.14	41.98	45.56	75.19	75.46
Ouro	R1	S	46.16	63.50	45.39	49.06	75.95	76.50
Ouro	R1	C	45.79	62.72	45.39	47.95	75.35	76.77
Ouro	R1	M	46.03	62.43	42.49	46.16	75.57	76.22
Ouro	R1	I44	45.79	62.72	45.39	47.95	75.35	76.77
Ouro	R1	C-CM	45.80	62.69	44.71	47.95	75.57	76.66
Ouro	R2	S	45.92	62.76	45.14	48.98	74.86	75.24
Ouro	R2	C	46.30	62.63	45.73	47.87	74.86	75.30
Ouro	R2	M	46.75	63.58	44.97	47.01	74.27	74.37
Ouro	R2	I44	46.30	62.63	45.73	47.87	74.86	75.30
Ouro	R2	C-CM	46.24	62.59	45.73	48.29	74.54	75.03

Table 10: Primary S versus C task tests using normalized accuracy. Differences are percentage points; p-values are exact paired McNemar values after Holm correction over the 36 primary S–C task contrasts.

Model	Init.	Task	$S - C$ (pp)	p_{Holm}	Signif.
OPT	I	ARC	-0.51	1	no
OPT	I	HS	+1.21	4.02e-07	yes
OPT	I	PIQA	-0.76	1	no
OPT	R1	ARC	-0.94	1	no
OPT	R1	HS	+0.49	0.24	no
OPT	R1	PIQA	+0.11	1	no
OPT	R2	ARC	+0.43	1	no
OPT	R2	HS	+0.14	1	no
OPT	R2	PIQA	+0.54	1	no
Llama	I	ARC	+6.57	4.37e-09	yes
Llama	I	HS	+9.35	1.19e-191	yes
Llama	I	PIQA	+3.92	5.61e-06	yes
Llama	R1	ARC	+1.19	1	no
Llama	R1	HS	+1.20	9.74e-10	yes
Llama	R1	PIQA	+0.05	1	no
Llama	R2	ARC	+1.11	1	no
Llama	R2	HS	+0.89	8.6e-06	yes
Llama	R2	PIQA	+0.33	1	no
Qwen	I	ARC	-0.60	1	no
Qwen	I	HS	+2.14	7.54e-19	yes
Qwen	I	PIQA	+1.20	0.853	no
Qwen	R1	ARC	-0.77	1	no
Qwen	R1	HS	+0.47	0.537	no
Qwen	R1	PIQA	+0.65	1	no
Qwen	R2	ARC	-1.79	0.802	no
Qwen	R2	HS	+0.83	0.000972	yes
Qwen	R2	PIQA	+0.05	1	no

Table 10 (continued)

Model	Init.	Task	$S - C$ (pp)	p_{Holm}	Signif.
Ouro	I	ARC	+1.19	1	no
Ouro	I	HS	+1.64	4.28e-14	yes
Ouro	I	PIQA	+0.00	1	no
Ouro	R1	ARC	+1.11	1	no
Ouro	R1	HS	+0.79	0.000972	yes
Ouro	R1	PIQA	-0.27	1	no
Ouro	R2	ARC	+1.11	1	no
Ouro	R2	HS	+0.13	1	no
Ouro	R2	PIQA	-0.05	1	no

Table 11: Recorded fitting-cost control. Layer ratios are C-CM cost divided by S's complete solver-endpoint cost; the matching window is $[0.95, 1.05]$. Target and Effective sum S and C-CM costs over layers, respectively. Qwen identity retains six unmatched layers, so this configuration is not fully cost-matched.

Model	Init.	Matched layers	Min ratio	Max ratio	Target (s)	Effective (s)
OPT	I	24/24	0.952	1.030	2979.2	2933.9
OPT	R1	24/24	0.952	1.025	3539.8	3500.5
OPT	R2	24/24	0.957	1.031	3510.5	3486.5
Llama	I	28/28	0.968	1.005	8960.9	8951.4
Llama	R1	28/28	0.996	1.005	10001.3	10001.8
Llama	R2	28/28	0.996	1.004	9854.1	9856.3
Qwen	I	30/36	0.946	1.059	19236.5	19389.1
Qwen	R1	36/36	0.951	1.048	14433.4	14487.5
Qwen	R2	36/36	0.954	1.043	14703.9	14561.7
Ouro	I	24/24	0.960	1.021	2644.6	2620.3
Ouro	R1	24/24	0.955	1.018	3208.0	3200.5
Ouro	R2	24/24	0.951	1.031	3146.4	3127.5

C.3 Input-weighted diagnostics and fresh confirmation

The diagnostic suite compares S against both M and C in every model–bank cell. Table 12 reports all 24 pairs and all three diagnostics. The signs of calibration weight error, diagonal input-weighted error, and output reconstruction error agree with the fresh-confirmation NLL contrast in 14, 16, and 17 of 24 pairs, respectively. These counts describe agreement between the rankings of fixed endpoints. They do not establish a selection rule that generalizes to new samples. The table also reports the remaining ten, eight, and seven sign disagreements. The associated absolute values and paired NLL intervals appear in Tables 13 and 14. The latter uses 100,000 paired document bootstrap draws, seed 20261082, with a 24-member family correction.

Table 12: All 24 endpoint pairs with three local diagnostics. Values are S minus control; positive means more error or higher NLL for S. W, D, and O denote M_W, M_{diag}, and M_{out}. Weight error is independent of the evaluation split; diagonal-input-weighted and output reconstruction errors retain calibration (cal.) and confirmation (conf.) values. Sign agreement is descriptive, not a selector validation.

Model	Init.	Pair	ΔW	ΔD_{cal}	ΔO_{cal}	ΔD_{conf}	ΔO_{conf}	ΔNLL
OPT	I	S-C	-0.00787	-0.02162	-0.02041	-0.02174	-0.02051	-0.0861
OPT	I	S-M	-0.00469	+0.01116	+0.01237	+0.01132	+0.01249	+0.0338
OPT	R1	S-C	-0.00700	-0.00508	-0.00359	-0.00507	-0.00359	-0.0152
OPT	R1	S-M	-0.00698	-0.00516	-0.00367	-0.00515	-0.00367	-0.0183
OPT	R2	S-C	-0.00700	-0.00509	-0.00355	-0.00508	-0.00355	-0.0112

Table 12 (continued)

Model	Init.	Pair	ΔW	ΔD_{cal}	ΔO_{cal}	ΔD_{conf}	ΔO_{conf}	ΔNLL
OPT	R2	S-M	-0.00697	-0.00497	-0.00360	-0.00496	-0.00360	-0.0105
Llama	I	S-C	-0.00541	-0.00639	-0.00338	-0.00642	-0.00340	-0.2985
Llama	I	S-M	-0.00303	-0.00016	+0.00061	-0.00008	+0.00068	+0.2020
Llama	R1	S-C	-0.00490	-0.00453	-0.00235	-0.00452	-0.00236	-0.0161
Llama	R1	S-M	-0.00488	-0.00451	-0.00228	-0.00451	-0.00229	-0.0206
Llama	R2	S-C	-0.00490	-0.00452	-0.00235	-0.00452	-0.00235	-0.0223
Llama	R2	S-M	-0.00487	-0.00450	-0.00231	-0.00449	-0.00231	-0.0347
Qwen	I	S-C	-0.00437	-0.00505	-0.00408	-0.00505	-0.00408	-0.0496
Qwen	I	S-M	-0.00119	+0.00124	+0.00419	+0.00130	+0.00425	+0.0692
Qwen	R1	S-C	-0.00387	-0.00371	-0.00220	-0.00371	-0.00220	-0.0141
Qwen	R1	S-M	-0.00392	-0.00375	-0.00218	-0.00374	-0.00217	-0.0045
Qwen	R2	S-C	-0.00387	-0.00372	-0.00214	-0.00372	-0.00214	+0.0006
Qwen	R2	S-M	-0.00391	-0.00376	-0.00217	-0.00376	-0.00217	+0.0006
Ouro	I	S-C	-0.00755	-0.00847	-0.00480	-0.00849	-0.00482	-0.0299
Ouro	I	S-M	-0.00532	-0.00271	-0.00020	-0.00266	-0.00015	+0.0393
Ouro	R1	S-C	-0.00703	-0.00652	-0.00321	-0.00651	-0.00321	+0.0248
Ouro	R1	S-M	-0.00702	-0.00648	-0.00315	-0.00648	-0.00316	+0.0010
Ouro	R2	S-C	-0.00703	-0.00651	-0.00317	-0.00650	-0.00318	+0.0210
Ouro	R2	S-M	-0.00701	-0.00648	-0.00316	-0.00648	-0.00318	+0.0363

Table 13: NLL effects and values are in nats/token. Absolute NLL on the diagnostic-suite 64-document confirmation cohort. FP is a single reference per model; C is I12R32. Token counts are model-tokenizer dependent.

Model	Init.	Route	Tokens	NLL
OPT	–	FP	16488	2.92837
OPT	I	S	16488	3.17283
OPT	I	C	16488	3.25897
OPT	I	M	16488	3.13900
OPT	R1	S	16488	3.01843
OPT	R1	C	16488	3.03360
OPT	R1	M	16488	3.03678
OPT	R2	S	16488	3.02717
OPT	R2	C	16488	3.03832
OPT	R2	M	16488	3.03762
Llama	–	FP	16114	2.61442
Llama	I	S	16114	2.94891
Llama	I	C	16114	3.24738
Llama	I	M	16114	2.74693
Llama	R1	S	16114	2.73145
Llama	R1	C	16114	2.74759
Llama	R1	M	16114	2.75205
Llama	R2	S	16114	2.72240
Llama	R2	C	16114	2.74467
Llama	R2	M	16114	2.75712
Qwen	–	FP	16246	2.75023
Qwen	I	S	16246	2.93188
Qwen	I	C	16246	2.98150
Qwen	I	M	16246	2.86265
Qwen	R1	S	16246	2.85836
Qwen	R1	C	16246	2.87247
Qwen	R1	M	16246	2.86285
Qwen	R2	S	16246	2.85572
Qwen	R2	C	16246	2.85509

Table 13 (continued)

Model	Init.	Route	Tokens	NLL
Qwen	R2	M	16246	2.85508
Ouro	–	FP	16709	2.72688
Ouro	I	S	16709	3.04720
Ouro	I	C	16709	3.07708
Ouro	I	M	16709	3.00792
Ouro	R1	S	16709	3.07517
Ouro	R1	C	16709	3.05035
Ouro	R1	M	16709	3.07418
Ouro	R2	S	16709	3.02987
Ouro	R2	C	16709	3.00889
Ouro	R2	M	16709	2.99354

Table 14: All fresh-confirmation S minus control NLL contrasts, on 64 documents. The 24 contrasts form one separately controlled family.

Model	Init.	Contrast	Estimate	95% CI	Adjusted CI
OPT	I	S-C	-0.0861	[-0.0992, -0.0728]	[-0.1060, -0.0655]
OPT	I	S-M	+0.0338	[+0.0237, +0.0434]	[+0.0181, +0.0484]
OPT	R1	S-C	-0.0152	[-0.0181, -0.0122]	[-0.0197, -0.0105]
OPT	R1	S-M	-0.0183	[-0.0220, -0.0146]	[-0.0242, -0.0124]
OPT	R2	S-C	-0.0112	[-0.0147, -0.0077]	[-0.0167, -0.0059]
OPT	R2	S-M	-0.0105	[-0.0145, -0.0065]	[-0.0168, -0.0041]
Llama	I	S-C	-0.2985	[-0.3231, -0.2744]	[-0.3371, -0.2622]
Llama	I	S-M	+0.2020	[+0.1798, +0.2244]	[+0.1687, +0.2370]
Llama	R1	S-C	-0.0161	[-0.0194, -0.0130]	[-0.0211, -0.0112]
Llama	R1	S-M	-0.0206	[-0.0276, -0.0138]	[-0.0314, -0.0103]
Llama	R2	S-C	-0.0223	[-0.0258, -0.0189]	[-0.0279, -0.0172]
Llama	R2	S-M	-0.0347	[-0.0400, -0.0297]	[-0.0430, -0.0271]
Qwen	I	S-C	-0.0496	[-0.0541, -0.0450]	[-0.0567, -0.0424]
Qwen	I	S-M	+0.0692	[+0.0625, +0.0760]	[+0.0587, +0.0800]
Qwen	R1	S-C	-0.0141	[-0.0171, -0.0111]	[-0.0189, -0.0092]
Qwen	R1	S-M	-0.0045	[-0.0082, -0.0006]	[-0.0103, +0.0016]
Qwen	R2	S-C	+0.0006	[-0.0022, +0.0035]	[-0.0039, +0.0051]
Qwen	R2	S-M	+0.0006	[-0.0033, +0.0045]	[-0.0056, +0.0066]
Ouro	I	S-C	-0.0299	[-0.0352, -0.0242]	[-0.0382, -0.0210]
Ouro	I	S-M	+0.0393	[+0.0287, +0.0498]	[+0.0231, +0.0557]
Ouro	R1	S-C	+0.0248	[+0.0193, +0.0302]	[+0.0159, +0.0330]
Ouro	R1	S-M	+0.0010	[-0.0080, +0.0098]	[-0.0131, +0.0145]
Ouro	R2	S-C	+0.0210	[+0.0155, +0.0264]	[+0.0126, +0.0295]
Ouro	R2	S-M	+0.0363	[+0.0283, +0.0437]	[+0.0235, +0.0477]

C.4 Error direction and Q context at matched amplitude

The direction experiment is restricted to Llama and Ouro, banks 0 and 1. It uses eight frozen physical layers per model: $\{0, 3, 7, 11, 15, 19, 23, 27\}$ for Llama and $\{0, 3, 6, 9, 13, 16, 19, 23\}$ for Ouro. Each selected layer is intervened on separately, with all other layers and V held at S. For the K error $E_d = \overline{W}_{k,d} - W_k$, the ideal constructed weight is $W_k + \|E_a\|_F E_d / \|E_d\|_F$, using the nonzero error norms of the installed endpoints. As in Equation (23), a scalar search compensates for the BF16 cast. The actual installed error must match the target norm within relative error 10^{-3} and have cosine at least 0.995 with the source direction; natural anchors reuse their installed tensors. We do not evaluate NLL for constructions that fail these checks. Q is installed from S or M at the same layer, and weights are restored to S after each condition. Let $F(q, d, a)$ be the token-weighted NLL averaged over these eight separate single-layer interventions, with Q context $q \in \{S, M\}$, K

direction $d \in \{S, M\}$, and amplitude source $a \in \{S, M\}$. Loss sums and token counts are pooled across the eight layers for each document before resampling; the experiment does not replace all eight layers jointly. The reported quantities are

$$D_{q,a} = F(q, M, a) - F(q, S, a), \tag{18}$$

$$J_a = D_{M,a} - D_{S,a}. \tag{19}$$

Thus negative D favors the M direction at the specified fixed amplitude and Q context; positive J means that changing Q from S to M increases the M-minus-S direction contrast. Table 15 gives all 16 direction contrasts and eight context interactions. It uses a separate 24-member family, 100,000 shared paired document resamples, and seed 20261082. The table is scaled by 10^3 for readability. The experiment measures the effects of local interventions at these fixed layers and endpoints; it does not reveal the optimization trajectory that produced them.

Table 15: All amplitude-matched K-direction and Q-context contrasts. In $D_{q,a}$, q denotes the Q context and a the amplitude source; J_a compares the two Q contexts. Values and intervals are in 10^{-3} nats/token; the adjusted interval controls the 24-member direction family.

Model	Init.	Contrast	Estimate	95% CI	Adjusted CI
Llama	I	$D_{M,M}$	-2.638	[-3.684, -1.588]	[-4.279, -1.028]
Llama	I	$D_{M,S}$	-1.936	[-2.896, -0.981]	[-3.415, -0.465]
Llama	I	$D_{S,M}$	-5.921	[-7.216, -4.631]	[-7.941, -3.932]
Llama	I	$D_{S,S}$	-5.712	[-6.873, -4.556]	[-7.518, -3.940]
Llama	I	J_M	+3.283	[+2.820, +3.739]	[+2.579, +3.997]
Llama	I	J_S	+3.777	[+3.306, +4.237]	[+3.045, +4.486]
Llama	R1	$D_{M,M}$	+0.449	[+0.170, +0.716]	[+0.019, +0.868]
Llama	R1	$D_{M,S}$	+0.478	[+0.252, +0.702]	[+0.129, +0.837]
Llama	R1	$D_{S,M}$	+0.508	[+0.293, +0.724]	[+0.177, +0.844]
Llama	R1	$D_{S,S}$	+0.316	[+0.016, +0.623]	[-0.155, +0.808]
Llama	R1	J_M	-0.059	[-0.313, +0.183]	[-0.455, +0.306]
Llama	R1	J_S	+0.162	[-0.117, +0.437]	[-0.278, +0.599]
Ouro	I	$D_{M,M}$	+3.678	[+3.031, +4.267]	[+2.643, +4.575]
Ouro	I	$D_{M,S}$	+3.397	[+2.814, +3.936]	[+2.471, +4.219]
Ouro	I	$D_{S,M}$	+1.668	[+1.086, +2.195]	[+0.756, +2.477]
Ouro	I	$D_{S,S}$	+1.694	[+1.112, +2.240]	[+0.778, +2.522]
Ouro	I	J_M	+2.010	[+1.736, +2.276]	[+1.577, +2.427]
Ouro	I	J_S	+1.703	[+1.392, +2.004]	[+1.215, +2.174]
Ouro	R1	$D_{M,M}$	-3.771	[-4.177, -3.374]	[-4.410, -3.151]
Ouro	R1	$D_{M,S}$	-3.289	[-3.647, -2.933]	[-3.841, -2.742]
Ouro	R1	$D_{S,M}$	-4.025	[-4.484, -3.569]	[-4.740, -3.329]
Ouro	R1	$D_{S,S}$	-3.455	[-3.894, -3.025]	[-4.148, -2.782]
Ouro	R1	J_M	+0.253	[+0.026, +0.473]	[-0.100, +0.600]
Ouro	R1	J_S	+0.166	[-0.149, +0.485]	[-0.321, +0.676]

C.5 Complete MMLU evaluation

Tables 16 and 17 report all evaluated MMLU configurations: S, M, and C in banks 0 and 1 of each model, plus one FP reference per model. Bank 2 was not evaluated on MMLU. The frozen test manifest has 14,042 questions across 57 subjects. Scoring is zero-shot with unnormalized answer log-likelihood for A/B/C/D. Micro accuracy weights questions equally; macro accuracy weights subjects equally. Of the 16 paired S–control contrasts, nine point estimates favor S. After Holm correction, six contrasts significantly favor S and three significantly favor the control; seven are not significant. The three significant losses are S versus M at bank 0 for Llama, Qwen, and Ouro. Descriptive confidence intervals use 10,000 subject-stratified paired bootstrap draws; significance uses exact two-sided McNemar tests with Holm correction over the 16 contrasts.

Table 16: Complete 0-shot MMLU on 14,042 test questions in 57 subjects, scored by raw A/B/C/D answer log-likelihood. Micro and subject-macro accuracy are percentages; Trunc. counts truncated examples. Only initializations 0 and 1 were evaluated; FP is one shared reference per model.

Model	Init.	Route	Correct	Micro (%)	Macro (%)	Trunc.
OPT	–	FP	3487	24.83	24.97	0
OPT	I	S	3328	23.70	23.79	0
OPT	I	C	3430	24.43	24.44	0
OPT	I	M	3355	23.89	24.21	0
OPT	R1	S	3479	24.78	24.82	0
OPT	R1	C	3489	24.85	24.99	0
OPT	R1	M	3499	24.92	24.67	0
Llama	–	FP	7581	53.99	55.54	0
Llama	I	S	5556	39.57	41.09	0
Llama	I	C	4551	32.41	33.48	0
Llama	I	M	6943	49.44	50.95	0
Llama	R1	S	6859	48.85	50.69	0
Llama	R1	C	6721	47.86	49.20	0
Llama	R1	M	6649	47.35	49.02	0
Qwen	–	FP	9988	71.13	73.63	0
Qwen	I	S	9132	65.03	68.08	0
Qwen	I	C	8827	62.86	65.75	0
Qwen	I	M	9430	67.16	69.67	0
Qwen	R1	S	9555	68.05	70.87	0
Qwen	R1	C	9486	67.55	70.19	0
Qwen	R1	M	9481	67.52	70.25	0
Ouro	–	FP	9588	68.28	69.09	0
Ouro	I	S	6869	48.92	49.76	0
Ouro	I	C	6636	47.26	48.79	0
Ouro	I	M	7533	53.65	55.34	0
Ouro	R1	S	7496	53.38	55.10	0
Ouro	R1	C	7322	52.14	53.79	0
Ouro	R1	M	7462	53.14	54.61	0

Table 17: All 16 MMLU comparisons. Differences and descriptive subject-stratified paired bootstrap intervals are percentage points. Statistical tests are exact paired McNemar with Holm correction over all 16 comparisons.

Model	Init.	Contrast	Diff. (pp)	95% CI	p_{Holm}
OPT	I	S-C	-0.73	[-1.46, -0.01]	0.377
OPT	I	S-M	-0.19	[-0.83, +0.45]	1
OPT	R1	S-C	-0.07	[-0.72, +0.59]	1
OPT	R1	S-M	-0.14	[-0.83, +0.54]	1
Llama	I	S-C	+7.16	[+6.27, +8.06]	2.6e-51
Llama	I	S-M	-9.88	[-10.77, -8.98]	1.37e-99
Llama	R1	S-C	+0.98	[+0.44, +1.51]	0.00224
Llama	R1	S-M	+1.50	[+0.88, +2.12]	2.22e-05
Qwen	I	S-C	+2.17	[+1.59, +2.76]	4.52e-12
Qwen	I	S-M	-2.12	[-2.76, -1.48]	1.11e-09
Qwen	R1	S-C	+0.49	[-0.02, +1.00]	0.415
Qwen	R1	S-M	+0.53	[-0.06, +1.12]	0.444
Ouro	I	S-C	+1.66	[+1.15, +2.16]	2.58e-09
Ouro	I	S-M	-4.73	[-5.58, -3.87]	3.27e-25
Ouro	R1	S-C	+1.24	[+0.69, +1.79]	9.06e-05
Ouro	R1	S-M	+0.24	[-0.46, +0.94]	1

C.6 FROZEN-ENDPOINT BRANCH DECOMPOSITION

The final intervention suite holds S and C endpoints fixed and substitutes their Q, K, and V branches across all physical layers. For $T \subseteq \{q, k, v\}$, let $f(T)$ denote token-weighted NLL when the branches in T come from C and the remaining branches from S. Then $f(\varnothing) = f(S)$ and $f(\{q, k, v\}) = f(C)$. Define

$$a(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} f(U), \qquad T \neq \varnothing, \tag{20}$$

$$\Delta = f(C) - f(S), \qquad A = a(\{q\}) + a(\{k\}) + a(\{v\}), \qquad H = \Delta - A. \tag{21}$$

The sign convention here differs from the S-minus-control tables above: positive Δ favors S. All 12 cells are shown in Tables 18–20. The corrected analysis uses 100,000 paired token-weighted document bootstrap draws (seed 20261091) with shared indices across states, banks, and models. The 36 $\Delta/A/H$ quantities and 84 nonempty branch terms form separate Bonferroni-adjusted families. These corrections do not jointly control the error rate across the two families. The confirmation split has 64 documents; the separate 16-document discovery split is not pooled into the confirmation intervals.

Eight adjusted Δ intervals favor S, two favor C (Ouro banks 1 and 2), and two contain zero (OPT bank 2 and Qwen bank 2). Only three adjusted H intervals exclude zero: positive for OPT bank 0 and Ouro bank 0, and negative for Llama bank 0. No cell provides statistical support for opposite signs of A and Δ. An interval containing zero does not establish additivity or equivalence.

The endpoint audit found V weights bitwise equal between S and C at every physical layer in all 12 cells. Consequently, the 48 V-containing branch terms and all 12 V Shapley allocations are exactly zero in the frozen results. Tables 20 and 21 omit those identical zero rows; their original full CSVs remain in the portable source bundle. V is unchanged in these endpoints; this does not establish insensitivity to V errors.

Table 18: NLL effects and values are in nats/token. Absolute NLL in the causal-suite confirmation split, with ordinary 95% bootstrap intervals. All rows use 64 documents. FP is shared across banks; S and C correspond to the all-S and all-C branch states.

Model	Init.	Route	NLL	95% CI
OPT	I	C	3.11881	[+3.01288, +3.23281]
OPT	–	FP	2.82379	[+2.71787, +2.93845]
OPT	I	S	3.03704	[+2.93230, +3.15086]
OPT	R1	C	2.92654	[+2.82205, +3.04035]
OPT	R1	S	2.90993	[+2.80558, +3.02354]
OPT	R2	C	2.93294	[+2.82858, +3.04618]
OPT	R2	S	2.92209	[+2.81878, +3.03454]
Llama	I	C	3.06843	[+2.95334, +3.19051]
Llama	–	FP	2.41584	[+2.23541, +2.57818]
Llama	I	S	2.76985	[+2.65595, +2.88703]
Llama	R1	C	2.58110	[+2.44072, +2.71770]
Llama	R1	S	2.55809	[+2.41112, +2.69858]
Llama	R2	C	2.56922	[+2.42972, +2.70462]
Llama	R2	S	2.54663	[+2.39994, +2.68613]
Qwen	I	C	2.78929	[+2.67577, +2.90611]
Qwen	–	FP	2.56609	[+2.45830, +2.67772]
Qwen	I	S	2.74532	[+2.63265, +2.86157]
Qwen	R1	C	2.68955	[+2.57921, +2.80331]
Qwen	R1	S	2.67894	[+2.56919, +2.79190]
Qwen	R2	C	2.67835	[+2.56698, +2.79285]
Qwen	R2	S	2.67424	[+2.56409, +2.78803]
Ouro	I	C	2.93454	[+2.81365, +3.05983]
Ouro	–	FP	2.58072	[+2.46253, +2.70335]
Ouro	I	S	2.89487	[+2.77439, +3.01960]

Table 18 (continued)

Model	Init.	Route	NLL	95% CI
Ouro	R1	C	2.88171	[+2.76063, +3.00722]
Ouro	R1	S	2.91426	[+2.79347, +3.03911]
Ouro	R2	C	2.86698	[+2.74557, +2.99316]
Ouro	R2	S	2.88516	[+2.76223, +3.01284]

Table 19: NLL effects and values are in nats/token. Frozen-endpoint branch decomposition on the new 64-document confirmation cohort. $\Delta = f(C) - f(S)$, A sums single-branch replacements, and $H = \Delta - A$. Positive Δ favors S. Adjusted intervals use the joint 36-member family of Δ, A, H.

Model	Init.	Term	Estimate	95% CI	Adjusted CI
OPT	I	Δ	+0.08177	[+0.06036, +0.10643]	[+0.04862, +0.12440]
OPT	I	A	+0.07572	[+0.05325, +0.10211]	[+0.04113, +0.12148]
OPT	I	H	+0.00606	[+0.00255, +0.00932]	[+0.00022, +0.01117]
OPT	R1	Δ	+0.01660	[+0.01213, +0.02095]	[+0.00929, +0.02379]
OPT	R1	A	+0.01632	[+0.01141, +0.02110]	[+0.00824, +0.02416]
OPT	R1	H	+0.00028	[-0.00134, +0.00197]	[-0.00234, +0.00304]
OPT	R2	Δ	+0.01085	[+0.00403, +0.01718]	[-0.00047, +0.02107]
OPT	R2	A	+0.00728	[-0.00110, +0.01486]	[-0.00703, +0.01903]
OPT	R2	H	+0.00357	[+0.00121, +0.00611]	[-0.00020, +0.00791]
Llama	I	Δ	+0.29859	[+0.26397, +0.33949]	[+0.24661, +0.37077]
Llama	I	A	+0.31845	[+0.27823, +0.36443]	[+0.25839, +0.39995]
Llama	I	H	-0.01987	[-0.03095, -0.00800]	[-0.03790, -0.00044]
Llama	R1	Δ	+0.02300	[+0.01385, +0.03509]	[+0.00959, +0.04511]
Llama	R1	A	+0.02496	[+0.01557, +0.03747]	[+0.01139, +0.04780]
Llama	R1	H	-0.00196	[-0.00368, -0.00024]	[-0.00482, +0.00081]
Llama	R2	Δ	+0.02260	[+0.01287, +0.03497]	[+0.00896, +0.04512]
Llama	R2	A	+0.02212	[+0.01243, +0.03408]	[+0.00846, +0.04351]
Llama	R2	H	+0.00047	[-0.00117, +0.00208]	[-0.00233, +0.00319]
Qwen	I	Δ	+0.04397	[+0.03601, +0.05206]	[+0.03122, +0.05720]
Qwen	I	A	+0.04078	[+0.03347, +0.04838]	[+0.02904, +0.05332]
Qwen	I	H	+0.00319	[+0.00013, +0.00612]	[-0.00191, +0.00798]
Qwen	R1	Δ	+0.01061	[+0.00602, +0.01555]	[+0.00317, +0.01908]
Qwen	R1	A	+0.00979	[+0.00510, +0.01491]	[+0.00226, +0.01853]
Qwen	R1	H	+0.00082	[-0.00089, +0.00252]	[-0.00199, +0.00372]
Qwen	R2	Δ	+0.00411	[-0.00119, +0.00957]	[-0.00460, +0.01311]
Qwen	R2	A	+0.00455	[-0.00136, +0.01052]	[-0.00511, +0.01435]
Qwen	R2	H	-0.00044	[-0.00216, +0.00125]	[-0.00328, +0.00234]
Ouro	I	Δ	+0.03968	[+0.03121, +0.04814]	[+0.02596, +0.05354]
Ouro	I	A	+0.03111	[+0.02265, +0.03950]	[+0.01754, +0.04472]
Ouro	I	H	+0.00856	[+0.00469, +0.01259]	[+0.00232, +0.01550]
Ouro	R1	Δ	-0.03254	[-0.03940, -0.02543]	[-0.04366, -0.02099]
Ouro	R1	A	-0.03235	[-0.03993, -0.02449]	[-0.04473, -0.01935]
Ouro	R1	H	-0.00019	[-0.00363, +0.00265]	[-0.00624, +0.00410]
Ouro	R2	Δ	-0.01817	[-0.02630, -0.01033]	[-0.03142, -0.00555]
Ouro	R2	A	-0.01593	[-0.02460, -0.00752]	[-0.03018, -0.00222]
Ouro	R2	H	-0.00225	[-0.00477, +0.00015]	[-0.00648, +0.00154]

Table 20: NLL effects and values are in nats/token. All potentially nonzero branch terms. Every one of the 48 terms containing V is exactly zero in the frozen tables (point estimate and both intervals); those identical rows are compressed here. Adjusted intervals retain the original 84-member family, including the zero V terms.

Model	Init.	Term	Estimate	95% CI	Adjusted CI
OPT	I	$a(k)$	+0.05467	[+0.03542, +0.07670]	[+0.02262, +0.09663]
OPT	I	$a(q)$	+0.02104	[+0.01576, +0.02691]	[+0.01204, +0.03166]
OPT	I	$a(q,k)$	+0.00606	[+0.00255, +0.00932]	[-0.00016, +0.01158]
OPT	R1	$a(k)$	+0.01745	[+0.01366, +0.02121]	[+0.01086, +0.02411]
OPT	R1	$a(q)$	-0.00114	[-0.00364, +0.00128]	[-0.00557, +0.00307]
OPT	R1	$a(q,k)$	+0.00028	[-0.00134, +0.00197]	[-0.00255, +0.00327]
OPT	R2	$a(k)$	+0.00620	[+0.00035, +0.01172]	[-0.00410, +0.01550]
OPT	R2	$a(q)$	+0.00109	[-0.00280, +0.00505]	[-0.00551, +0.00810]
OPT	R2	$a(q,k)$	+0.00357	[+0.00121, +0.00611]	[-0.00038, +0.00837]
Llama	I	$a(k)$	+0.27814	[+0.24081, +0.32146]	[+0.21990, +0.36143]
Llama	I	$a(q)$	+0.04031	[+0.03094, +0.05138]	[+0.02498, +0.06114]
Llama	I	$a(q,k)$	-0.01987	[-0.03095, -0.00800]	[-0.03950, +0.00090]
Llama	R1	$a(k)$	+0.02134	[+0.01578, +0.02834]	[+0.01320, +0.03512]
Llama	R1	$a(q)$	+0.00362	[-0.00238, +0.01051]	[-0.00651, +0.01661]
Llama	R1	$a(q,k)$	-0.00196	[-0.00368, -0.00024]	[-0.00504, +0.00102]
Llama	R2	$a(k)$	+0.01646	[+0.01252, +0.02129]	[+0.01055, +0.02578]
Llama	R2	$a(q)$	+0.00567	[-0.00162, +0.01390]	[-0.00627, +0.02166]
Llama	R2	$a(q,k)$	+0.00047	[-0.00117, +0.00208]	[-0.00261, +0.00339]
Qwen	I	$a(k)$	+0.03070	[+0.02708, +0.03471]	[+0.02457, +0.03802]
Qwen	I	$a(q)$	+0.01008	[+0.00431, +0.01587]	[+0.00001, +0.02050]
Qwen	I	$a(q,k)$	+0.00319	[+0.00013, +0.00612]	[-0.00235, +0.00844]
Qwen	R1	$a(k)$	+0.00696	[+0.00492, +0.00896]	[+0.00335, +0.01052]
Qwen	R1	$a(q)$	+0.00283	[-0.00138, +0.00750]	[-0.00417, +0.01137]
Qwen	R1	$a(q,k)$	+0.00082	[-0.00089, +0.00252]	[-0.00230, +0.00391]
Qwen	R2	$a(k)$	+0.00433	[+0.00249, +0.00623]	[+0.00101, +0.00774]
Qwen	R2	$a(q)$	+0.00023	[-0.00499, +0.00558]	[-0.00908, +0.00964]
Qwen	R2	$a(q,k)$	-0.00044	[-0.00216, +0.00125]	[-0.00340, +0.00253]
Ouro	I	$a(k)$	+0.02467	[+0.01874, +0.03088]	[+0.01449, +0.03585]
Ouro	I	$a(q)$	+0.00644	[+0.00138, +0.01106]	[-0.00266, +0.01441]
Ouro	I	$a(q,k)$	+0.00856	[+0.00469, +0.01259]	[+0.00183, +0.01613]
Ouro	R1	$a(k)$	+0.00108	[-0.00464, +0.00722]	[-0.00848, +0.01231]
Ouro	R1	$a(q)$	-0.03343	[-0.03974, -0.02683]	[-0.04475, -0.02216]
Ouro	R1	$a(q,k)$	-0.00019	[-0.00363, +0.00265]	[-0.00670, +0.00444]
Ouro	R2	$a(k)$	-0.00485	[-0.01074, +0.00076]	[-0.01547, +0.00478]
Ouro	R2	$a(q)$	-0.01108	[-0.01841, -0.00399]	[-0.02407, +0.00097]
Ouro	R2	$a(q,k)$	-0.00225	[-0.00477, +0.00015]	[-0.00681, +0.00180]

Table 21: NLL effects and values are in nats/token. Descriptive Shapley allocations of the frozen endpoint difference. V is exactly zero in all 12 cells and is omitted. Ordinary 95% intervals are not used for a main significance claim.

Model	Init.	Branch	Allocation	95% CI
OPT	I	k	+0.05770	[+0.03882, +0.07896]
OPT	I	q	+0.02407	[+0.01950, +0.02897]
OPT	R1	k	+0.01760	[+0.01413, +0.02103]
OPT	R1	q	-0.00099	[-0.00354, +0.00150]
OPT	R2	k	+0.00798	[+0.00285, +0.01294]
OPT	R2	q	+0.00287	[-0.00055, +0.00643]
Llama	I	k	+0.26821	[+0.23385, +0.30865]
Llama	I	q	+0.03038	[+0.02030, +0.04218]

Table 21 (continued)

Model	Init.	Branch	Allocation	95% CI
Llama	R1	k	+0.02036	[+0.01521, +0.02697]
Llama	R1	q	+0.00265	[-0.00353, +0.00952]
Llama	R2	k	+0.01669	[+0.01268, +0.02174]
Llama	R2	q	+0.00590	[-0.00140, +0.01430]
Qwen	I	k	+0.03230	[+0.02855, +0.03631]
Qwen	I	q	+0.01168	[+0.00555, +0.01777]
Qwen	R1	k	+0.00737	[+0.00543, +0.00935]
Qwen	R1	q	+0.00324	[-0.00096, +0.00784]
Qwen	R2	k	+0.00411	[+0.00263, +0.00565]
Qwen	R2	q	+0.00001	[-0.00501, +0.00520]
Ouro	I	k	+0.02895	[+0.02326, +0.03483]
Ouro	I	q	+0.01072	[+0.00539, +0.01580]
Ouro	R1	k	+0.00098	[-0.00389, +0.00602]
Ouro	R1	q	-0.03353	[-0.04030, -0.02668]
Ouro	R2	k	-0.00597	[-0.01169, -0.00056]
Ouro	R2	q	-0.01220	[-0.01941, -0.00532]

C.7 Depth paths and background dependence

Physical layers are partitioned into four consecutive groups $G_j = \{\lfloor jL/4 \rfloor, \ldots, \lfloor (j+1)L/4 \rfloor - 1\}$, where $j = 0, 1, 2, 3$ and L is the number of physical layers. The E path replaces S by C in order G_0, G_1, G_2, G_3; the R path uses the reverse order. Each path marginal is the NLL change upon the indicated group replacement, conditional on the groups already replaced along that path. The four marginals sum to Δ in each path by telescoping. Table 22 displays all 96 marginals with both ordinary and 96-family adjusted intervals. Table 23 retains all 48 E-minus-R differences with exploratory, unadjusted 95% intervals. These comparisons show how a group's effect depends on which other groups have already been replaced. They do not assign each layer a unique contribution that is independent of replacement order. For Ouro, replacing one physical layer affects all its recurrent applications; the intervention does not isolate recurrence as a cause.

Table 22: NLL effects and values are in nats/token. All 96 depth-path marginals. E replaces consecutive physical-layer groups from early to late; R reverses the order. Groups G0–G3 are indexed by depth, not by step in R. Positive marginals increase NLL while replacing S by C. Adjusted intervals use the 96-member family.

Model	Init.	Path	Group	Marginal	95% CI	Adjusted CI
OPT	I	E	G0	+0.02168	[+0.01744, +0.02679]	[+0.01508, +0.03158]
OPT	I	E	G1	+0.04784	[+0.03993, +0.05635]	[+0.03476, +0.06399]
OPT	I	E	G2	+0.01797	[+0.01462, +0.02102]	[+0.01185, +0.02318]
OPT	I	E	G3	-0.00572	[-0.02202, +0.01176]	[-0.03382, +0.02674]
OPT	I	R	G0	+0.03445	[+0.02783, +0.04226]	[+0.02388, +0.04960]
OPT	I	R	G1	+0.03583	[+0.03093, +0.04047]	[+0.02718, +0.04412]
OPT	I	R	G2	+0.01414	[+0.01154, +0.01665]	[+0.00958, +0.01859]
OPT	I	R	G3	-0.00265	[-0.01857, +0.01485]	[-0.02980, +0.03009]
OPT	R1	E	G0	+0.00137	[+0.00015, +0.00255]	[-0.00076, +0.00342]
OPT	R1	E	G1	+0.00142	[-0.00058, +0.00349]	[-0.00201, +0.00520]
OPT	R1	E	G2	+0.00369	[+0.00183, +0.00555]	[+0.00044, +0.00715]
OPT	R1	E	G3	+0.01012	[+0.00709, +0.01318]	[+0.00478, +0.01555]
OPT	R1	R	G0	+0.00118	[-0.00002, +0.00233]	[-0.00095, +0.00318]
OPT	R1	R	G1	+0.00178	[-0.00035, +0.00399]	[-0.00188, +0.00565]
OPT	R1	R	G2	+0.00389	[+0.00193, +0.00586]	[+0.00030, +0.00749]
OPT	R1	R	G3	+0.00976	[+0.00670, +0.01277]	[+0.00434, +0.01519]
OPT	R2	E	G0	+0.00219	[+0.00073, +0.00379]	[-0.00020, +0.00516]
OPT	R2	E	G1	+0.00267	[+0.00070, +0.00469]	[-0.00085, +0.00624]

Table 22 (continued)

Model	Init.	Path	Group	Marginal	95% CI	Adjusted CI
OPT	R2	E	G2	+0.00202	[-0.00043, +0.00435]	[-0.00215, +0.00600]
OPT	R2	E	G3	+0.00397	[-0.00099, +0.00876]	[-0.00513, +0.01213]
OPT	R2	R	G0	+0.00365	[+0.00213, +0.00534]	[+0.00115, +0.00674]
OPT	R2	R	G1	+0.00144	[-0.00110, +0.00393]	[-0.00309, +0.00583]
OPT	R2	R	G2	+0.00303	[+0.00047, +0.00544]	[-0.00147, +0.00717]
OPT	R2	R	G3	+0.00273	[-0.00271, +0.00783]	[-0.00748, +0.01121]
Llama	I	E	G0	+0.23247	[+0.20518, +0.26239]	[+0.18668, +0.28735]
Llama	I	E	G1	+0.04446	[+0.03226, +0.05893]	[+0.02514, +0.07229]
Llama	I	E	G2	+0.01622	[+0.01395, +0.01885]	[+0.01245, +0.02114]
Llama	I	E	G3	+0.00543	[+0.00362, +0.00733]	[+0.00223, +0.00886]
Llama	I	R	G0	+0.22130	[+0.19620, +0.25043]	[+0.17984, +0.27609]
Llama	I	R	G1	+0.05161	[+0.03871, +0.06932]	[+0.03365, +0.08767]
Llama	I	R	G2	+0.01872	[+0.01391, +0.02505]	[+0.01146, +0.03147]
Llama	I	R	G3	+0.00695	[+0.00521, +0.00872]	[+0.00378, +0.01024]
Llama	R1	E	G0	+0.00076	[-0.00340, +0.00546]	[-0.00635, +0.00963]
Llama	R1	E	G1	+0.00957	[+0.00539, +0.01396]	[+0.00239, +0.01764]
Llama	R1	E	G2	+0.01281	[+0.00731, +0.02078]	[+0.00521, +0.02920]
Llama	R1	E	G3	-0.00014	[-0.00262, +0.00190]	[-0.00499, +0.00308]
Llama	R1	R	G0	+0.00030	[-0.00366, +0.00452]	[-0.00680, +0.00803]
Llama	R1	R	G1	+0.00979	[+0.00447, +0.01604]	[+0.00071, +0.02195]
Llama	R1	R	G2	+0.01100	[+0.00718, +0.01596]	[+0.00530, +0.02085]
Llama	R1	R	G3	+0.00191	[+0.00064, +0.00319]	[-0.00042, +0.00424]
Llama	R2	E	G0	-0.00052	[-0.00375, +0.00359]	[-0.00571, +0.00784]
Llama	R2	E	G1	+0.00960	[+0.00656, +0.01278]	[+0.00442, +0.01531]
Llama	R2	E	G2	+0.00763	[+0.00373, +0.01207]	[+0.00135, +0.01614]
Llama	R2	E	G3	+0.00588	[+0.00161, +0.01185]	[-0.00004, +0.01819]
Llama	R2	R	G0	-0.00041	[-0.00450, +0.00470]	[-0.00724, +0.01026]
Llama	R2	R	G1	+0.00990	[+0.00682, +0.01311]	[+0.00464, +0.01579]
Llama	R2	R	G2	+0.00807	[+0.00288, +0.01485]	[+0.00024, +0.02220]
Llama	R2	R	G3	+0.00503	[+0.00229, +0.00816]	[+0.00062, +0.01120]
Qwen	I	E	G0	+0.00676	[+0.00392, +0.00960]	[+0.00176, +0.01187]
Qwen	I	E	G1	+0.02423	[+0.02004, +0.02865]	[+0.01696, +0.03243]
Qwen	I	E	G2	-0.00172	[-0.00646, +0.00314]	[-0.01002, +0.00683]
Qwen	I	E	G3	+0.01470	[+0.01143, +0.01797]	[+0.00861, +0.02073]
Qwen	I	R	G0	+0.00682	[+0.00374, +0.00999]	[+0.00141, +0.01265]
Qwen	I	R	G1	+0.01970	[+0.01502, +0.02473]	[+0.01171, +0.02888]
Qwen	I	R	G2	+0.00192	[-0.00317, +0.00704]	[-0.00704, +0.01092]
Qwen	I	R	G3	+0.01553	[+0.01260, +0.01872]	[+0.01050, +0.02163]
Qwen	R1	E	G0	+0.00118	[-0.00100, +0.00337]	[-0.00280, +0.00519]
Qwen	R1	E	G1	+0.00420	[+0.00119, +0.00716]	[-0.00129, +0.00950]
Qwen	R1	E	G2	-0.00149	[-0.00425, +0.00143]	[-0.00626, +0.00388]
Qwen	R1	E	G3	+0.00672	[+0.00421, +0.00925]	[+0.00211, +0.01137]
Qwen	R1	R	G0	+0.00146	[-0.00052, +0.00347]	[-0.00194, +0.00504]
Qwen	R1	R	G1	+0.00416	[+0.00146, +0.00689]	[-0.00065, +0.00908]
Qwen	R1	R	G2	-0.00123	[-0.00405, +0.00174]	[-0.00608, +0.00407]
Qwen	R1	R	G3	+0.00622	[+0.00382, +0.00869]	[+0.00192, +0.01072]
Qwen	R2	E	G0	+0.00040	[-0.00136, +0.00213]	[-0.00286, +0.00354]
Qwen	R2	E	G1	+0.00332	[+0.00067, +0.00579]	[-0.00160, +0.00752]
Qwen	R2	E	G2	+0.00148	[-0.00243, +0.00550]	[-0.00515, +0.00848]
Qwen	R2	E	G3	-0.00108	[-0.00370, +0.00165]	[-0.00563, +0.00391]
Qwen	R2	R	G0	+0.00016	[-0.00186, +0.00206]	[-0.00356, +0.00350]
Qwen	R2	R	G1	+0.00267	[+0.00063, +0.00451]	[-0.00124, +0.00582]
Qwen	R2	R	G2	+0.00109	[-0.00289, +0.00506]	[-0.00584, +0.00796]
Qwen	R2	R	G3	+0.00018	[-0.00229, +0.00289]	[-0.00406, +0.00522]
Ouro	I	E	G0	-0.02013	[-0.02663, -0.01376]	[-0.03187, -0.00918]
Ouro	I	E	G1	+0.00720	[+0.00246, +0.01226]	[-0.00097, +0.01619]
Ouro	I	E	G2	+0.04308	[+0.03666, +0.04918]	[+0.03155, +0.05411]

Table 22 (continued)

Model	Init.	Path	Group	Marginal	95% CI	Adjusted CI
Ouro	I	E	G3	+0.00951	[+0.00571, +0.01370]	[+0.00292, +0.01726]
Ouro	I	R	G0	-0.03148	[-0.03851, -0.02471]	[-0.04454, -0.01968]
Ouro	I	R	G1	+0.01211	[+0.00555, +0.01875]	[+0.00069, +0.02375]
Ouro	I	R	G2	+0.04381	[+0.03726, +0.05007]	[+0.03231, +0.05519]
Ouro	I	R	G3	+0.01523	[+0.01133, +0.01986]	[+0.00884, +0.02405]
Ouro	R1	E	G0	-0.00477	[-0.00954, +0.00039]	[-0.01291, +0.00472]
Ouro	R1	E	G1	-0.02612	[-0.03196, -0.02063]	[-0.03709, -0.01655]
Ouro	R1	E	G2	+0.00071	[-0.00180, +0.00338]	[-0.00361, +0.00568]
Ouro	R1	E	G3	-0.00237	[-0.00493, +0.00012]	[-0.00700, +0.00191]
Ouro	R1	R	G0	-0.00729	[-0.01142, -0.00287]	[-0.01440, +0.00074]
Ouro	R1	R	G1	-0.02586	[-0.03188, -0.01995]	[-0.03676, -0.01546]
Ouro	R1	R	G2	+0.00167	[-0.00094, +0.00443]	[-0.00279, +0.00665]
Ouro	R1	R	G3	-0.00107	[-0.00351, +0.00129]	[-0.00563, +0.00327]
Ouro	R2	E	G0	+0.00552	[+0.00062, +0.01047]	[-0.00302, +0.01441]
Ouro	R2	E	G1	-0.02130	[-0.02578, -0.01658]	[-0.02907, -0.01289]
Ouro	R2	E	G2	-0.00381	[-0.00669, -0.00093]	[-0.00895, +0.00137]
Ouro	R2	E	G3	+0.00142	[-0.00083, +0.00367]	[-0.00259, +0.00549]
Ouro	R2	R	G0	+0.00264	[-0.00225, +0.00722]	[-0.00619, +0.01046]
Ouro	R2	R	G1	-0.01980	[-0.02403, -0.01536]	[-0.02693, -0.01186]
Ouro	R2	R	G2	-0.00325	[-0.00661, -0.00011]	[-0.00932, +0.00218]
Ouro	R2	R	G3	+0.00225	[-0.00065, +0.00482]	[-0.00348, +0.00667]

Table 23: NLL effects and values are in nats/token. All 48 E-minus-R marginal differences. These are exploratory point estimates with ordinary 95% intervals, without family correction; they describe dependence on the replacement background.

Model	Init.	Group	E-R	95% CI
OPT	I	G0	-0.01277	[-0.01648, -0.00902]
OPT	I	G1	+0.01202	[+0.00708, +0.01834]
OPT	I	G2	+0.00383	[+0.00008, +0.00723]
OPT	I	G3	-0.00307	[-0.00511, -0.00100]
OPT	R1	G0	+0.00019	[-0.00089, +0.00126]
OPT	R1	G1	-0.00036	[-0.00149, +0.00073]
OPT	R1	G2	-0.00020	[-0.00135, +0.00102]
OPT	R1	G3	+0.00037	[-0.00094, +0.00166]
OPT	R2	G0	-0.00146	[-0.00311, +0.00002]
OPT	R2	G1	+0.00123	[-0.00026, +0.00281]
OPT	R2	G2	-0.00101	[-0.00259, +0.00072]
OPT	R2	G3	+0.00125	[-0.00058, +0.00297]
Llama	I	G0	+0.01118	[+0.00023, +0.02434]
Llama	I	G1	-0.00716	[-0.01656, +0.00173]
Llama	I	G2	-0.00249	[-0.00802, +0.00168]
Llama	I	G3	-0.00152	[-0.00324, +0.00031]
Llama	R1	G0	+0.00046	[-0.00133, +0.00235]
Llama	R1	G1	-0.00022	[-0.00296, +0.00213]
Llama	R1	G2	+0.00181	[-0.00049, +0.00519]
Llama	R1	G3	-0.00204	[-0.00418, -0.00029]
Llama	R2	G0	-0.00011	[-0.00211, +0.00186]
Llama	R2	G1	-0.00030	[-0.00218, +0.00156]
Llama	R2	G2	-0.00044	[-0.00365, +0.00176]
Llama	R2	G3	+0.00085	[-0.00163, +0.00473]
Qwen	I	G0	-0.00006	[-0.00226, +0.00209]
Qwen	I	G1	+0.00453	[+0.00154, +0.00763]
Qwen	I	G2	-0.00365	[-0.00648, -0.00077]

Table 23 (continued)

Model	Init.	Group	E-R	95% CI
Qwen	I	G3	-0.00083	[-0.00250, +0.00071]
Qwen	R1	G0	-0.00028	[-0.00180, +0.00122]
Qwen	R1	G1	+0.00004	[-0.00168, +0.00173]
Qwen	R1	G2	-0.00026	[-0.00179, +0.00131]
Qwen	R1	G3	+0.00051	[-0.00089, +0.00189]
Qwen	R2	G0	+0.00024	[-0.00124, +0.00179]
Qwen	R2	G1	+0.00064	[-0.00123, +0.00246]
Qwen	R2	G2	+0.00038	[-0.00126, +0.00196]
Qwen	R2	G3	-0.00127	[-0.00262, -0.00001]
Ouro	I	G0	+0.01135	[+0.00756, +0.01508]
Ouro	I	G1	-0.00490	[-0.00838, -0.00150]
Ouro	I	G2	-0.00073	[-0.00335, +0.00194]
Ouro	I	G3	-0.00572	[-0.00867, -0.00264]
Ouro	R1	G0	+0.00252	[-0.00019, +0.00525]
Ouro	R1	G1	-0.00027	[-0.00282, +0.00215]
Ouro	R1	G2	-0.00096	[-0.00277, +0.00087]
Ouro	R1	G3	-0.00130	[-0.00324, +0.00062]
Ouro	R2	G0	+0.00289	[+0.00065, +0.00523]
Ouro	R2	G1	-0.00151	[-0.00377, +0.00062]
Ouro	R2	G2	-0.00056	[-0.00229, +0.00127]
Ouro	R2	G3	-0.00082	[-0.00239, +0.00112]

D SHARING DURATION AND EVALUATION SETTING

The schedule sweep tests whether the benefit observed at one release time persists when we shorten or lengthen the shared phase. It uses OPT, Llama, Qwen, and Ouro, two random orthogonal initialization configurations (start identifiers 100 and 101), and release times $t \in \{4, 12, 32\}$. Each configuration allocates 44 rounds per branch in total. In S(t), Q and K share a transform for the first t rounds, while V is fitted independently. The resulting QK transform is copied into separate Q and K variables, after which all three branches are fitted independently for $44 - t$ rounds. The control C(t) uses independent transforms in both phases and the same phase-entry reinitialization of the alternating solver at round t. Thus C(t) is a restart-matched control, not uninterrupted I44. The comparison matches allocated rounds, not measured fitting time.

The sweep is a separate development experiment from the fixed-release confirmation study in the main comparison. It uses two additional random starts and 1,024 C4 documents, so its point estimates are not pooled with the 128-document confirmation estimates. C4 documents are truncated to at most 2,048 tokens; WikiText-2 uses chunks of at most 512 tokens. Each S/C pair uses identical items and scored tokens within its evaluation setting. The same fitted endpoints are scored in both settings without refitting. Cross-setting ranking differences do not isolate corpus effects from context length and segmentation. Table 24 reports the scored token counts and full-precision references. Ouro uses four recurrent steps and NLL computed from the final-step logits, with each physical weight modified once and reused across steps.

Table 24: Schedule-sweep evaluation coverage and full-precision reference NLL. Each model uses the same evaluation tokens for every schedule. C4 uses 1,024 documents capped at 2,048 tokens; WikiText-2 uses contiguous chunks of at most 512 tokens.

	C4			WikiText-2	
Model	Scored tokens	FP NLL	Chunks	Scored tokens	FP NLL
OPT	418,788	2.75629	556	284,058	2.98217
Llama	405,280	2.39152	565	288,372	2.32399
Qwen	411,340	2.56710	584	298,355	2.33053
Ouro	426,926	2.55342	593	302,674	2.44784

Table 25 gives all 24 paired configurations, corresponding to 48 quantized endpoints. S has lower point-estimate NLL in 23 of 24 C4 comparisons and 20 of 24 WikiText-2 comparisons. These counts summarize the tested configurations; they are not independent replications or a significance test. OPT and Llama favor S throughout the grid. Ouro at start 101 and $t = 4$ favors C on both corpora, whereas longer shared prefixes favor S. Qwen at start 100 favors S on C4 but C on WikiText-2 at every evaluated release time. These examples show that the ranking can depend on both sharing duration and the evaluation setting. Extending the shared prefix changes the observed Ouro ranking to favor S, while Qwen's ranking still differs between settings.

The portable supplement contains the per-document and per-chunk summed losses and scored-token counts. Every NLL in these tables is recomputed as total loss divided by total scored tokens, after checking that the same evaluation items and token counts are used within each S–C pair. The table and figure report point estimates; they do not supply additional confidence intervals or a validated rule for selecting the release time on a new model.

Table 25: Schedule sweep: token-averaged NLL (↓). S and C share the same initialization and split time. $\Delta = \text{NLL(C)} - \text{NLL(S)}$, so positive differences favor the shared prefix. Entries are point estimates, not significance claims.

			C4			WikiText-2		
Model	Start ID	t	S	C	Δ	S	C	Δ
OPT	100	4	2.86696	2.87849	+0.01153	3.14080	3.15968	+0.01888
OPT	100	12	2.86635	2.87849	+0.01215	3.14126	3.15970	+0.01844
OPT	100	32	2.86742	2.87849	+0.01107	3.14590	3.15969	+0.01379
OPT	101	4	2.85908	2.87603	+0.01695	3.12539	3.14025	+0.01486
OPT	101	12	2.85719	2.87603	+0.01884	3.12470	3.14025	+0.01555
OPT	101	32	2.85732	2.87613	+0.01881	3.12271	3.14021	+0.01750
Llama	100	4	2.54463	2.55727	+0.01264	2.44924	2.47011	+0.02087
Llama	100	12	2.54281	2.55729	+0.01448	2.44808	2.47011	+0.02203
Llama	100	32	2.54341	2.55726	+0.01385	2.44656	2.47018	+0.02362
Llama	101	4	2.53421	2.55602	+0.02181	2.44713	2.46633	+0.01920
Llama	101	12	2.53051	2.55627	+0.02576	2.44635	2.46632	+0.01997
Llama	101	32	2.53295	2.55615	+0.02321	2.44745	2.46631	+0.01886
Qwen	100	4	2.68901	2.69405	+0.00504	2.46645	2.45534	-0.01111
Qwen	100	12	2.69084	2.69457	+0.00373	2.46115	2.45601	-0.00514
Qwen	100	32	2.69017	2.69444	+0.00427	2.46209	2.45591	-0.00618
Qwen	101	4	2.69279	2.69297	+0.00018	2.46560	2.46979	+0.00420
Qwen	101	12	2.68520	2.69305	+0.00785	2.45451	2.47005	+0.01554
Qwen	101	32	2.68763	2.69295	+0.00532	2.46159	2.46981	+0.00822
Ouro	100	4	2.87662	2.89082	+0.01420	2.80516	2.84179	+0.03663
Ouro	100	12	2.86327	2.89082	+0.02755	2.79308	2.84179	+0.04871
Ouro	100	32	2.87715	2.89086	+0.01371	2.81599	2.84171	+0.02572
Ouro	101	4	2.89102	2.88094	-0.01008	2.84858	2.84184	-0.00674
Ouro	101	12	2.87273	2.88101	+0.00828	2.82839	2.84173	+0.01334
Ouro	101	32	2.87616	2.88092	+0.00476	2.82731	2.84182	+0.01451

E Exploratory Applications

We tested whether selecting among fitted endpoints or transferring transforms improves quantization beyond the controlled comparisons in the main text.

Depth-wise endpoint selection. For Llama-3.2-3B and Qwen3-4B-Base, four contiguous, equal-sized layer groups each take all three QKV operators from $S(12)$ or $C(12)$, giving 16 combinations. Here $S(t)$ and $C(t)$ use a transition after t rounds within a total allocation of 44 rounds; all endpoints use initialization configuration 100. Selection minimizes NLL on 256 calibration documents, followed by evaluation on 2,048 separate confirmation documents. Comparators select either the better unmixed parent or one of the six whole-model schedules $S(t), C(t)$ with $t \in \{4, 12, 32\}$ on the same calibration set.

Both stages score documents capped at 2,048 tokens. Normalized-text hashes are disjoint across selection, confirmation, and the two inherited development cohorts. A portable audit in the supplementary archive recomputes calibration selections, token-weighted NLLs, and paired intervals directly from per-document losses.

Model	Selected parent	Whole schedule	Depth combination
Llama-3.2-3B	2.541908	2.542224	2.539611
Qwen3-4B-Base	2.677741	2.676122	2.677470

Table 26: Confirmation NLL (↓, nats/token) for QKV-only ternary endpoint selection. Each model's columns use the same 2,048 documents. The selected whole schedules are $S(32)$ for Llama and $S(4)$ for Qwen.

The depth combination lowers NLL relative to the selected parent by 0.002297 on Llama and 0.000271 on Qwen. Paired document-bootstrap intervals are $[0.001378, 0.003217]$ and $[-0.000244, 0.000780]$, respectively, with Bonferroni adjustment over the two comparisons (5,000 resamples; 97.5% per-model intervals). Qwen's whole-schedule selection performs better than its depth combination. The results therefore support a small confirmed Llama gain over its parent, but no consistent advantage over whole-model schedule selection.

Exploratory QKV replacement and transform-transfer tests on Qwen3-4B-Base did not establish an improvement over TWLA (Zhao et al., 2026). Replacement changed both the quantizer and transform representation; the separate transform-transfer pilot remained calibration-only. The benefits reported here are therefore established within the studied quantizer and fitting setup.

F Historical Experiments and Numerical Controls

This supplement retains the earlier evidence that adds distinct initialization, solver, or numerical information. These experiments remain valid within their recorded protocols; they are not pooled with the four-model study. Complete earlier suites, including all branch replacements, direction tests, initialization sensitivities, numerical-work counts and direct-control tables, are preserved in `supplementary_data/historical_v11_12/`. Machine-readable direct-control data remain in `current_10_6/`, `matched_10_5/` and `stable_10_4/` under `supplementary_data/`.

F.1 Initial partitions and continuation

The initial comparison includes Mistral-7B-v0.1 (Jiang et al., 2023) and fits every nonempty QKV subset from identity, with at most 100 rounds per group, relative tolerance 10^{-7} and patience five. Different group counts imply different total fitting budgets. Only QKV weights are ternary; other tensors retain backbone precision. Ouro uses FP32 here, with its 24 physical layers tied across four recurrent steps and final-step logits; the other models use BF16. NLL is in nats per scored token. Physical-layer and branch averages weight each branch equally, including under grouped-query attention.

Table 27: Complete initial 128-document partition comparison (NLL, lower is better). Direct has no rotation; Random uses unfitted Gaussian-QR rotations, seed 20260910, shared across QKV and equal-width layers.

Model	FP	Direct	Random	Common	QK|V	QV|K	Q|KV	Ind.
Qwen	2.5336	5.3878	3.1851	2.8311	2.7520	2.7488	2.7761	2.7551
OPT	2.7215	5.0476	7.5987	6.6468	3.4128	4.3399	6.2444	3.0437
Ouro	2.5195	4.4783	3.9019	3.2065	2.9303	3.1556	3.1085	2.9319
Llama	2.3916	6.0619	3.8036	3.3789	2.8747	3.3498	3.0846	2.9854
Mistral	2.0883	6.0508	4.4981	4.9107	2.2100	4.5746	5.1626	2.2746

The corresponding pre-cast relative squared errors for QK|V/Independent are .063272/.052953 (Qwen), .082704/.069926 (OPT), .082351/.067574 (Ouro), .055092/.044281 (Llama), and .049589/.040794 (Mistral). These use the average in Eq. 2 on FP32 effective weights before installation, whereas later M_W uses installed weights. All five independent fits reduce this error; their NLL ordering varies by model.

Continuation uses FP32 proposals, the same 100-round limit and stopping thresholds, and a newly selected 128-document C4 cohort (65,222 Llama and 65,271 OPT prediction tokens). A1 retains sharing, B1 releases it from the same solution, and C1 continues the original independent solution. Initial-solution costs differ.

Table 28: Continuation NLL and paired differences; intervals are marginal 95% intervals.

Model	A1	B1	C1	B1−A1 [95% CI]	B1−C1 [95% CI]
Llama	2.8840	2.7118	2.9845	$-.1722$ $[-.1905, -.1554]$	$-.2727$ $[-.2953, -.2519]$
OPT	3.4336	2.9375	3.0731	$-.4960$ $[-.5127, -.4802]$	$-.1356$ $[-.1438, -.1280]$

In ARC-Challenge/HellaSwag/PIQA order, normalized accuracies (%) are Llama A1: 34.30/55.50/71.60, B1: 37.37/62.17/74.21, C1: 31.31/52.59/69.15; OPT A1: 25.68/41.70/67.52, B1: 26.88/47.89/70.18, C1: 27.39/46.45/70.29. Every Llama B1–C1 task gain survives Holm correction; only HellaSwag does on OPT. Fully resolved runtime settings were not saved, so these are within-run comparisons. NLL and task correction families retain the omitted adapted-FPT contrasts (five and fifteen tests); that local ternary adaptation is not a reproduction of FPTQuant.

F.2 Baseline and stabilized solvers

D fits independently; S uses QK|V then splits; V uses QV|K then splits; M selects branches from two independent starts by repaired pre-cast error, with first-start tie breaking. The baseline FP32 solver allocates

$$3(40) + 3 = 2(12) + 3(32) + 3 = 3(20) + 3(19) + 6 = 123 \tag{22}$$

factorizations for D, S/V and M, including final repairs. It stops at the first rejected proposal or five accepted improvements no larger than $10^{-7}\max(|f_{\text{prop}}|, 10^{-12})$, with acceptance tolerance $64\epsilon_{64}(1 + |f_{\text{cur}}|)$. Unused allocations are not transferred. Splitting copies raw rotations and recomputes codes without intermediate repair; final FP64 polar repair precedes FP32 reconstruction and BF16 installation. Slot 0 uses identity and Gaussian-QR seed $20260913 + 10000j + 10l + 1$ ($j = 0$ OPT, 1 Llama); D/S/V use identity and M uses both. QR uses FP64 CPU NumPy draws with fixed diagonal signs. Equal factorization ceilings do not imply equal runtime.

Table 29: Complete baseline-solver endpoints: installed M_W, NLL and normalized task accuracy (%). ARC-C denotes ARC-Challenge; HS denotes HellaSwag.

Model	Route	$M_W \downarrow$	NLL↓	ARC-C↑	HS↑	PIQA↑
Llama	D	.044286	2.984915	30.72	52.69	69.26
	S	.038873	2.713691	37.71	61.93	73.94
	V	.039193	2.941492	33.19	56.20	71.38
	M	.041904	2.566732	43.00	70.17	76.06
OPT	D	.069930	3.139846	27.30	46.43	70.35
	S	.062050	3.006090	27.13	47.89	69.26
	V	.062195	3.067512	27.22	47.88	70.29
	M	.066726	2.982659	26.11	48.44	70.73

Stabilized numerical update. Every phase initializes groups by FP64 polar decomposition and retains FP64 states, products, projection and objective reduction. Fast proposals use FP32 accumulation/SVD followed by two FP64 Newton–Schulz iterations. For $\|R^\top R - I\|_F/\sqrt{d}$, retraction entry/exit tolerances are $10^{-2}/10^{-10}$. A failed tolerance or $f_{\text{prop}} \geq f_{\text{cur}} - \tau$, $\tau = 64\epsilon_{64}(1 + |f_{\text{cur}}|)$,

triggers FP64 Procrustes rechecking; all later proposals for that group use FP64. Acceptance requires $f_{\text{prop}} \le f_{\text{cur}} + \tau$. FP64 rejection or the five-small-improvement rule above stops the group; the best accepted state is returned. Split groups receive phase-entry polar repair and recomputed codes. Final FP64 polar repair is per distinct selected raw tensor; FP32 reconstructed weights are cast to BF16. These tests do not certify stationarity.

S allocates $2(12) + 3(32) = 120$ group updates; same-start I40 allocates $3(40) = 120$, versus M's $3(20 + 19) = 117$. Phase-entry/final repairs and rechecks are additional work; the stabilized solver has no hard SVD ceiling. Bank 0 uses identity first and the earlier FP32 QR start promoted to FP64 second. Fresh starts use seed $2026104000 + 100000j + 1000b + 10l + t$, with $j = 1/2$ for OPT/Llama and $t = 1, 2$; S/I40 use the first tensor and M uses both. The same FP64 Gaussian-QR sign convention applies.

F.3 All seven stabilized S–M fits

C4 uses 128 new individually tokenized documents capped at 512 tokens, excluding earlier offsets/hashes (46,445 OPT; 45,635 Llama prediction tokens). WikiText-2 uses the first 128 non-overlapping 512-token windows of concatenated test text without special tokens (65,408 predictions). Paired ratio bootstraps resample whole units, with 10,000 draws and seed 20261042. Intervals are marginal and conditional on fitted candidates, not simultaneous intervals or uncertainty across starts.

Table 30: Every stabilized S–M fit, S minus M. Bank 0 begins from identity; the other banks use fresh random starts.

Model	Bank	ΔM_W	C4 ΔNLL [95% CI]	WikiText ΔNLL [95% CI]
OPT	0	$-.00469$	$.04882$ [$.04070, .05689$]	$.04853$ [$.04185, .05533$]
OPT	1	$-.00698$	$-.01679$ [$-.02053, -.01323$]	$-.03215$ [$-.03570, -.02863$]
OPT	2	$-.00697$	$-.01243$ [$-.01672, -.00826$]	$-.02675$ [$-.03062, -.02297$]
OPT	3	$-.00696$	$-.00438$ [$-.00900, .00021$]	$-.01637$ [$-.02058, -.01208$]
Llama	0	$-.00303$	$.17512$ [$.14728, .20860$]	$.20862$ [$.19555, .22189$]
Llama	1	$-.00488$	$-.02260$ [$-.03070, -.01514$]	$-.02713$ [$-.03090, -.02319$]
Llama	2	$-.00487$	$-.03458$ [$-.04506, -.02564$]	$-.02331$ [$-.02685, -.01975$]

S lowers weight error throughout, but worsens both NLLs at bank 0; OPT bank 3 C4 remains unresolved. Earlier bank-0 HellaSwag also favors M: S-only/M-only correct counts are 243/345 (OPT) and 160/931 (Llama), with Holm-adjusted $p = 2.98085 \times 10^{-5}$ and 2.03473×10^{-132}. Random initialization improves I40 more than S in the observed banks. Across 360 S/M layer–arm fits, fitting consumes 32,122 FP32 and 7,005 FP64 SVDs, 819 rechecks and 25.7108 summed hours.

F.4 Local diagnostics and execution precision

The baseline-solver direction study selects two sites per model using 32 scout documents, then evaluates 64 separate confirmation documents. Inputs are captured with all-S weights installed. The ideal intervention is

$$W_k^{(d,a)} = W_k + \|E_a\|_F E_d / \|E_d\|_F, \quad E_d = \overline{W}_{k,d} - W_k, \tag{23}$$

with $d, a \in \{S, M\}$. Native casting is norm-compensated; natural endpoints are reused exactly. Acceptance requires relative installed-norm error $\le 10^{-3}$ and direction cosine $\ge .995$; achieved extrema are 7.81×10^{-5} and .999952. These constructed weights need not be ternary-representable. Averaging the two amplitudes at Q=S gives direction effects $-.05557, +.00732, -.00804, +.01066$ at Llama L5/L8 and OPT L7/L23, respectively; Holm-8 p values are .00080/.12149/.00080/.01120. Changing Q to M changes these effects by $+.02580, +.02315, +.00010, +.00037$ (adjusted p=.00080/.00080/.81522/.81332). Thus three direction effects and both Llama context changes survive correction. The offline FP32 reference shares the tested Q context, original K and V_S; it is not the full floating-point block. Centering subtracts each query's valid-key mean; raw/centered errors have different denominators. KL uses masked log-softmax; output errors include W_o.

Table 31: All four local sites: M-minus-S direction differences at Q=S and S's natural amplitude. Errors are relative squared errors; KL and NLL are in nats. The primary direction test averages two amplitudes, unlike this table.

Site	K act.	Raw score	Centered	KL	Post-W_o	NLL
Llama L5	-.00145	-.01041	-.00298	-.15058	-.23356	-.05375
Llama L8	-.00193	-.00730	-.00221	-.05180	-.04367	+.00669
OPT L7	-.06393	-.06466	-.03760	-.03115	-.06938	-.00721
OPT L23	+.00006	-.00026	+.73583	+2.28482	+.42941	+.00977

The complete 32-condition data are in `local_fixed_table.csv`. Llama L8's primary NLL direction effect is unresolved after Holm-8 despite local improvements; its Q-context interaction is significant. A separate stabilized-endpoint check covers 256 condition/document rows. Native BF16 fails the offline-agreement thresholds (relative L2 $\leq .005$, cosine $\geq .9999$): maximum L2 is .11308 and minimum cosine .9946897. Offline/native FP32 maximum discrepancies are 3.6254×10^{-5} (OPT) and 1.1197×10^{-6} (Llama), versus native BF16/FP32 .10648/.02347. This supports a precision explanation on that path; local diagnostics remain descriptive, not validated native-attention mediators.

F.5 Archived direct controls and cost boundary

The complete earlier I40/I44/I-CM suite is retained under `supplementary_data/current_10_6/`: `TABLE_CONTRASTS.csv`, `TABLE_CANDIDATE_LEVELS.csv`, `TABLE_FP_NLL.csv`, `TABLE_COMPUTE_MATCH.csv`, `TABLE_HELLASWAG.csv` and `MCNEMAR_10_6.json`. All 42 NLL contrasts favor S with marginal intervals excluding zero (seed 20261061); they cover seven fits, not 42 independent replications. All twelve HellaSwag contrasts survive Holm-12. That execution uses `lm-eval` 0.4.9 (Gao et al., 2023), batch 16, length 2,048, truncation, zero shots, no chat template and all four seeds 1234. Earlier S/M tasks used batch 8/defaults; slight OPT score differences remain unexplained, so executions are not pooled.

I-CM selects a complete-sweep checkpoint solely by closeness of measured prefix-plus-final-repair cost to frozen S cost. Persistence, loading, evaluation and unselected checkpoint work are excluded. Of 180 layers, 119 match within $\pm 5\%$ and 61 stop under budget; all remain in evaluation. Stopping is not optimality. Detailed controls are archived as whole suites rather than selected favorable rows. Dense quality evaluation and offline fitting time do not establish deployment acceleration; the earlier packed correctness path showed no general quality/latency/resident-memory advantage, and QK|V resident memory exceeded FP slightly on Llama and Mistral.