RAZOR: PRUNING REPLACEABLE EXPERTS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-experts (MoE) models activate few experts per token but store the full expert pool. Pruning aims to reduce this burden at a fixed budget while preserving output distributions and, for reasoning models, reasoning ability. Deletion damage depends on functional replaceability by surviving computation, not usage or contribution magnitude alone. We introduce RAZOR, a training-free method that scores functional replaceability using consensus residuals, the deviations of expert outputs from the original weighted mixture. At a fixed layer input, an exact single-deletion identity accounts for survivor renormalization and router-selected refill. We aggregate these local scores over calibration tokens for budgeted pruning without gradients or recovery training. On GLM-4.7-Flash, Qwen3.6-35B-A3B, DeepSeek-V4-Flash-0731, and Hy3 at 25% and 50% expert removal, RAZOR achieves the highest nine-task macro average among the evaluated pruning methods in all eight settings. On the two backbones with matched REAP benchmark runs, it exceeds REAP by 2.12–5.59 points and wins all 36 paired task comparisons. It also lowers reverse KL relative to REAP in all four matched GLM-4.7-Flash and Qwen3.6-35B-A3B model–budget settings. Responses generated by Qwen3.6-35B-A3B nevertheless change in diversity, formatting, and termination, showing that task retention and predictive fidelity do not ensure generation stability.

1 INTRODUCTION

Mixture-of-experts (MoE) models increase capacity while activating only a small subset of experts per token, yet sparse computation does not reduce the cost of storing the full expert pool (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022). Whole-expert pruning addresses this burden by reducing the pool itself (Lu et al., 2024; Muzio et al., 2024; Lasby et al., 2026; Liu et al., 2024). Our focus is pruning reasoning models while retaining their reasoning ability. At a fixed budget, we seek to preserve the output distribution by assessing *functional replaceability*, not simply contribution magnitude.

Common pruning scores characterize experts by selection frequency or output magnitude (Muzio et al., 2024; Jaiswal et al., 2025; Lasby et al., 2026). These statistics do not capture how expert outputs combine, even when the output norm is weighted by routing mass. A large contribution may be replaceable by the surviving mixture, whereas a small one may supply a component that the survivors cannot recover. Isolated importance therefore differs from deletion damage, which depends on the computation available after removal.

Deletion changes a routed mixture by renormalizing the survivors' weights and, when the router refills the vacated slot, promoting an unselected expert. Figure 1 isolates both effects in a constructed single-token example. Without refill, panel (a) shows that removing the expert closest to the original mixture is less harmful than removing the one with the smallest output norm. With refill, panel (b) shows that a promoted expert can compensate for a removed balancing contribution and change the preferred deletion. Output magnitude, fixed-support damage, and refill-aware damage can therefore select different experts. More generally, identical gates and output norms can correspond to different least-harmful deletions (Proposition 1).

We explore functional replaceability through three consensus-residual scores. The reference is the original weighted MoE output, which we call the *consensus* without assuming expert agreement. RCS measures each expert's routing-weighted deviation from this mixture. RCS-LOO accounts for survivor renormalization, and RCS-Refill additionally includes router-selected replacement. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

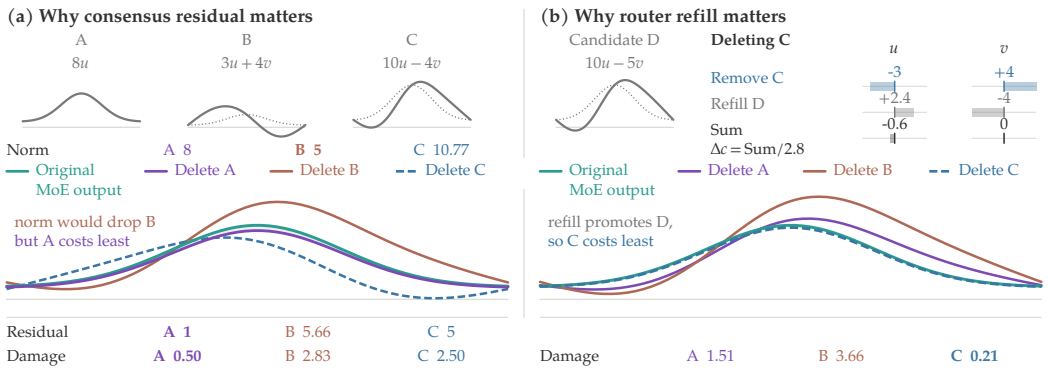


Figure 1: **Replaceability depends on output geometry and router refill.** A constructed single-token example without (a) and with (b) refill. Damage is L2 distance to the original mixture. Bold marks row minima. Appendix A gives the construction and plotting details.

latter two recover exact single-deletion output changes under their respective fixed-input assumptions. Comparing them tests whether a richer local counterfactual improves pruning. We aggregate these scores over calibration tokens and retain the highest-scoring experts within each layer's budget, using only forward computation, without gradients, subset search, or recovery training. Local change remains a surrogate for the model-level objective because joint deletions interact and altered activations propagate across layers.

To assess reasoning-model compression, we evaluate task retention alongside predictive fidelity and generation behavior. On GLM-4.7-Flash, Qwen3.6-35B-A3B, DeepSeek-V4-Flash-0731, and Hy3 at 25% and 50% expert removal, RCS-Refill has the highest nine-task macro average among the evaluated pruning methods in all eight settings, exceeding REAP by 2.12–5.59 points on the two backbones where REAP was also benchmarked. Aggressive pruning is not lossless, as all four backbones lose macro performance relative to their originals at 50% removal. On matched GLM-4.7-Flash and Qwen3.6-35B-A3B checkpoints, RCS-Refill also lowers reverse KL relative to REAP in all four model–budget settings. Analyses of responses generated by Qwen3.6-35B-A3B show why these forms of retention remain distinct from generation stability. Diversity does not follow the task ranking. At 50% removal, RCS-Refill produces fewer stray closing delimiters than REAP but more length-limited finishes than the unpruned model.

2 Methodology

2.1 Pruning objective and routed mixture

We retain B of E routed experts per layer, with $k \leq B < E$, without recovery training. To preserve the original output distribution at this budget, we use local Euclidean output change as a surrogate, not a global guarantee.

For token representation x, the router selects k experts $S(x)$. Expert i produces $f_i(x)$ with normalized weight $w_i(x) \geq 0$, where $\sum_{i \in S(x)} w_i(x) = 1$. The routed output is

$$c(x) = \sum_{i \in S(x)} w_i(x) f_i(x). \tag{1}$$

We call this mixture the *consensus*, without assuming that its experts agree. Unchanged residual-connection, dense, and shared-expert branches cancel in the local comparison. A common routed-output scale does not change within-layer rankings (Appendix B.3).

The statistic $w_i \| f_i \|_2$ measures the magnitude of expert i's weighted output contribution. Deletion also renormalizes the survivors, so its local effect depends on information that even the complete set of gates and output norms cannot supply.

Proposition 1 (Magnitude does not identify the least harmful deletion)**.** *Even with equal known gates, the expert-output norms do not in general determine which single-expert deletion minimizes the local output shift after survivor renormalization without refill.*

Proof. Consider equally weighted scalar outputs $(8, 3, 10)$ and $(8, 3, -10)$. Both have the same output norms and gates. Their original mixtures are 7 and $1/3$. The absolute shifts after deleting the first, second, or third expert are $(1/2, 2, 3/2)$ in the first case and $(23/6, 4/3, 31/6)$ in the second. The unique least harmful deletion is therefore the first expert in one case and the second in the other. A rule observing only the gates and norms cannot distinguish them. □

Reweighting output norms cannot recover the missing directional information. Given f_i and w_i, however, the shared vector c determines the survivor output.

2.2 From consensus residual to exact deletion effect

To isolate one deletion, we hold the layer input and expert outputs fixed and remove a selected expert with $w_i(x) < 1$. We first model router refill, then recover deletion without replacement as a special case.

Assumption 1 (Refill)**.** Deleting $i \in S(x)$ promotes the highest-ranked unselected expert r under the router's selection rule. Its nonnegative mixture score, divided by the original selected-score sum, defines the pseudo-weight $w_r(x)$. The surviving scores and this promoted score are then renormalized over their combined mass $1 - w_i(x) + w_r(x)$.

Assumption 2 (Fixed support)**.** The surviving weights are renormalized over $S(x) \setminus \{i\}$ without promoting an unselected expert. Equivalently, $w_r(x) \equiv 0$ in Assumption 1.

We write $\tilde{c}^{-i}$ for the refilled mixture and c^{-i} for the fixed-support one. The first uses the routed outputs and the promoted expert's output at the same layer input, whereas the second needs only the routed outputs. Neither requires evaluating a separate end-to-end pruned model for each expert.

Remark 1. Nonnegative $w_r(x)$ and $w_i(x) < 1$ ensure $D_i(x) \equiv 1 - w_i(x) + w_r(x) \geq 1 - w_i(x) > 0$. This argument does not require the selection ranking to match the ordering of mixture weights, which may differ when routing uses a selection-only correction bias (Appendix A.4).

Proposition 2 (Single-expert deletion under refill)**.** *Under Assumption 1, the token-level deletion damage is*

$$\delta_i(x) \equiv \left\| c(x) - \tilde{c}^{-i}(x) \right\|_2 = \frac{\left\| w_i(x) r_i(x) - w_r(x) r_r(x) \right\|_2}{D_i(x)}, \tag{2}$$

where $r_j(x) = f_j(x) - c(x)$ is the consensus residual of expert j.

Proof. Suppressing x, the promoted slot carries mass w_r while the surviving selected mass is $1 - w_i$, so $\tilde{c}^{-i} = (c - w_i f_i + w_r f_r)/D_i$. Using $D_i - 1 = -w_i + w_r$,

$$c - \tilde{c}^{-i} = \frac{D_i c - c + w_i f_i - w_r f_r}{D_i} = \frac{w_i(f_i - c) - w_r(f_r - c)}{D_i}.$$

Taking norms yields the result. □

Setting $w_r = 0$ recovers the fixed-support case in closed form.

Proposition 3 (Single-expert leave-one-out impact)**.** *Under Assumption 2, the token-level deletion damage is*

$$\delta_i^{\mathrm{loo}}(x) \equiv \left\| c(x) - c^{-i}(x) \right\|_2 = \frac{w_i(x)}{1 - w_i(x)} \left\| f_i(x) - c(x) \right\|_2. \tag{3}$$

Proof. Setting $w_r = 0$ in Proposition 2 gives denominator $1 - w_i$ and numerator $w_i \| f_i - c \|_2$. □

Equivalently, fixed-support damage is $w_i \| f_i - c^{-i} \|_2$. Distance to the survivor output measures local functional replaceability, and routing mass converts that distance into mixture change. A small change

can result from small distance or small routing mass. The LOO factor $1/(1 - w_i)$ corrects for self-inclusion in c. It follows from the deletion counterfactual rather than serving as a tunable importance weight (Appendix A.2). Refill adds an opposing weighted residual and changes the normalizer. The promoted expert can compensate for the removed contribution, but the change depends on residual magnitudes and routing weights as well as direction. An exact match $w_i r_i = w_r r_r$ leaves the local mixture unchanged.

Exact single-deletion effects are not additive. Without refill, removing a routed subset $\mathcal{A} \subseteq S(x)$ with total weight $W_{\mathcal{A}} = \sum_{i \in \mathcal{A}} w_i < 1$ gives

$$c - c^{-\mathcal{A}} = \frac{\sum_{i \in \mathcal{A}} w_i (f_i - c)}{1 - W_{\mathcal{A}}}. \tag{4}$$

Residuals can reinforce or cancel, while removed mass changes the denominator (Appendix A.3). Scalar ranking does not optimize this joint effect, and refill does not restore additivity. Promoting several ranks at once couples the removed experts through both the numerator and the set-dependent normalizer. If $W_{\mathcal{A}} = 1$, no routing mass survives and the fixed-support counterfactual is undefined, even if $B \geq k$ experts remain in the layer. Deployment reselects top-k from the retained pool and propagates changed activations across layers, so matching its refill rule at a single token still does not guarantee checkpoint fidelity.

2.3 Expert scoring and pruning

RCS uses $w_i \|r_i\|_2$, RCS-LOO uses δ_i^{loo}, and RCS-Refill uses δ_i. Unless explicitly varied, all three aggregate their token quantities by root mean square (RMS) over calibration tokens routed to each expert. Aggregation details and comparisons appear in Appendices C.2 and C.3. Table 5 maps the criteria to their experimental coverage.

We retain the highest-scoring experts within each layer's budget and deploy top-k routing over the retained pools. Appendix A.4 gives the implementation and pseudocode.

3 Experiments

Four research questions guide our evaluation of reasoning-model compression.

- **RQ1.** How well does pruning preserve output distributions across domains and routing conditions?
- **RQ2.** How well does it preserve downstream performance across four backbones and two removal budgets?
- **RQ3.** How does it affect response diversity, form, and termination separately from accuracy?
- **RQ4.** Do refinements to the local deletion score improve pruning, and how do scoring and aggregation choices affect expert selection and deletion-set interactions?

3.1 Experimental Setup

Models and pruning methods. We compare RCS, RCS-LOO, and Razor on GLM-4.7-Flash, Qwen3.6-35B-A3B, DeepSeek-V4-Flash-0731, and Hy3 at 25% and 50% expert removal, without recovery training. All three use conditional RMS unless an aggregation ablation is specified. *We call RCS-Refill with conditional RMS aggregation* Razor. Each method retains the highest-scoring experts per layer. Matched checkpoint and component studies use GLM-4.7-Flash and Qwen3.6-35B-A3B, whose routers select 4/64 and 8/256 experts, respectively.

Baselines. We compare against routing frequency (Frequency), a conditional-mean adaptation of expert activation norm (EAN) (Jaiswal et al., 2025), and REAP (Lasby et al., 2026), which averages routing-weighted output norms conditional on use. Downstream benchmark comparisons with Frequency, EAN, and REAP cover GLM-4.7-Flash and Qwen3.6-35B-A3B only. Separate archived, mask-based domain-wise and routing-stratified diagnostics include REAP on all four backbones (Appendices E.2 and E.3). REAP* is a separate external checkpoint calibrated with 24,576 samples, not the matched REAP baseline.



Figure 2: **Domain-wise reverse KL at 25% (top) and 50% (bottom) removal.** Values are in nats, with lower (inward) better. Compare methods along each spoke, not by polygon area. Whiskers show pointwise 95% paired-batch intervals. Shading distinguishes ID/OOD domains.

Data and metrics. Calibration uses the 2,048-example, seven-domain RAZORCAL pool with model-specific chat templates. Predictive fidelity is reverse KL, $D_{\mathrm{KL}}(q\|p)$, from original (p) and pruned (q) predictions on shared reference prefixes. Assistant-token NLL gives $\Delta\mathrm{NLL} = \mathrm{NLL}_q - \mathrm{NLL}_p$ and relative excess PPL, $\exp(\Delta\mathrm{NLL}) - 1$. The eight axes comprise four calibration types (mathematics, code, instruction following, and tools) and four held-out types (chat, creative writing, safety, and SQL). The matched component sweep pools tokens within axes and weights axes equally, whereas domain-wise and routing-stratified studies use grouped estimators.

The nine downstream tasks are AIME'26 (Mathematical Association of America, n.d.), IFEval (Zhou et al., 2023), IFBench (Pyatkin et al., 2025), SuperGPQA (Du et al., 2025), BFCL v4 (Patil et al., 2025; Mao et al., 2025), LongBench v2 (Bai et al., 2025), HumanEval+ (Liu et al., 2023), LiveCodeBench (2026 latest) (Jain et al., 2025), and SWE-bench Verified (Jimenez et al., 2024; Chowdhury et al., 2024). The suite emphasizes mathematics, programming, and long-context reasoning, complemented by instruction following, knowledge, and tool use. Appendix B.5 details task versions and score aggregation. Response analyses complement task scores by measuring diversity, stray closing thinking delimiters, code-fence structure, length, and length-limit finishes. These are observable response properties, not measures of latent reasoning.

Hyperparameters and evaluation protocol. Every scoring-rule comparison matches calibration inputs and layerwise budgets. For downstream benchmarks, we use temperature 0, except for SWE-bench Verified, where we use 0.7. AIME'26 reports avg@8, and the other eight benchmarks report avg@3. Diversity and response-form diagnostics follow the separate protocol of Section 3.4. Context and output limits, the thinking setting, and cross-study matching are detailed in Appendices B.1–B.5. Diagnostic sampling and uncertainty are described in Appendices E.2–E.6.

3.2 PREDICTIVE FIDELITY (RQ1)

Figure 2 compares REAP and the three residual criteria on eight domains, with backbones by column and removal budgets by row. Lower reverse KL indicates less drift. Spokes are scaled independently, mostly from nonzero origins, so methods are comparable on the same spoke, not by polygon area. Table 6 gives matched equal-axis GLM-4.7-Flash and Qwen3.6-35B-A3B estimates. Figure 8 compares REAP, RCS, RCS-LOO, and RAZOR across routing-concentration deciles under both fidelity measures, separating calibration-covered from held-out domains. All three scored criteria use conditional RMS there, so those curves isolate the token quantity.

Table 1: Nine-task benchmark results on GLM-4.7-Flash and Qwen3.6-35B-A3B at 25% and 50% removal. Scores are scaled by 100. Bold marks the best pruned score in each column within a model–budget block.

Model		Method	Math	Instruction Following			Know.	Tool	LC	Coding				*Overall Avg*
			AIME'26	IFEval	IFBench	*Avg*	SuperGPQA	BFCL v4	LongB. v2	HE+	LCB	SWE	*Avg*	
GLM-4.7-Flash	0%	—	31.7	79.7	40.0	59.9	38.5	65.3	25.3	79.9	37.4	5.2	40.8	44.8
	25%	Frequency	26.3	77.3	37.3	57.3	29.5	62.8	21.9	74.4	29.3	**5.2**	36.3	40.4
		EAN	30.4	76.3	39.3	57.8	31.5	64.7	27.2	75.0	33.8	4.8	37.9	42.6
		REAP*	30.0	74.9	36.7	55.8	32.6	64.9	27.2	75.0	30.6	3.6	36.4	41.7
		REAP	30.8	74.9	37.3	56.1	33.7	65.0	27.0	75.0	34.3	3.6	37.6	42.4
		RCS	29.6	77.1	36.0	56.6	33.3	**65.7**	26.0	73.2	33.8	4.2	37.1	42.1
		RCS-LOO	31.3	76.7	39.7	58.2	**34.3**	64.8	27.4	75.6	34.8	4.2	38.2	43.2
		RAZOR	**33.3**	**78.2**	**40.3**	**59.3**	34.2	65.6	**28.5**	**76.2**	**39.8**	4.6	**40.2**	**44.5**
	50%	Frequency	0.0	62.8	30.3	46.6	15.7	51.1	18.7	54.9	17.8	0.2	24.3	27.9
		EAN	15.8	70.6	29.0	49.8	25.4	54.8	23.7	59.2	**33.9**	2.6	31.9	35.0
		REAP	23.3	61.0	25.3	43.2	27.8	56.0	22.1	67.1	26.7	1.2	31.7	34.5
		RCS	29.6	63.8	27.3	45.6	24.9	**59.1**	19.5	71.3	27.6	3.0	34.0	36.2
		RCS-LOO	26.3	67.5	30.7	49.1	29.0	58.6	20.9	72.0	29.4	3.2	34.9	37.5
		RAZOR	**30.8**	**72.0**	**32.3**	**52.2**	**33.7**	**59.1**	**25.9**	**73.2**	29.8	**4.0**	**35.7**	**40.1**
Qwen3.6-35B-A3B	0%	—	76.3	83.9	33.7	58.8	62.8	67.6	50.1	91.5	71.9	17.0	60.1	61.6
	25%	Frequency	75.0	79.1	34.3	56.7	51.4	**67.2**	45.9	90.9	70.1	12.0	57.7	58.4
		EAN	75.4	81.7	32.3	57.0	57.4	65.8	48.7	92.7	64.8	16.6	58.0	59.5
		REAP	77.1	79.5	33.7	56.6	57.0	65.8	48.8	90.9	70.9	14.6	58.8	59.8
		RCS	76.7	80.0	38.0	59.0	57.3	66.3	48.5	92.1	71.9	18.0	60.7	61.0
		RCS-LOO	76.3	80.6	34.7	57.7	57.1	66.4	49.7	92.7	**72.6**	17.4	60.9	60.8
		RAZOR	**77.5**	**81.9**	**38.3**	**60.1**	**57.8**	67.0	**49.9**	**93.3**	**72.6**	**20.2**	**62.0**	**62.1**
	50%	Frequency	68.8	71.0	31.7	51.4	41.1	43.1	42.5	89.6	55.0	7.2	50.6	50.0
		EAN	72.1	74.1	30.3	52.2	46.1	58.7	46.1	88.4	47.1	8.4	48.0	52.4
		REAP	71.7	73.2	31.0	52.1	46.3	63.9	47.5	88.4	66.4	12.6	55.8	55.7
		RCS	73.3	74.3	**36.3**	55.3	48.3	64.1	**48.3**	91.5	67.6	12.8	57.3	57.4
		RCS-LOO	**75.4**	78.2	32.0	55.1	47.0	64.5	47.9	**92.1**	67.9	13.6	57.9	57.6
		RAZOR	75.0	**79.3**	34.0	**56.7**	**49.2**	**65.0**	**48.3**	**92.1**	**69.4**	**15.8**	**59.1**	**58.7**

Obs 1. RAZOR reduces predictive drift, but KL and likelihood need not agree. Across the four matched GLM-4.7-Flash and Qwen3.6-35B-A3B settings, RAZOR has lower reverse KL than REAP in all four and than RCS-LOO in three (Table 6). A separate routing-stratified comparison across all four backbones illustrates the distinction between fidelity measures. At 50% removal, RAZOR lowers KL relative to REAP in 73 of 80 model–domain-group–decile estimates, but keeps PPL closer to the original in only 47. The corresponding RCS-LOO counts are 72 and 47 (Appendix E.4). Reduced distributional drift therefore need not imply better preservation of reference-token likelihood.

Obs 2. KL improvements are not uniform across domains. Of eight domains, the counts where RAZOR lowers KL relative to REAP at 25% and 50% removal are, respectively, 5 and 7 for GLM-4.7-Flash, 6 and 5 for Qwen3.6-35B-A3B, 6 and 8 for DeepSeek-V4-Flash-0731, and 6 and 7 for Hy3. Yet Qwen3.6-35B-A3B's equal-domain held-out mean at 50% is 0.2% higher than REAP's, and the relative-change interval includes zero. Aggregate gains thus coexist with domain-level regressions. Appendix E.2 gives the corresponding RCS-LOO results. These domain-wise and routing-stratified views reweight overlapping predictions rather than provide independent replications.

3.3 DOWNSTREAM PERFORMANCE (RQ2)

Tables 1 and 2 group the four backbones by model and removal budget, with unpruned references at 0%. Scores are scaled by 100. Overall Avg weights all nine tasks equally, while the two *Avg* columns average instruction-following and coding tasks. Bold marks the best pruned score in each column within a model–budget block. Macro gaps use unrounded task-score averages.

Obs 3. RAZOR leads the macro comparison across all eight model–budget settings. Among the evaluated pruning methods, RAZOR has the highest macro score in every setting, exceeding RCS-LOO by 0.74–2.58 points across all eight. Against RCS-LOO, it wins 61 of 72 paired tasks, ties two and loses nine, so the macro advantage does not imply a uniform taskwise ordering. REAP comparisons cover only GLM-4.7-Flash and Qwen3.6-35B-A3B, giving four model–budget settings. Here RAZOR exceeds REAP by 2.12–5.59 points and wins all 36 paired task comparisons. In comparison, RCS-LOO exceeds REAP by 0.80–3.01 points while losing three of the same 36 tasks.

Table 2: Nine-task benchmark results on DeepSeek-V4-Flash-0731 and Hy3 at 25% and 50% removal. Conventions follow Table 1.

Model		Method	Math	Instruction Following			Know.	Tool	LC	Coding				*Overall Avg*
			AIME'26	IFEval	IFBench	*Avg*	SuperGPQA	BFCL v4	LongB. v2	HE+	LCB	SWE	*Avg*	
DeepSeek-V4-Flash-0731	0%	—	67.9	86.1	37.7	61.9	60.3	72.6	33.8	83.5	69.3	21.2	58.0	59.2
	25%	**RCS**	73.8	82.6	37.0	59.8	55.4	70.2	35.4	91.5	68.3	12.4	57.4	58.5
		RCS-LOO	76.7	83.7	**38.7**	**61.2**	55.6	70.4	35.2	89.6	68.8	**18.8**	59.1	59.7
		RAZOR	**81.3**	**84.7**	36.3	60.5	**56.0**	**72.8**	**37.4**	**93.3**	**69.1**	18.2	**60.2**	**61.0**
	50%	**RCS**	75.8	78.4	36.7	57.6	**46.7**	66.5	30.4	84.2	58.0	8.4	50.2	53.9
		RCS-LOO	76.7	79.8	**39.0**	59.4	46.5	68.1	29.6	85.4	**58.4**	**10.4**	**51.4**	54.9
		RAZOR	**78.8**	**80.4**	38.7	**59.6**	46.4	**72.1**	**30.8**	**86.0**	58.2	9.2	51.1	**55.6**
Hy3	0%	—	82.9	90.4	46.7	68.6	64.7	64.2	50.0	90.8	68.3	15.4	58.2	63.7
	25%	**RCS**	81.5	88.0	46.3	67.2	57.9	63.9	46.5	89.6	66.2	14.0	56.6	61.5
		RCS-LOO	83.0	89.1	46.0	67.6	58.3	64.2	47.3	90.2	65.7	13.4	56.4	61.9
		RAZOR	**88.9**	**89.5**	**47.3**	**68.4**	**59.2**	**64.3**	**48.2**	**90.8**	**66.9**	**15.8**	**57.8**	**63.4**
	50%	**RCS**	86.8	81.6	42.7	62.2	45.5	59.3	42.7	87.2	58.6	10.0	51.9	57.2
		RCS-LOO	91.6	83.7	44.7	64.2	46.1	59.6	**44.7**	89.0	58.2	10.0	52.4	58.6
		RAZOR	**92.3**	**84.3**	**45.0**	**64.7**	**46.3**	**60.2**	44.3	**90.2**	**61.7**	**11.2**	**54.4**	**59.5**

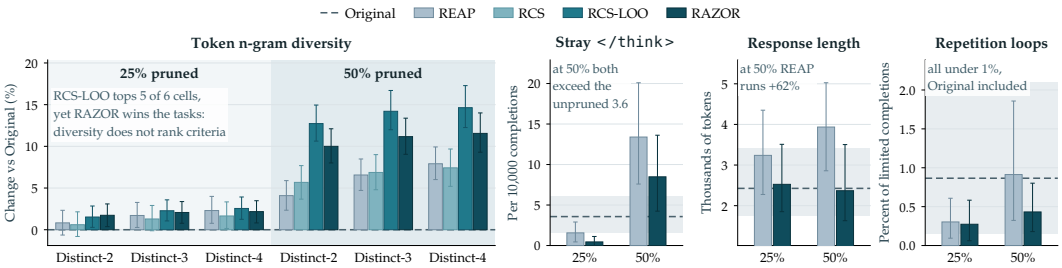


Figure 3: **Qwen3.6-35B-A3B response diversity and response form.** The left panel shows Distinct-n changes from Original. The right panels show stray </think> per 10,000 LiveCodeBench completions, median token length, and loop share among length-limited completions. Dashed lines mark Original, with 95% bands in the form panels. Whiskers show pointwise 95% bootstrap intervals.

These results favor refill for macro task utility within the residual family. Additional contrasts and checkpoint-matching limits appear in Appendix B.5.

Obs 4. Near-original performance at 25% gives way to consistent losses at 50%. The 25% RAZOR checkpoints remain close to or above the originals in macro score, whereas 50% removal is lossy on all four backbones. For example, GLM-4.7-Flash falls from 44.5 at 25% to 40.1 at 50%, against 44.8 unpruned. Qwen3.6-35B-A3B falls from 62.1 to 58.7, against 61.6. These small gains above the original macro scores do not establish that pruning improves the original models.

3.4 GENERATION BEHAVIOR (RQ3)

Beyond task accuracy, we examine response behavior in reasoning models. Figure 3 separates four properties of Qwen3.6-35B-A3B generations. The left panel compares all four pruning criteria by changes from Original in Distinct-n ($n = 2, 3, 4$), using 16 fixed responses per question and weighting AIME'26, HumanEval+, and LiveCodeBench equally. Shading separates the two budgets on a shared vertical scale. The three right panels compare REAP and RAZOR on LiveCodeBench using 256 shared sample positions per question (44,800 completions per variant), with separate scales for stray closing delimiters, median token length, and loop share among length-limited responses. Sampling uses temperature 0.7, top-$p = 0.95$, top-$k = 20$, an 8,192-token cap, and thinking disabled, probing response behavior rather than explicit reasoning traces. Diversity intervals use paired-question resampling, whereas form intervals resample questions within each variant. Both condition on the fixed generations (Appendices E.5 and E.6).

Obs 5. Higher diversity does not identify the best pruning criterion. RCS-LOO has the highest Distinct-n in five of the six combinations of removal budget and n-gram order and exceeds REAP in all six, with a median gap of 3.71 percentage points in relative change from Original, despite RAZOR's

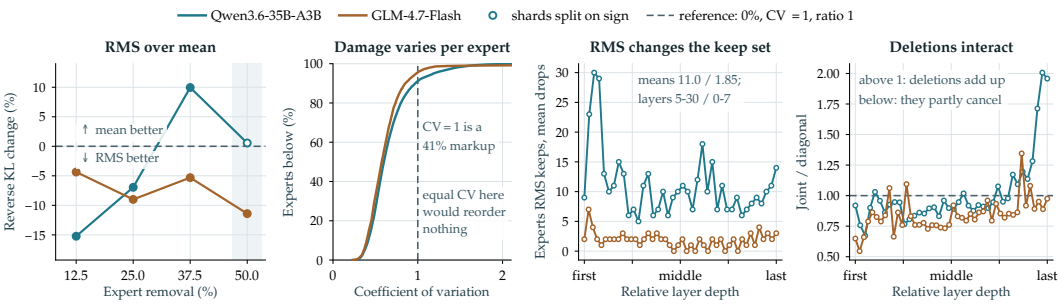


Figure 4: **RMS aggregation and expert selection for RAZOR.** Left to right, panels show relative KL change from mean (negative favors RMS), expert CV, RMS-only retained experts per layer, and deletion-set Gram ratios. Shading marks 50% removal, used in the last three panels.

stronger macro task performance. The exception is Distinct-2 at 25%, where RAZOR changes by $+1.73\%$ from Original versus $+1.53\%$ for RCS-LOO. Changes are smaller at 25% than at 50%. In a separate four-task analysis, Qwen3.6-35B-A3B's RCS-LOO–REAP Distinct-4 gap reverses under a 128-token prefix control that also changes question eligibility, so the reversal cannot be attributed to length alone (Appendix E.5). Higher diversity therefore establishes neither correctness nor a ranking robust to this joint control.

Obs 6. Closer response form does not imply preserved termination rates. At 50% removal, stray `</think>` emissions fall from REAP's 13.4 to RAZOR's 8.5 per 10,000 completions, whereas RCS-LOO reaches 68.5 (Table 10). Median token length is also closer to Original, at 2.37k for RAZOR versus 2.43k unpruned and 3.94k for REAP. These response-form comparisons use a different sample pool from the three-benchmark diversity result and do not control its length dependence.

Termination remains affected, with length-limited finishes reaching 25.8% for RAZOR and 24.2% for REAP, against 16.0% unpruned. The rightmost panel shows that few capped responses meet the loop criterion (fewer than one quarter of token 4-grams unique). Among length-limited completions, loop shares are 0.27%/0.43% for RAZOR and 0.30%/0.91% for REAP at 25%/50% removal. Thus loops under this definition account for less than 1% of capped responses, without ruling out other termination problems (Appendix E.6).

3.5 DISCUSSION (RQ4)

3.5.1 SCORING COMPONENTS

Table 6 compares REAP and the three residual criteria on matched GLM-4.7-Flash and Qwen3.6-35B-A3B checkpoints at both budgets. The criteria progressively represent more of one local deletion counterfactual. RCS scores the routing-weighted consensus residual, RCS-LOO adds survivor renormalization, and RAZOR adds router-selected replacement. Component contrasts in Appendix C.6 hold other choices fixed. The factorial study (Appendix C.5) uses partly different shards and is analyzed separately.

Obs 7. Consensus subtraction helps, while counterfactual refinements have mixed effects. Consensus subtraction lowers KL in all four model–budget settings when routing weights and mean aggregation are held fixed. This matched contrast isolates the reference change, unlike REAP-to-RCS, which also changes aggregation. Subsequent refinements are model-dependent. Under conditional RMS, adding the LOO factor lowers KL by 0.6%/5.6% on GLM-4.7-Flash but raises it by 7.1%/11.8% on Qwen3.6-35B-A3B at 25%/50% removal. Refill then improves on RCS-LOO in three of four settings, with GLM-4.7-Flash at 25% the exception ($+0.5\%$ KL). Thus a more complete single-deletion counterfactual need not produce a better jointly pruned checkpoint. The choice of RAZOR rests on the benchmark comparisons as well as these fidelity results, rather than local exactness alone.

3.5.2 EXPERT SELECTION

Figure 4 connects aggregation to selection for RAZOR through RMS-versus-mean KL across four budgets, per-expert coefficient of variation (CV), per-layer keep-set swaps, and deletion-set Gram

ratios. Hollow markers in the KL panel indicate disagreement in sign across its four evaluation shards. The latter three panels use 50% removal. Table 7 compares RAZOR and RCS-LOO selection statistics side by side.

Obs 8. RMS changes retained sets, with model-dependent fidelity benefits. RMS yields lower KL than mean aggregation at all four budgets on GLM-4.7-Flash, but higher KL above 25% removal on Qwen3.6-35B-A3B. This differs from RCS-LOO, whose RMS-over-mean gains are small on Qwen3.6-35B-A3B (Appendix C.6). Variation in the RMS premium $\sqrt{1 + \mathrm{CV}^2}$ reweights experts rather than uniformly rescaling them. At 50% removal, RMS retains an average of 11.0 experts per layer that mean would discard on Qwen3.6-35B-A3B and 1.85 on GLM-4.7-Flash, from retention budgets of 128 and 32, respectively. The per-layer ranges are 5–30 and 0–7. This establishes changed selection, not that the swaps cause benchmark gains. Appendix C.3 gives further aggregation contrasts.

Obs 9. Local diagnostics expose interactions that scalar rankings do not resolve. The Gram-ratio panel shows net reinforcement in four GLM-4.7-Flash layers for RAZOR, whereas RCS-LOO-selected sets show net cancellation in all 46 GLM-4.7-Flash layers and 29 of 40 Qwen3.6-35B-A3B layers (Appendix C.7.2). These interactions concern the fixed-support numerator, constrained by $\sum_i w_i(f_i - c) = 0$. The denominator and refill remain outside this probe. Single-deletion rankings also show no consistent advantage over REAP, despite favorable RCS-LOO comparisons in concentrated-routing bins (Appendices C.7.1 and C.7.3). Thus the probes expose interactions without establishing optimal joint selection or the cause of the benchmark gains.

4 Related Work

MoE compression methods differ in whether they score experts independently, model relations among them, or optimize retained sets. We organize the closest work along these choices, separating RAZOR's fixed-input deletion–refill score from merging, recovery, and combinatorial selection. Appendix F provides a broader treatment, including budget allocation and finer-grained compression.

Expert scoring. At the individual-expert level, pruning methods use routing, activation, or weight statistics (Muzio et al., 2024; Jaiswal et al., 2025; Liu et al., 2026a). Deletion-aware methods also model the computation after removal. REAP includes promoted substitution and survivor renormalization but retains only the removed expert's routing-weighted output norm (Lasby et al., 2026), while a unified framework derives exact damage yet omits the rerouting residual from practical scores (Liu et al., 2026b). RAZOR instead scores the full fixed-input consensus-relative deletion–refill change.

Expert relations. Independent scores can miss structure shared across experts. STUN and HC-SMoE cluster router behavior or expert outputs, while SHAPE, ConMoE, and MAESTRO use routing co-occurrence, parameter-space replaceability, or cross-layer transitions (Lee et al., 2025; Chen et al., 2025; Zhang, 2026; Yao et al., 2026; Goel et al., 2026). These relations support selection, merging, or path modeling, whereas RAZOR scores deletion with survivor renormalization and refill, without recovery.

Candidate-set search. Beyond relational ranking, other methods evaluate retained sets through reconstruction, joint pruning–merging, or layerwise and blockwise search (Lu et al., 2024; Liu et al., 2024; Yang et al., 2024). RAZOR performs no subset search. Its exactness for one fixed-input deletion does not solve joint pruning, because deletions interact and a refill candidate may itself be removed.

5 Conclusion

We presented RAZOR for training-free expert pruning of reasoning MoE models. Its consensus residuals score functional replaceability under survivor renormalization and router refill. Across four post-trained MoE models and two removal budgets, it achieves the highest nine-task macro average among the evaluated methods in all eight settings and lowers reverse KL against REAP in all four matched GLM-4.7-Flash and Qwen3.6-35B-A3B comparisons. These gains do not guarantee stable generation, motivating evaluation of both retained capability and changed behavior.

AI USE STATEMENT

Generative AI tools assisted with the refinement of academic prose, focusing on clarity, coherence, and precision of expression. They also supported literature search and coding.

ETHICS STATEMENT

This work studies expert pruning of post-trained MoE language models, aiming to reduce model storage requirements while retaining their capabilities. Removing experts can also alter behaviors shaped by post-training, including safety-related responses, even when aggregate task performance is preserved. Agreement with the original model on safety-related inputs measures predictive fidelity, but does not establish safety or preservation of alignment. We therefore emphasize independent safety and alignment evaluation before deployment, particularly in safety-sensitive applications. The study uses previously released calibration data (Appendix B.1) and involves no human-subject research.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix A provide the scoring criterion and derivations. Appendix A.4 gives the algorithm and implementation details. Appendix B documents calibration data, model configurations, evaluation protocols, score aggregation, and uncertainty estimates. Appendix B.5 records unavailable evaluator settings and cross-study matching limits that prevent exact reproduction of all comparisons. Model checkpoints are not redistributable by us.

REFERENCES

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Yuxiao Dong, Jie Tang, Lei Hou, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

I-Chun Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, and Chun-Yi Lee. Retraining-free merging of sparse MoE via hierarchical clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 8594–8620. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25aq.html.

Neil Chowdhury, James Aung, Chan Jun Shern, Oliver Jaffe, Dane Sherburn, Giulio Starace, Evan Mays, Rachel Dias, Marwan Aljubeh, Mia Glaese, Carlos E. Jimenez, John Yang, Leyton Ho, Tejal Patwardhan, Kevin Liu, and Aleksander Madry. Introducing SWE-bench Verified. OpenAI, August 2024. URL https://openai.com/index/introducing-swe-bench-verified/.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Qinrui Li, Rui Li, Sirun Li, Yizhi Li, Yunwen Li, Yiyan Liao, David Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zekun Moore Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Zili Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xingjian Zhang, Xiyue Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Meng Cao, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Xiang Yue, Zhaoxiang Zhang, Wangchunshu

Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-3766. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a3c5af1f56fc73eef1ba0f442739f5ca-Paper-Datasets_and_Benchmarks_Track.pdf.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

GLM-4.5 Team. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*, 2025. URL https://arxiv.org/abs/2508.06471.

Palaash Goel, Ayush Maheshwari, and Tanmoy Chakraborty. It takes a MAESTRO to prune bad experts. *arXiv preprint arXiv:2607.08601*, 2026. URL https://arxiv.org/abs/2607.08601.

Hugging Face Smol Models Research. SmolTalk2. Hugging Face dataset card, 2025. URL https://huggingface.co/datasets/HuggingFaceTB/smoltalk2. Accessed 2026-09-14.

Sieun Hyeon and Jaeyoung Do. Is retraining-free enough? the necessity of router calibration for efficient MoE compression. *arXiv preprint arXiv:2603.02217*, 2026. URL https://arxiv.org/abs/2603.02217.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

Ajay Jaiswal, Jianyu Wang, Yixiao Li, Pingzhi Li, Tianlong Chen, Zhangyang Wang, Chong Wang, Ruoming Pang, and Xianzhi Du. Finding fantastic experts in MoEs: A unified study for expert dropping strategies and observations. *arXiv preprint arXiv:2504.05586*, 2025. URL https://arxiv.org/abs/2504.05586.

Saurav Jha, Maryam Hashemzadeh, Ali Saheb Pasand, Ali Parviz, Min-Joong Lee, and Boris Knyazev. REAM: Merging improves pruning of experts in LLMs. *arXiv preprint arXiv:2604.04356*, 2026. URL https://arxiv.org/abs/2604.04356.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. REAP the experts: Why pruning prevails for one-shot MoE compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ukGxWd2aDG.

Jaeseong Lee, Seung-won Hwang, Aurick Qiao, Daniel F Campos, Zhewei Yao, and Yuxiong He. STUN: Structured-then-unstructured pruning for scalable MoE pruning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13660–13676. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.acl-long.671. URL https://aclanthology.org/2025.acl-long.671/.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. doi: 10.52202/068431-0278. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119. Association for Computational Linguistics, 2016. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Ke Li, Zheng Yang, Zhongbin Zhou, Feng Xue, Zhonglin Jiang, and Wenxiao Wang. HEAPr: Hessian-based efficient atomic expert pruning in output space. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JAbMgS7gl6.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient SMoE with hints from its routing policy. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=eFWG9Cy3WK.

Enshu Liu, Junyi Zhu, Zinan Lin, Xuefei Ning, Matthew B. Blaschko, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs. *arXiv preprint arXiv:2407.00945*, 2024. URL https://arxiv.org/abs/2407.00945.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572. Curran Associates, Inc., 2023. doi: 10.52202/075280-0943. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/43e9d647ccd3e4b7b5baab53f0368686-Paper-Conference.pdf.

Zongfang Liu, Guangyi Chen, Shengkun Tang, Yifan Shen, Huan Wang, and Xin Yuan. AIMER: Calibration-free task-agnostic MoE expert pruning. *arXiv preprint arXiv:2603.18492*, 2026a. URL https://arxiv.org/abs/2603.18492.

Zongfang Liu, Jinghui Zhang, Zijian Ma, Guangyi Chen, and Xin Yuan. How to score experts for one-shot MoE expert pruning: A unified formulation and selection principle. *arXiv preprint arXiv:2606.15716*, 2026b. URL https://arxiv.org/abs/2606.15716.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.334. URL https://aclanthology.org/2024.acl-long.334/.

Huanzhi Mao, Raymond Tsao, Jingzhuo Zhou, Shishir G. Patil, and Joseph E. Gonzalez. BFCL V4: Agentic part 1: Web search, July 2025. URL https://gorilla.cs.berkeley.edu/blogs/15_bfcl_v4_web_search.html.

Mathematical Association of America. MAA invitational competitions. Official competition website, n.d. URL https://maa.org/maa-invitational-competitions/.

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.

Alexandre Muzio, Alex Sun, and Churan He. SEER-MoE: Sparse expert efficiency through regularization for mixture-of-experts. *arXiv preprint arXiv:2404.05089*, 2024. URL https://arxiv.org/abs/2404.05089.

NVIDIA. Nemotron 3 Super: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for agentic reasoning. *arXiv preprint arXiv:2604.12374*, 2026. URL https://arxiv.org/abs/2604.12374.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1645. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/46499a0622ecf568b72d17b61e45dbd5-Paper-Datasets_and_Benchmarks_Track.pdf.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961. Dataset card: https://huggingface.co/datasets/allenai/Dolci-Instruct-SFT.

Teknium. OpenHermes 2.5: An open dataset of synthetic data for generalist LLM assistants. Hugging Face dataset card, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

Tencent Hy Team. Hy3. Hugging Face model card, 2026. URL https://huggingface.co/tencent/Hy3.

Yanyue Xie, Zhi Zhang, Ding Zhou, Cong Xie, Ziang Song, Xin Liu, Yanzhi Wang, Xue Lin, and An Xu. MoE-Pruner: Pruning mixture-of-experts large language model using the hints from its router. *arXiv preprint arXiv:2410.12013*, 2024. URL https://arxiv.org/abs/2410.12013.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Yuanlin Duan, Wenqi Jia, Miao Yin, Yu Cheng, and Bo Yuan. MoE-I^2: Compressing mixture of experts models through inter-expert pruning and intra-expert low-rank decomposition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10456–10466. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.612. URL https://aclanthology.org/2024.findings-emnlp.612/.

Yilun Yao, Jiaming Pan, Elsie Dai, Peizhuang Cong, Yaoming Li, and Tong Yang. ConMoE: Expert-pool consolidation via prototype reassignment for MoE compression. *arXiv preprint arXiv:2605.29350*, 2026. URL https://arxiv.org/abs/2605.29350.

Yuhao Zhang. SHAPE: Coalition-aware expert pruning for sparse mixture-of-experts LLMs. *arXiv preprint arXiv:2606.09886*, 2026. URL https://arxiv.org/abs/2606.09886.

Zeliang Zhang, Nikhil Ghosh, Jiani Liu, Bin Yu, and Xiaodong Liu. Does a global perspective help prune sparse MoEs elegantly? *arXiv preprint arXiv:2604.06542*, 2026. URL https://arxiv.org/abs/2604.06542.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

Appendix A derives the local deletion quantities and describes how they are collected. Appendix B separates calibration, reference-response fidelity, and downstream benchmark protocols. Appendix C tests scoring choices and selection diagnostics. Appendix D examines calibration stability and configuration sensitivity. Appendix E details the output analyses, their aggregation units, and their uncertainty estimates. Appendix F expands the comparison with related compression methods.

A Additional Derivations and Implementation

A.1 Residual decomposition

Expanding the consensus residual gives

$$\|f_i - c\|_2^2 = \|f_i\|_2^2 - 2\langle f_i, c\rangle + \|c\|_2^2, \tag{5}$$

$$\langle f_i, c\rangle = \|f_i\|_2 \, \|c\|_2 \cos\theta_i. \tag{6}$$

The residual depends on direction as well as magnitude. At fixed norms, a positive inner product reduces it and a negative one increases it. Even without collinear or near-duplicate outputs, an aligned expert may be replaceable, while one balancing the others may be costly to lose. The mixture includes expert i, so a further self-inclusion correction is needed to obtain distance to the survivors. Across calibration tokens, changes in mixture norm, gates, and routed sets can alter rankings, so the geometric identity alone does not determine which retained set performs better.

Constructed example in Figure 1. Let u, v be orthonormal and let E_A, E_B, E_C have outputs $8u$, $3u + 4v$, and $10u - 4v$, with equal initial weights. Their consensus is $c = 7u$, the routed-expert mixture at one layer, not the model's final output. Residuals are $\|f_i - c\|_2$. Without refill, deleting E_A, E_B, E_C and renormalizing the survivors gives damage 0.5, $\sqrt{8} \approx 2.83$, and 2.5, respectively. With refill, the same initially unselected $E_D = 10u - 5v$ enters after every deletion, with survivor and replacement masses $1, 1, 0.8$. For deletion of E_C, the bars show $-(f_C - c)$, $0.8(f_D - c)$, and their unnormalized sum. The v terms cancel, giving $\Delta c = -0.6u/2.8$ and damage $3/14 \approx 0.21$, versus 1.51 for deleting E_A and 3.66 for deleting E_B. Thus magnitude, fixed-support damage, and refill damage favor E_B, E_A, E_C, respectively. This constructed contrast illustrates the role of output geometry and refill, not a measured checkpoint-level gain. Curves span output coordinates, with mixture curves using $2.5\times$ the expert vertical scale. Dotted curves isolate u components, and shading marks coordinate-wise differences.

A.2 Survivor-distance identity

For $w_i < 1$, starting from the fixed-support post-removal output gives

$$c^{-i} = \frac{c - w_i f_i}{1 - w_i}, \tag{7}$$

$$f_i - c^{-i} = f_i - \frac{c - w_i f_i}{1 - w_i} = \frac{f_i - c}{1 - w_i}. \tag{8}$$

Taking norms gives the survivor-distance identity used in the main text,

$$\|f_i - c\|_2 = (1 - w_i) \, \|f_i - c^{-i}\|_2, \tag{9}$$

and combining equation 9 with Proposition 3 yields

$$\|c - c^{-i}\|_2 = w_i \|f_i - c^{-i}\|_2, \tag{10}$$

which expresses damage as routing mass times distance to the survivors. The distance describes local functional replaceability at the same input, while the weight describes its effect on the mixture. Neither alone measures redundancy in the final retained model. The factor $1/(1 - w_i)$ is required to convert the observed consensus residual into this survivor distance. Omitting it gives $(1 - w_i)\delta_i^{\mathrm{loo}}$ at each token, a different proxy rather than the same deletion damage. Since w_i varies across tokens, this is not a common rescaling of the aggregated expert score. With inputs and gates fixed, adding a common vector g to all routed outputs shifts both mixtures by g and leaves deletion damage unchanged, although output norms may change. This translation invariance applies to the local comparison, not to model predictions.

For $\delta_i^{\mathrm{loo}} > 0$, expanding the refill numerator gives

$$\frac{\delta_i}{\delta_i^{\mathrm{loo}}} = \frac{1 - w_i}{1 - w_i + w_r} \sqrt{1 - 2\rho_i \cos\theta_i + \rho_i^2}, \qquad \rho_i = \frac{w_r \|r_r\|_2}{w_i \|r_i\|_2}, \tag{11}$$

where θ_i is the angle between nonzero residuals. When the promoted weighted residual is zero, the square-root factor is one. The denominator factor is at most one, and strictly smaller when $w_r > 0$. For $\rho_i > 0$, the numerator factor exceeds one exactly when $\rho_i > 2\cos\theta_i$. Thus routing mass, residual magnitudes, and alignment jointly determine whether refill increases local damage. Orthogonality makes the numerator factor exceed one when $\rho_i > 0$, but does not determine the full damage ratio. Appendix C.4 reports score ratios and partial geometric summaries, not a complete empirical decomposition of these factors.

A.3 Exact perturbation for a selected set

For a removed routed subset $\mathcal{A}$ with $W_{\mathcal{A}} < 1$, the survivor mixture is $c^{-\mathcal{A}} = (c - \sum_{j\in\mathcal{A}} w_j f_j)/(1 - W_{\mathcal{A}})$. Subtracting it from c gives Eq. 4. All residuals $r_j = f_j - c$ share the original consensus, so the squared numerator expands as

$$\Big\| \sum_{j\in\mathcal{A}} w_j r_j \Big\|_2^2 = \underbrace{\sum_{j\in\mathcal{A}} w_j^2 \|r_j\|_2^2}_{\text{diagonal}} + \underbrace{\sum_{j\neq l\in\mathcal{A}} w_j w_l \langle r_j, r_l\rangle}_{\text{interference}}. \tag{12}$$

Cross terms and the set-dependent denominator prevent an additive decomposition into independent removal costs. The diagonal term also sums experts within a token, whereas RMS aggregates tokens within an expert. Thus the expansion is neither a derivation of RMS nor an exact joint objective obtained by summing expert scores. Refill does not repair this. Removing $\mathcal{A}$ promotes the $|\mathcal{A}|$ highest-ranked unselected experts $\mathcal{R}$, so the numerator becomes $\| \sum_{j\in\mathcal{A}} w_j r_j - \sum_{r\in\mathcal{R}} w_r r_r\|_2$ and the denominator becomes $1 - W_{\mathcal{A}} + W_{\mathcal{R}}$. Both now depend on the whole set, and the expansion acquires cross terms within $\mathcal{R}$ and between $\mathcal{A}$ and $\mathcal{R}$ in addition to those in Eq. 12. These set-dependent interactions are not captured by summing scalar single-expert scores.

A.4 Implementation and complexity

A.4.1 Online statistics and bounded working buffers

Algorithm 1 changes only the token-level quantity across the three criteria, which share chunking, conditional RMS aggregation, ranking, and physical pruning. The RCS and RCS-LOO branches reconstruct only the routed expert outputs, whereas RCS-Refill additionally reconstructs the rank-$(k{+}1)$ promoted expert. Selecting RCS-Refill with conditional RMS gives RAZOR. For each learned-router layer, the collector obtains the routing context for the batch, partitions its token positions, and locally reconstructs the required expert outputs for each chunk. These outputs form the consensus and residuals before the collector advances to the next chunk. This is additional local computation alongside the original forward, not a separate end-to-end model evaluation for every candidate deletion.

With T packed token positions, top-k routing, and expert-output width d, storing the routed outputs together with the promoted one requires $O(T(k{+}1)d)$ space. Restricting the working set to C positions replaces this with $O(C(k+1)d)$, plus $O(Cd)$ for the consensus. Only $q_i = \sum_t \delta_i(t)^2$ and the valid routed-token count n_i persist for RMS scoring, requiring $O(E)$ state per layer. Disjoint calibration shards accumulate these quantities independently. Summing them before taking $\sqrt{q_i/\max(n_i, 1)}$ gives the pooled conditional RMS, unlike averaging shard-level RMS scores. Shards must share the same model and calibration configuration. Counts and moment sums are accumulated in double precision, but floating-point reduction order need not be bitwise identical across schedules. Unobserved experts receive zero score. We remove the lowest $E - B$ scores without a secondary ranking criterion for exact ties.

The instrumented collector also computes ablation moments for the fixed-support and gate-only instances, reusing the same reconstructed outputs without additional expert-output buffers. An

Algorithm 1 Chunked scoring and routing-aware physical pruning for the RCS family

Require: Model, calibration data $\mathcal{D}$ with valid-position masks, budgets B_ℓ, chunk size C, criterion $a \in \{\mathrm{RCS}, \mathrm{RCS\text{-}LOO}, \mathrm{RCS\text{-}Refill}\}$, clamp ϵ

Initialize per-shard $q_{\ell i} \leftarrow 0$, $n_{\ell i} \leftarrow 0$ for learned-router layers
for learned-router layer input from an unpruned forward over $\mathcal{D}$ **do**
  Obtain routed sets S_t and normalized selected weights w_t
  if $a = \mathrm{RCS\text{-}Refill}$ **then**
    Obtain the rank-$(k{+}1)$ expert r_t and its pseudo-weight $w_r(t)$
  end if
  for token chunk X of size at most C **do**
    Reconstruct routed outputs $f_i(t)$ and form $c_t = \sum_{i \in S_t} w_i(t) f_i(t)$ and $r_i(t) = f_i(t) - c_t$
    if $a = \mathrm{RCS}$ **then**
      $d_i(t) \leftarrow w_i(t) \|r_i(t)\|_2$
    else if $a = \mathrm{RCS\text{-}LOO}$ **then**
      $d_i(t) \leftarrow w_i(t) \|r_i(t)\|_2 / \max(1 - w_i(t), \epsilon)$
    else
      Reconstruct $f_{r_t}(t)$, set $r_{r_t}(t) = f_{r_t}(t) - c_t$
      $d_i(t) \leftarrow \|w_i(t) r_i(t) - w_r(t) r_{r_t}(t)\|_2 / \max(1 - w_i(t) + w_r(t), \epsilon)$
    end if
    For valid routed pairs, accumulate $q_{\ell i} \mathrel{+}= d_i(t)^2$, $n_{\ell i} \mathrel{+}= 1$; discard chunk-local outputs
  end for
end for
Sum q, n across disjoint shards; set $s_{\ell i} = \sqrt{q_{\ell i} / \max(n_{\ell i}, 1)}$
Retain the B_ℓ highest-scoring experts in each learned-router layer
For fixed-hash layers, select by table counts and remap to distinct surviving routes
return Model with consistently sliced expert tensors, router rows, and updated route indices

optional Gram matrix, enabled by default in this collector, adds $O(E^2)$ state per layer and a separately chunked workspace. These diagnostics are unnecessary for the RMS updates. Expert outputs are converted to single precision for residual calculations. Clamping each denominator from below by 10^{-6} gives $\max(1 - w_i + w_r, 10^{-6})$ for the refill score and $\max(1 - w_i, 10^{-6})$ for its fixed-support special case. The identity describes the unclamped quantity, so the computed score can differ from exact local damage when this safeguard is active. The buffer bounds exclude model weights, full-batch routing tensors, ordinary forward activations, expert-MLP intermediates, and attention workspaces. They describe the storage controlled by this design, not a total peak-memory guarantee.

A.4.2 Layer-wise execution

Output chunking does not address the memory occupied by model weights. For models that exceed the available device capacity, a layer-wise schedule loads one decoder layer, propagates every sequence in the current calibration shard through it, collects its statistics, and releases its weights before loading the next layer. The shard's hidden states and model-specific cross-layer state remain resident. Dividing calibration into smaller disjoint shards therefore controls a different memory cost from token chunking. For DeepSeek-V4-Flash-0731, attention and indexer queries are also chunked, avoiding full query-by-key score workspaces while retaining their key coverage. These schedules support the same score definition, but their availability does not establish which schedule was used by every reported checkpoint.

Calibration examples are tokenized and packed before collection. Streaming here concerns the online reduction of statistics and staged loading of weights, rather than the sampling of calibration examples.

A.4.3 Architecture-specific routing

The derivation requires normalized selected weights, not a particular router parameterization. We preserve each architecture's native selection rule and score function, using softmax probabilities for Qwen3.6-35B-A3B, sigmoid scores for GLM-4.7-Flash, DeepSeek-V4-Flash-0731's native score transformation, and sigmoid scores for Hy3. Where a correction bias determines selection, it is not added to the mixture weights. Selected unmodified scores are renormalized before evaluating

the damage formula. The implementation retains the routed-output scale λ, although omitting this common positive factor from the displayed score leaves within-layer rankings unchanged.

DeepSeek-V4-Flash-0731 includes three fixed-hash layers whose expert identities come from a frozen token-to-expert table. They are handled separately from learned-router saliency. To retain B experts, we select the most frequent experts by counting occurrences over the entire table, without weighting by calibration-token occurrence. Each table row first preserves its surviving expert identities. Removed entries are then replaced, in slot order, by the cosine-nearest surviving router-weight vector not already used in that row. With $B \geq k$, the result has k distinct routes. Reindexing uses the same survivor order as parameter slicing. This is a routing repair, not expert-weight merging or an optimal assignment. The routing-stratified study shares this hash-layer selection across criteria, but matching remains unconfirmed for all downstream checkpoints (Appendix B.5).

A.4.4 PHYSICAL PRUNING

We slice expert-indexed tensors and corresponding router rows with the same keep indices, leaving dense blocks and shared experts unchanged. Under a common retained width, the result is a smaller MoE checkpoint requiring neither runtime masks nor custom sparse kernels. A separate option retains hash layers at full width, producing a mixed-width model that requires a compatible architecture configuration and loading implementation. For Qwen3.6-35B-A3B, reducing 256 to 128 routed experts per layer yields an approximately 19B-parameter model.

Removing experts reduces stored parameters, but preserving top-k does not proportionally reduce active expert computation. We do not measure serving latency, energy use, or end-to-end acceleration.

B DATA AND EVALUATION PROTOCOL

We use open Nemotron post-training datasets. The Nemotron 3 Super technical report provides context for their development and reuse across the model family (NVIDIA, 2026). The specific dataset versions used here are listed below.

B.1 CALIBRATION DATA

RAZORCAL contains 2,048 chat-format examples across seven domains (Table 3). It is a calibration mixture assembled from existing instruction datasets, not a new independently collected corpus. Coding receives twice the allocation of each other domain to cover function completion, competitive programming, class-based solutions, software engineering, and tool use. Source-specific sampling applies quotas, length filters, and deduplication. Preprocessing removes dataset-specific harness prefixes where appropriate and caps examples at 32K characters, not tokens. Each checkpoint's chat template renders the preserved conversation roles before tokenization.

Sources and composition. Coding includes 205 Python-algorithm examples from Dolci-Instruct-SFT, the instruction-tuning mixture accompanying Olmo 3 (Team Olmo et al., 2025). It adds 141 Python and 115 C++ examples from Nemotron-SFT-Competitive-Programming-v2, 40 agentless software-engineering examples from Nemotron-SFT-SWE-v2, and 11 general tool-use examples from Nemotron-SFT-OpenCode-v1. Nemotron-SFT-SWE-v2 covers code localization, repair, and test generation, whereas Nemotron-SFT-OpenCode-v1 covers agent interactions.

Mathematics uses 256 AoPS examples from the `high_part00` split of Nemotron-Math-v2, which pairs problems from Art of Problem Solving with model-generated solution traces. The allocation retains 160 examples from the earlier calibration pool and adds 96 separately selected examples from the same source. High denotes the teacher's reasoning mode, not problem difficulty. Science combines 176 chemistry reasoning questions and 80 multiple-choice questions from Nemotron-Science-v1. Chinese-STEM uses translated mathematics, code, and STEM examples from Nemotron-SFT-Multilingual-v1 (86/85/85), which translates prompts and final answers rather than all reasoning traces. Instruction following draws 256 examples from a Dolci constraint-following pool containing Instruct, Think, and Persona examples (Team Olmo et al., 2025), rather than from the Instruct mixture alone. Tool calling uses 256 multi-turn agent–tool conversations from Nemotron-Agentic-v1. The world-knowledge allocation contains 256 OpenHermes-2.5 examples (Teknium, 2023), drawn

Table 3: RAZORCAL domain allocations and calibration sources. Counts denote source examples, not tokens.

Domain	Samples	Source
Coding	**512**	Dolci-Instruct-SFT and Nemotron competitive programming, SWE, and OpenCode
Mathematics	256	Nemotron-Math-v2, high-reasoning AoPS subset
Science	256	Nemotron-Science-v1
Chinese-STEM	256	Nemotron-SFT-Multilingual-v1
Instruction following	256	Dolci constraint-following mixture
Tool calling	256	Nemotron-Agentic-v1
World knowledge	256	OpenHermes-2.5

from the cleaned, non-thinking subset in SmolTalk2 (Hugging Face Smol Models Research, 2025). OpenHermes is a general-purpose synthetic instruction corpus, not a dedicated knowledge test. These quotas balance source examples rather than token counts.

Packing and scored positions. For the diagnostic collection, calibration examples are shuffled with seed 0 and rendered with thinking disabled. Each example fills the remaining space in a 4,096-token row, with any overflow discarded rather than continued in the next row. A masked separator position is left between examples when space permits. The packing mask does not isolate conversations within a row, so later examples can attend to earlier nonmasked positions. Calibration statistics include all nonpadding conversation positions, including prompts and role headers, unlike the assistant-only fidelity targets in Appendix B.2.

Scored coverage. This collection scores 192 packed sequences rather than the full source pool. Qwen3.6-35B-A3B uses 785,703 valid scored tokens from 921 unique source examples, while GLM-4.7-Flash uses 785,632 tokens from 992 examples. These counts record examples touched by packing, not necessarily complete conversations. The same row capacity therefore gives different valid-token counts and source-example coverage under the two tokenizers. These settings describe the diagnostic collection, not an inferred configuration for every external benchmark checkpoint.

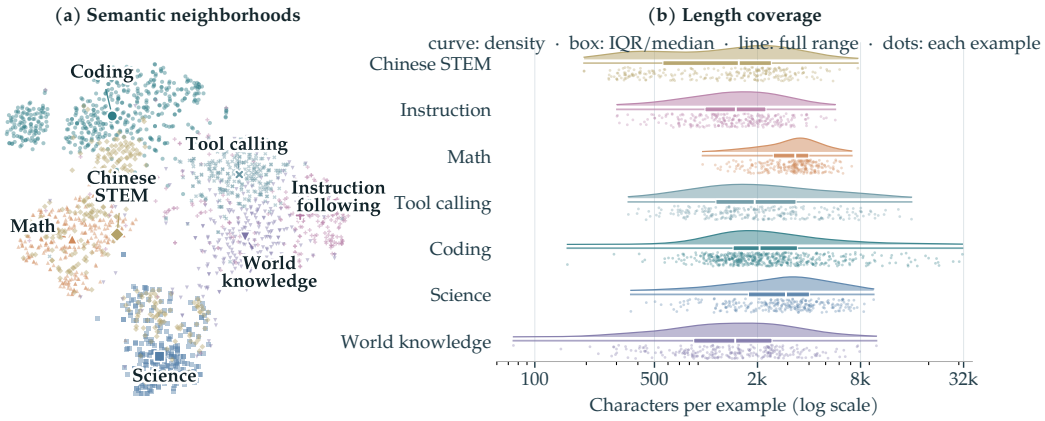


Figure 5: **Overview of the RAZORCAL calibration pool.** (a) Cosine t-SNE projection of Qwen3-Embedding-8B representations of the rendered conversations, with color and marker shape encoding domain and labels marking domain medians. (b) Per-domain character-length coverage over the full observed range. The curve is a kernel-density estimate in log-character space, the bar is the interquartile range with the median marked, the thin line spans minimum to maximum, and each dot is one of the 2,048 examples. Domain counts are given in Table 3. Generated from the calibration data and cached embeddings, not from benchmark outputs.

B.2 FIDELITY EVALUATION DATA

Reverse KL and reference-token perplexity use a pool of 8,192 conversations, with 1,024 examples on each of eight task axes. These are reference-response comparisons, not benchmark answer scores. Four axes cover calibration task types.

Mathematics. Mathematical problem-solving responses from the `medium` split of Nemotron-Math-v2. Medium denotes the teacher's reasoning mode, not problem difficulty.

Code. Python solutions from the `competitive_coding_python` split of Nemotron-SFT-Competitive-Programming-v2.

Instruction following. Constraint-following conversations from the `chat_if` split of Nemotron-Instruction-Following-Chat-v1, selected by its instruction-following capability label.

Tools. Agent–tool interaction trajectories from the `tool_calling` split of Nemotron-Agentic-v1.

The other four axes hold out task types.

Chat. Examples carrying the chat capability label in the same Chat-v1 `chat_if` split.

Creative writing. Examples selected by prompt keywords from the `reasoning_off` split of Nemotron-SFT-Instruction-Following-Chat-v2. Creative writing is our selection, not an official split.

Safety. Reference safety-aligned responses from the `train` split of Nemotron-SFT-Safety-v1.

SQL. Examples from the `text_to_sql` split of the competitive-programming source, separate from its Python subset, pairing natural-language tasks and database schemas with SQL. These are the sources' official split names under their default configurations, not train/test assignments introduced by this study. ID/OOD denotes task-type coverage, not disjoint source datasets or a formal distribution-shift guarantee. Both groups contain synthetic data.

For each axis, we retain the first 1,024 eligible examples in source order and shuffle the combined pool with seed 0. Eligibility uses a 13,200-character window budget, as a proxy for 4,000 tokens, and at least 726 characters of assistant content. Neither threshold guarantees a model-specific token length. Preprocessing normalizes message content, decodes encoded safety responses, and preserves tool calls and schemas. Conversations end on an assistant response. Plain-text assistant turns may be truncated, but structured tool calls remain intact. Creative-writing eligibility additionally uses a prompt keyword filter.

Evaluation packs examples into 4,096-token rows and scores assistant positions, excluding role headers and turn terminators. Chat rendering disables thinking and omits separately supplied reasoning fields. Fidelity is measured on fixed reference prefixes rather than generated reasoning trajectories. The routing study uses six batches per axis (Appendix E.3). Its evaluated batches, the full fidelity pool, the nine-task benchmark, and the diversity set are distinct collections. The fidelity pool has no exact-message overlap with calibration, although near-duplicates and benchmark contamination have not been ruled out.

B.3 EVALUATION CONFIGURATION

Table 4 summarizes the four main backbones. Qwen3.6-35B-A3B uses the hybrid attention–MoE architecture described for Qwen3.5 (Qwen Team, 2026). The GLM and DeepSeek model families are described in their technical reports (GLM-4.5 Team, 2025; DeepSeek-AI, 2026). The configurations below refer to the evaluated checkpoints. Hy3 is the public release (Tencent Hy Team, 2026). The architecture and routing settings describe the language-model decoder. Layer counts exclude auxiliary prediction modules. GLM-4.7-Flash and Hy3 each begin with one dense feed-forward layer. All 43 DeepSeek-V4-Flash-0731 layers are MoE layers, of which three use fixed-hash routing and the other 40 use learned-router selection. Its score transformation is $\sqrt{\mathrm{softplus}(z)}$. The nominal parameter counts exclude auxiliary MTP and DSpark modules.

Both removal ratios retain the original top-k and shared experts. Calibration uses the 2,048-example source pool, while the scored coverage in Appendix B.1 refers specifically to the Qwen3.6-35B-A3B and GLM-4.7-Flash diagnostic collection. For downstream benchmarks, we use temperature 0, except for SWE-bench Verified, where we use 0.7. AIME'26 reports avg@8, and the other eight benchmarks

report avg@3. For each benchmark, sampling parameters are shared across all four backbones. The maximum input length is 224K tokens for GLM-4.7-Flash and Qwen3.6-35B-A3B, and 128K tokens for DeepSeek-V4-Flash-0731 and Hy3. All four use a 32K-token output limit. These are evaluation limits, not the models' maximum context capacities. The Qwen3.6-35B-A3B and GLM-4.7-Flash benchmark settings enable thinking. Fidelity and response diversity use their own protocols, with generation analysis using temperature 0.7. Cross-study comparability is discussed in Appendix B.5.

Within a study, scoring-rule comparisons match calibration inputs and layerwise budgets, but this does not imply shared calibration inputs or retained expert sets across studies. Downstream and component comparisons do not repeat calibration and pruning runs, so their reported estimates do not measure variability over those runs, and the diagnostic bootstrap intervals do not capture calibration-draw variability.

B.4 Metric definitions

Let p_t and q_t be the unpruned and pruned next-token distributions on the same reference prefix. Reverse KL is

$$D_{\mathrm{KL}}(q_t \| p_t) = \sum_v q_t(v) \log \frac{q_t(v)}{p_t(v)}. \tag{13}$$

The main fidelity sweep weights assistant tokens within each evaluation axis and balances the eight axes equally.

For reference token y_t, define $\ell_p(t) = -\log p_t(y_t)$ and $\ell_q(t) = -\log q_t(y_t)$. Averaging these losses under the specified evaluation weights gives reference-token NLL and

$$\begin{aligned} \mathrm{PPL}_p &= \exp(\mathrm{NLL}_p), \qquad \mathrm{PPL}_q = \exp(\mathrm{NLL}_q), \\ \Delta\mathrm{NLL} &= \mathrm{NLL}_q - \mathrm{NLL}_p, \qquad \frac{\mathrm{PPL}_q}{\mathrm{PPL}_p} - 1 = \exp(\Delta\mathrm{NLL}) - 1. \end{aligned} \tag{14}$$

The final quantity is excess perplexity. These likelihood metrics use observed reference tokens, not samples from p_t or q_t. Negative ΔNLL indicates higher reference-token likelihood under the pruned model, not necessarily better downstream capability.

For normalized selected routing weights and $k > 1$, routing entropy is

$$H_r(x) = -\frac{\sum_{j \in S(x)} w_j(x) \log w_j(x)}{\log k}. \tag{15}$$

Smaller H_r means more concentrated routing. Predictive entropy instead summarizes the vocabulary distribution,

$$H(p_t) = -\sum_v p_t(v) \log p_t(v), \qquad \Delta H_t = H(q_t) - H(p_t). \tag{16}$$

Absolute entropy drift averages predictive entropies within routing cells and across valid layers, takes the absolute pruned–original difference in each batch–decile cell, and then averages these differences. Taking the absolute value after cell aggregation distinguishes this measure from the mean per-token $|\Delta H_t|$. Predictive entropy also differs from reference-token NLL, so its exponential is not the PPL reported here. Appendix E specifies the sampling units and uncertainty estimates for each output analysis.

B.5 Downstream benchmark protocol and scope

The benchmark evaluates four backbones at two removal ratios each. Table 1 covers GLM-4.7-Flash and Qwen3.6-35B-A3B, and Table 2 covers DeepSeek-V4-Flash-0731 and Hy3. The full method comparison includes Frequency, EAN, REAP, and all three residual scores on GLM-4.7-Flash and Qwen3.6-35B-A3B. RCS, RCS-LOO, and RAZOR use conditional RMS. DeepSeek-V4-Flash-0731 and Hy3 compare the three residual scores against their unpruned references only. REAP, Frequency, and EAN were not evaluated in the nine-task benchmark on these two backbones. These models therefore contribute to the RCS-LOO-relative and unpruned-relative benchmark comparisons, but not to any REAP-relative benchmark range or win count.

Table 4: Architectures and pruning settings of the four main backbones. Parameter counts are nominal backbone sizes, and expert counts are per MoE layer.

	Qwen3.6-35B-A3B	GLM-4.7-Flash	DeepSeek-V4-Flash-0731	Hy3
Architecture				
Total parameters	35B	30B	284B	295B
Active parameters per token	∼3B	∼3B	∼13B	∼21B
Decoder layers	40	47	43	80
MoE layers	40	46	43	79
Routed experts	256	64	256	192
Expert intermediate width	512	1,536	2,048	1,536
Shared experts	1	1	1	1
Routing				
Top-k	8	4	6	8
Router scores	softmax	sigmoid	sqrtsoftplus	sigmoid
Routed scale λ	1.0	1.8	1.5	2.826
Fixed-hash layers	0	0	3	0
Pruning				
Pruned fractions		25% and 50%		
Retained routed experts (25%)	192	48	192	144
Retained routed experts (50%)	128	32	128	96

Baseline scores. For $n_i = |\mathcal{D}_i| > 0$ routed calibration tokens, the baseline scores are, up to common positive layer scales that do not affect ranking,

$$s_i^{\text{Frequency}} = n_i, \qquad s_i^{\text{EAN}} = \frac{1}{n_i} \sum_{x \in \mathcal{D}_i} \|f_i(x)\|_2, \qquad s_i^{\text{REAP}} = \frac{1}{n_i} \sum_{x \in \mathcal{D}_i} w_i(x) \|f_i(x)\|_2. \quad (17)$$

Frequency counts hard routing selections, as in frequency-based pruning (Muzio et al., 2024; Jaiswal et al., 2025). EAN is our conditional-mean adaptation of activation-norm pruning (Jaiswal et al., 2025; Lasby et al., 2026). We divide by each expert's routed-token count, whereas the accumulated-norm EAN in REAP's baseline definition does not. This expert-specific normalization is not a common rescaling and can change the ranking. REAP follows the routed-token conditional mean in its original definition (Lasby et al., 2026). All three scores are computed without recovery training, and zero-observation experts receive zero score.

Mathematics and instruction following. The nine-task suite covers six capability categories. AIME'26 uses the 2026 American Invitational Mathematics Examination (Mathematical Association of America, n.d.) to assess competition mathematics. IFEval (Zhou et al., 2023) checks compliance with programmatically verifiable instructions. IFBench (Pyatkin et al., 2025) extends this evaluation to additional verifiable constraints, testing instruction-following generalization beyond IFEval.

Knowledge, tools, and long context. SuperGPQA (Du et al., 2025) evaluates graduate-level knowledge across 285 disciplines. The Berkeley Function Calling Leaderboard, version 4 (BFCL v4, Patil et al., 2025; Mao et al., 2025), evaluates function calling and agentic tool use. LongBench v2 (Bai et al., 2025) tests comprehension and reasoning over long contexts.

Coding tasks and versions. The three coding tasks cover function-level generation, competition programming, and repository-level issue resolution. HumanEval+ (Liu et al., 2023) uses EvalPlus's expanded tests to assess the functional correctness of generated Python functions. LiveCodeBench (Jain et al., 2025) draws from time-indexed programming competitions. The main benchmark uses its 2026 evaluation-time latest version, abbreviated LCB in the tables, which is distinct from the 175-question subset used for the generation diagnostics (Appendix E.5). SWE-bench Verified (Jimenez et al., 2024; Chowdhury et al., 2024), abbreviated SWE, is the 500-issue subset of SWE-bench screened by software developers for well-specified issues and appropriately scoped tests. The reported benchmark designation is Verified for all four backbones, but the retained score summaries

do not independently establish the evaluated issue sets. Scores report the percentage of resolved issues averaged over three evaluations.

Score aggregation. Instruction-following and coding averages give equal weight to their two and three tasks, respectively, while Overall weights all nine equally. Averages and paired macro gaps are computed from task scores rather than rounded displayed means. The recorded task-score table yields a 33/0/3 win/tie/loss count for RCS-LOO against REAP and 36/0/0 for RAZOR across GLM-4.7-Flash and Qwen3.6-35B-A3B at two budgets. These counts are arithmetic summaries of the recorded task scores, but the current artifact does not include raw evaluator reports for an independent rederivation. Both sets of 36 comparisons reuse the same task suite rather than provide independent replications, and neither extends to the two backbones without a REAP benchmark run. AIME'26 reports avg@8, and the other eight benchmarks report avg@3. The benchmark comparisons do not repeat calibration and pruning runs and report no task-level uncertainty. The three residual criteria and REAP use matched RAZORCAL inputs, while REAP* is an external GLM-4.7-Flash checkpoint calibrated with 24,576 samples.

Additional task contrasts. On Hy3, RAZOR scores 63.4 and 59.5 at 25% and 50% removal, respectively, versus 63.7 unpruned. On DeepSeek-V4-Flash-0731, the corresponding scores are 61.0 and 55.6 versus 59.2 unpruned (Table 2). The macro ordering is neither uniform across tasks nor monotonic across scoring refinements. At 25% on Qwen3.6-35B-A3B, RCS slightly exceeds RCS-LOO in macro average (61.0 versus 60.8). At 50% on GLM-4.7-Flash, EAN exceeds RAZOR on LCB (33.9 versus 29.8).

For RCS-LOO versus REAP, both benchmarked backbones show larger macro gaps at 50% than at 25% (0.80 to 3.01 on GLM-4.7-Flash and 1.02 to 1.96 on Qwen3.6-35B-A3B). Whether this widening extends to DeepSeek-V4-Flash-0731 and Hy3 is untested, since REAP was not benchmarked there. On DeepSeek-V4-Flash-0731 at 25%, RCS-LOO gains on SWE (12.4 to 18.8) but loses on HE+ (91.5 to 89.6) relative to RCS, the nearest measured criterion on that backbone. At 50%, SWE suffers the largest proportional loss from the original for RCS-LOO on both Hy3 and DeepSeek-V4-Flash-0731. Thus even favorable macro comparisons coexist with capability-specific trade-offs.

Comparability. Using the same residual criterion in the benchmark and diagnostic studies does not establish that their retained expert sets or calibration tokens are identical. This applies to all three residual criteria. For DeepSeek-V4-Flash-0731, fixed-hash selections are matched across criteria in the routing study, but this matching is not established for every benchmark checkpoint. Consequently, the full DeepSeek-V4-Flash-0731 benchmark gap cannot be assigned solely to learned-router saliency.

The "2026 latest" designation does not specify an immutable LiveCodeBench snapshot or contest-date range. Exact evaluator revisions, benchmark-specific seeds, the AIME'26 data redistribution, and the SWE agent scaffold and tool budget are also unavailable. Input and output limits and shared sampling settings are specified in Appendix B.3. Public model-card settings cannot fill the remaining experiment-specific gaps, which limit exact reproduction and cross-study attribution.

C SCORING PROXIES AND EXPERT SELECTION

C.1 SCORING CRITERIA AND BUDGET COMPARISONS

Table 5 lists the three consensus-residual scoring rules and the magnitude baseline REAP-RMS, all with conditional RMS, alongside their experimental coverage. The comparisons below separate the token-level quantity from its aggregation across routed tokens.

The fixed-support deletion identity determines the combination of consensus residual and LOO factor, but neither the aggregation rule nor the performance of the selected checkpoint. All mean and RMS criteria condition on tokens routed to the expert. Our EAN adaptation averages $\|f_i\|_2$, and REAP averages $w_i\|f_i\|_2$. Appendix B.5 gives the source definitions and the EAN normalization difference.

RCS uses the routing-weighted consensus residual. RCS-LOO includes survivor renormalization, and RCS-Refill additionally models the promoted expert. The token quantities of RCS-LOO and RCS-Refill are exact under their respective deletion assumptions, but the three rules do not form a guaranteed performance hierarchy. REAP-RMS instead uses the output norm and differs from

Table 5: **Scoring criteria and experimental coverage.**

Instance	Token quantity	Evidence in this paper
RCS-Refill (RAZOR)	$\frac{\|w_i r_i - w_r r_r\|_2}{1 - w_i + w_r}$	Nine-task benchmarks on all four backbones (Tables 1, 2). Reverse KL (Figure 2, Table 6), score distributions (Table 7), and refill mechanism (Table 8). Mean–RMS reduction and factorial comparisons (Figure 6, Table 9). Per-axis, routing-stratified and PPL fidelity, selection probes, and Qwen3.6-35B-A3B response behavior (Table 10).
RCS-LOO	$\frac{w_i \|r_i\|_2}{1 - w_i}$	Nine-task benchmarks on four backbones (Tables 1, 2). Output-fidelity and diversity figures, response behavior, reverse KL, score distributions, and calibration-budget and corpus diagnostics.
RCS	$w_i \|r_i\|_2$	Nine-task benchmarks on all four backbones (Tables 1, 2). Reverse KL (Figure 2, Table 6) and the reduction axis (Figure 6).
REAP-RMS	$w_i \|f_i\|_2$	Reverse KL in the separate factorial collection (Appendix C.5). No benchmarks.

REAP solely in replacing conditional mean with RMS. Its results belong to the factorial collection in Appendix C.5, where Table 9 crosses f_i versus $f_i - c$ and w_i versus $w_i/(1 - w_i)$ across the available sum, mean, and RMS reductions, with refill damage as an additional quantity.

Table 6 compares REAP, RCS, RCS-LOO, and RAZOR under the same checkpoint-evaluation protocol. Moving from REAP to RCS lowers KL on both models, but changes both the reference and mean-to-RMS aggregation, so it does not isolate consensus subtraction. The matched mean-aggregation contrast appears in Appendix C.6. Adding the LOO denominator raises KL on Qwen3.6-35B-A3B, while adding refill lowers it in three of four settings. These estimates pool tokens within each axis and weight the eight axes equally. The Δ vs REAP column gives RCS-LOO's relative change from REAP, while refill gain gives RAZOR's relative reduction from RCS-LOO. Negative values favor RCS-LOO in both columns, and percentages use unrounded KL.

C.2 RMS AGGREGATION AND SCORE DISTRIBUTIONS

For routed calibration tokens $\mathcal{D}_i = \{x \in \mathcal{D} : i \in S(x)\}$, RAZOR uses the conditional root mean square of refill damage,

$$s_i = \sqrt{\frac{1}{|\mathcal{D}_i|} \sum_{x \in \mathcal{D}_i} \left(\frac{\|w_i(x) r_i(x) - w_r(x) r_r(x)\|_2}{1 - w_i(x) + w_r(x)} \right)^2}. \tag{18}$$

This definition applies when $|\mathcal{D}_i| > 0$, with unseen experts handled as specified in Appendix A.4. The token damage follows from Proposition 2, but conditioning and RMS are aggregation choices, not consequences of that identity. The same aggregation is applied to $w_i \|r_i\|_2$ for RCS and δ_i^{loo} for RCS-LOO.

RMS gives larger observed shifts more weight than the conditional mean. With empirical conditional mean μ_i and population standard deviation σ_i of token-level damage over $\mathcal{D}_i$,

$$s_i = \sqrt{\mu_i^2 + \sigma_i^2} = \mu_i \sqrt{1 + \mathrm{CV}_i^2}, \qquad \mathrm{CV}_i = \sigma_i / \mu_i \quad (\mu_i > 0). \tag{19}$$

For $\mu_i < \mu_j$, RMS ranks i above j exactly when $\sigma_i^2 - \sigma_j^2 > \mu_j^2 - \mu_i^2$. Such reversals change the retained set only across the budget boundary. This preference for variable damage motivates RMS as an empirical design choice rather than a guarantee of better pruning. Appendix C.6 tests that choice by comparing mean and RMS while holding the token quantity fixed.

We recover each expert's mean and standard deviation from its routed-token count and the accumulated first and second moments of token-level damage, without retaining per-token outputs. Both fixed-support and refill instances have these accumulators, allowing Table 7 to compare RCS-LOO and

Table 6: **Reverse KL across scoring rules in nats (lower is better).**

Removal	REAP	RCS	RCS-LOO	RAZOR	Δ vs REAP	Refill gain
Qwen3.6-35B-A3B						
25%	0.07461	0.06669	0.07141	0.06872	−4.3%	+3.8%
50%	0.45654	0.40337	0.45108	0.44397	−1.2%	+1.6%
GLM-4.7-Flash						
25%	0.22850	0.20481	0.20351	0.20459	−10.9%	−0.5%
50%	1.01690	0.92310	0.87103	0.85457	−14.3%	+1.9%

RAZOR. Equal-expert quantiles cover all observed experts without a count threshold, comprising 10,240 on Qwen3.6-35B-A3B and 2,944 on GLM-4.7-Flash. The CV panel of Figure 4 uses the same experts. Within-layer dispersion is $\mathrm{median}_\ell[\mathrm{std}_i(\mathrm{CV}_{\ell i})/\,\mathrm{mean}_i(\mathrm{CV}_{\ell i})]$, using population standard deviations.

The median fixed-support premium is 1.189 on Qwen3.6-35B-A3B and 1.155 on GLM-4.7-Flash. Its variation across experts, rather than its absolute size, can change a ranking. We quantify the resulting boundary swaps by comparing direct top-B mean and RMS rankings on the same experts and averaging swap counts equally across layers. Each entry reports the RMS-only retained count over the retained-set size. Qwen3.6-35B-A3B has more exchanges at both budgets. Under refill, the median premium is 1.166 on Qwen3.6-35B-A3B and 1.153 on GLM-4.7-Flash, and the swap counts differ from the fixed-support counts by less than one expert per layer. These summaries show similar relative variability and mean–RMS selection differences in the two calibration collections. They do not imply equal damage magnitudes or equal second moments across counterfactuals.

A boundary-crossing example comes from Qwen3.6-35B-A3B layer 0, the first layer with a mean–RMS disagreement at 25% removal in the 4,096-token-row diagnostic. We choose the opposite-status pair with the closest conditional means, without using downstream outcomes. Expert 208 has mean/CV/RMS 0.03323/0.741/0.04136 over 32,041 routed tokens, compared with 0.03359/0.637/0.03982 over 23,766 tokens for expert 38. Their mean-to-RMS ranks move from 194 to 190 and from 189 to 201, respectively, crossing the 192-expert retention boundary. The pair illustrates how variability changes selection near a small mean-score margin. It does not measure the utility of the exchange or verify its presence in the benchmark checkpoints.

C.3 CONDITIONAL RMS, MEAN, AND CORPUS-SUM REDUCTIONS

Conditioning on expert use is distinct from squaring the damage. Let g_i be any non-negative token quantity with conditional RMS s_i, let $\pi_i = |\mathcal{D}_i|/|\mathcal{D}|$, and set g_i to zero on tokens not routed to i. The empirical average squared damage over *all* calibration tokens is

$$\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \mathbf{1}\{i \in S(x)\} g_i(x)^2 = \pi_i s_i^2. \tag{20}$$

An all-token RMS criterion would therefore rank $\sqrt{\pi_i}s_i$. Ranking s_i instead measures severity conditional on use, without discounting infrequent but severe observed deletions. It neither minimizes total calibration error nor protects unobserved capabilities. Comparing conditional mean and RMS tests the second moment, whereas comparing conditional and all-token reductions tests the treatment of exposure.

To isolate aggregation, we fix the RCS token quantity $g_i(t) = w_i(t)\|f_i(t) - c(t)\|_2$ and compare three reductions of the observed damage at four pruning budgets,

$$\begin{aligned} &\text{conditional RMS } \sqrt{\tfrac{1}{n_i}\textstyle\sum_{t\in\mathcal{D}_i} g_i(t)^2}, && \text{conditional mean } \tfrac{1}{n_i}\textstyle\sum_{t\in\mathcal{D}_i} g_i(t), \\ &\text{corpus sum } \textstyle\sum_{t\in\mathcal{D}} g_i(t)^2 = |\mathcal{D}|\,\pi_i s_i^2, \end{aligned} \tag{21}$$

where $\mathcal{D}_i$ contains tokens routed to expert i, $n_i = |\mathcal{D}_i|$, and $g_i(t) = 0$ elsewhere. The corpus sum differs from the all-token mean squared damage in Eq. 20 only by the common factor $|\mathcal{D}|$, so they induce the same ranking. We exclude selection frequency n_i because it discards g_i entirely. Including

Table 7: Damage variability and mean–RMS selection differences for RCS-LOO and RAZOR.

	Qwen3.6-35B-A3B		GLM-4.7-Flash	
	RCS-LOO	**RAZOR**	**RCS-LOO**	**RAZOR**
Sample				
Scored tokens	785,703		785,632	
Coefficient of variation				
CV median	0.643	0.599	0.577	0.574
CV p_5 / p_{95}	0.370 / 1.255	0.373 / 1.184	0.350 / 1.002	0.362 / 0.978
Within-layer dispersion of CV	0.317	0.305	0.221	0.211
RMS premium $\sqrt{1+\mathrm{CV}^2}$				
Median	1.189	1.166	1.155	1.153
p_{95}	1.605	1.550	1.416	1.398
Boundary swaps				
25% removal	7.95 / 192	7.53 / 192	1.39 / 48	1.33 / 48
50% removal	10.80 / 128	11.00 / 128	1.70 / 32	1.85 / 32

it would conflate a change in token quantity with a change in aggregation. Table 1 compares frequency as a criterion in its own right.

For the RCS quantity, reverse KL follows the strict ordering RMS $<$ mean $<$ sum in all eight model–budget settings (Figure 6). Under this protocol, conditioning on use gives better checkpoint fidelity than weighting squared severity by exposure. This empirical result does not invalidate the all-token error quantity, since the two scores target different aspects of deletion damage. The size of the aggregation effect depends on the backbone. Between the extreme reductions, KL spans $2.4\times$ and $1.7\times$ on Qwen3.6-35B-A3B and $1.15\times$ and $1.05\times$ on GLM-4.7-Flash at 25% and 50% removal. The corresponding spans among REAP, RCS, and RCS-LOO in Table 6 are $1.12\times$, $1.13\times$, $1.12\times$, and $1.17\times$. The reduction span is therefore larger on Qwen3.6-35B-A3B, comparable on GLM-4.7-Flash at 25%, and smaller there at 50%. The criterion span changes several scoring choices, so it does not isolate the reference.

These fidelity comparisons use one local proxy and equal layerwise budgets, evaluated on the same four shards as Table 6. The four leave-one-out and refill curves apply the two conditional reductions to their respective residuals on those shards. For refill damage, RMS lowers reverse KL relative to the conditional mean at all four GLM-4.7-Flash budgets, by 4.4%–11.4%, and at 12.5% and 25% on Qwen3.6-35B-A3B. It raises KL at 37.5% and 50% on Qwen3.6-35B-A3B (0.20744 versus 0.18874, and 0.44397 versus 0.44143). Thus RMS aggregation of refill damage improves checkpoint fidelity in six of eight settings. Applied to the leave-one-out residual, it lowers KL in seven of eight settings, failing only at 37.5% on Qwen3.6-35B-A3B (0.20504 versus 0.19196). This four-budget collection contains no corpus-sum arm for either residual, unlike the separate two-budget factorial study (Appendix C.5). No downstream accuracy comparison accompanies this reduction study.

C.4 WHAT THE REFILL TERM CHANGES

Table 8 describes how refill changes expert scores and token-level damage profiles, using the same accumulated calibration moments as Table 7. Router averages are token-weighted over routed events, whereas score ratios and token-axis cosines use unweighted expert quantiles. The routed-event-weighted mean $w_i = 1/k$ is a renormalization consistency check, since each token's k selected weights sum to one.

Score scale and residual geometry. The median refill-to-fixed-support score ratio corresponds to a 2.81% increase on Qwen3.6-35B-A3B and a 2.56% decrease on GLM-4.7-Flash. Refill raises scores for 85.2% of Qwen3.6-35B-A3B experts and 14.3% of GLM-4.7-Flash experts. Substituting the reported mean routing weights into $(1-w_i)/(1-w_i+w_r)$ gives 0.922 and 0.850, respectively. These values illustrate the denominator's attenuation at average weights. They are neither token-averaged factors nor a decomposition of the median score ratio.

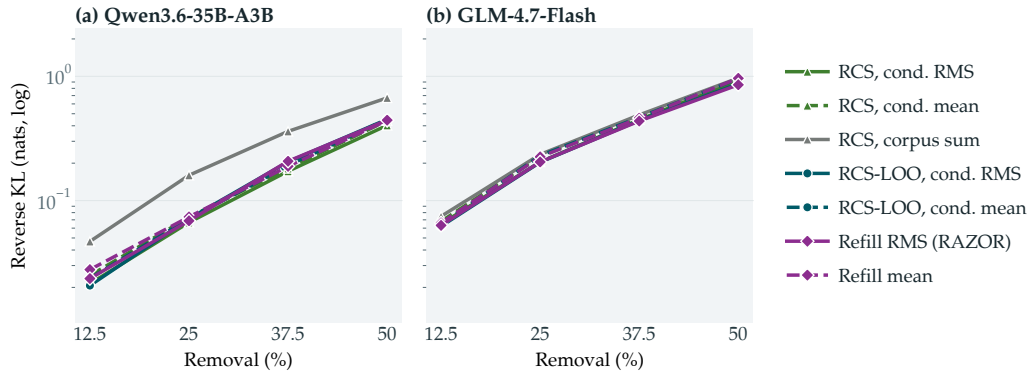


Figure 6: **Reverse KL under different calibration reductions (lower is better).** The RCS quantity uses conditional RMS, conditional mean, and corpus sum. The RCS-LOO and RCS-Refill quantities each use the two conditional reductions on the same shards, with no corpus-sum arm in this four-budget collection. Selection frequency is excluded because it discards the token quantity rather than reducing it.

The mean residual cosines are 0.108 on Qwen3.6-35B-A3B and 0.103 on GLM-4.7-Flash. A low positive cosine is compatible with either an increase or a decrease in the numerator, depending on the relative residual magnitude (Eq. 11). The available summaries do not retain the promoted residual norms needed to reconstruct ρ_i, so they do not resolve that contribution. Moreover, the score ratios summarize experts whereas the routing statistics summarize events, so combining their averages cannot recover the nonlinear factors.

Damage-profile similarity and selection. Treating δ_i^{loo} and δ_i as vectors over each expert's routed tokens, their cosine has median 0.9942 on Qwen3.6-35B-A3B and 0.9952 on GLM-4.7-Flash, with fifth percentiles 0.9630 and 0.9715. Most profiles are therefore close in direction, although the minima of 0.7062 and 0.5440 show substantial changes for some experts. Even nearly proportional profiles can change expert rankings if their scale factors differ. These statistics describe local score changes, not their contribution to checkpoint quality. Despite the opposite median score shifts on the two models, refill lowers KL at both Qwen3.6-35B-A3B budgets and only one GLM-4.7-Flash budget.

C.5 Reference, routing factor, and aggregation

Table 9 reports fourteen scoring configurations for two models at 25% and 50% removal. Its core grid crosses two output references with two routing factors. Rows within each block compare token quantities, while blocks compare reductions of the same quantity. The conditional mean of $w_i\|f_i\|_2$ is REAP, and the conditional RMS of $\frac{w_i}{1-w_i}\|f_i - c\|_2$ is RCS-LOO. Adding refill damage gives five token quantities across the corpus-sum $\sum_t g_i(t)^2$, conditional-mean, and conditional-RMS blocks. No corpus-sum result is available for $\frac{w_i}{1-w_i}\|f_i\|_2$.

The refill rows require a separate matching distinction. Their corpus-sum and RMS results share the grid's four evaluation shards. The conditional-mean row instead comes from the main sweep in Figure 6, which differs in one shard. Comparisons involving that row are therefore cross-collection contrasts, unlike the matched comparisons among the other thirteen rows.

Evaluation and aggregation. Within each model, thirteen of the fourteen rows share four evaluation shards with matched batch-domain sequences and token counts. The mean-refill row uses the separate main-sweep shards noted above. Let $L_{s,d}$ denote mean reverse KL in shard s and domain d, with $n_{s,d}$ scored tokens. The estimator pools tokens within domains and then weights the eight domains

Table 8: **Calibration statistics of routing weights and refill-induced score changes.**

Quantity	Qwen3.6-35B-A3B	GLM-4.7-Flash
Routing geometry (token-weighted)		
Selected weight w_i $(= 1/k)$	0.1250	0.2500
Promoted pseudo-weight w_r	0.0740	0.1325
$\cos(r_i, r_r)$	0.1080	0.1034
Score ratio $s_i^{\mathrm{rf}}/s_i^{\mathrm{loo}}$ *(per expert)*		
Median	1.0281	0.9744
Experts with ratio > 1	85.2%	14.3%
Token-axis $\cos(\delta_i^{\mathrm{loo}}, \delta_i)$ *(per expert)*		
Median	0.9942	0.9952
p_5	0.9630	0.9715
Minimum	0.7062	0.5440
Collection		
Top-k / experts per layer	8 / 256	4 / 64
Layers / scored experts	40 / 10,240	46 / 2,944
Routed token positions	785,703	785,632

Table 9: **Fourteen scoring configurations varying reference, routing factor, and aggregation.** Reverse KL is in nats at 25% and 50% removal, with lower values better. Thirteen rows share four evaluation shards. The conditional-mean refill row comes from the main sweep in Figure 6, which differs in one shard. Comparisons involving this row are not fully matched. Sum denotes $\sum_t g_i(t)^2$. No corpus-sum result is available for $\frac{w_i}{1-w_i}\|f_i\|_2$. Refill damage with RMS is RAZOR.

Token quantity	Aggregate	Qwen3.6-35B-A3B		GLM-4.7-Flash	
		25%	50%	25%	50%
$w_i\|f_i\|_2$	Sum	0.14712	0.67725	0.22309	0.89425
$w_i\|f_i - c\|_2$	Sum	0.16005	0.67173	0.23513	0.96880
$\frac{w_i}{1-w_i}\|f_i - c\|_2$	Sum	0.17659	0.69294	0.23339	0.91110
$\frac{\|w_i r_i - w_r r_r\|_2}{1-w_i+w_r}$	Sum	0.18798	0.70264	0.23176	0.88086
$w_i\|f_i\|_2$	Mean	0.07469	0.45691	0.22852	1.01692
$\frac{w_i}{1-w_i}\|f_i\|_2$	Mean	0.07882	0.48975	0.22922	0.97162
$w_i\|f_i - c\|_2$	Mean	0.07357	0.44075	0.22804	0.94483
$\frac{w_i}{1-w_i}\|f_i - c\|_2$	Mean	0.07151	0.45341	0.23017	0.95748
$\frac{\|w_i r_i - w_r r_r\|_2}{1-w_i+w_r}$	Mean	0.07386	0.44143	0.22481	0.96456
$w_i\|f_i\|_2$	RMS	0.06869	0.43948	0.20813	0.86866
$\frac{w_i}{1-w_i}\|f_i\|_2$	RMS	0.09535	0.64640	0.20865	0.90296
$w_i\|f_i - c\|_2$	RMS	0.06674	0.40355	0.20483	0.92318
$\frac{w_i}{1-w_i}\|f_i - c\|_2$	RMS	0.07144	0.45134	0.20353	0.87107
$\frac{\|w_i r_i - w_r r_r\|_2}{1-w_i+w_r}$	RMS	0.06874	0.44409	0.20461	0.85460

equally,

$$L = \frac{1}{8}\sum_{d=1}^{8} \frac{\sum_{s=0}^{3} n_{s,d} L_{s,d}}{\sum_{s=0}^{3} n_{s,d}}. \tag{22}$$

The entries are shard-based point estimates, not estimates of variability across independent runs.

Relation to the matched component study. This collection is not identical to the main sweep. For their shared criteria, per-domain reverse KL agrees exactly in three of four shards. The remaining Qwen3.6-35B-A3B shard has 120 batches and 529,093 tokens here, compared with 119 batches and

532,523 tokens in the main sweep. Both GLM-4.7-Flash collections have 121 batches in that shard, but contain 528,964 and 530,090 tokens, respectively. We therefore analyze the collections separately. Their overlap also precludes treating these results as an independent replication.

Component effects. Even with mean aggregation fixed, the effect of the routing factor depends on the output reference. On GLM-4.7-Flash at 50% removal, replacing w_i with $w_i/(1-w_i)$ lowers reverse KL from 1.01692 to 0.97162 for output magnitudes, but raises it from 0.94483 to 0.95748 for consensus residuals. For fixed-support deletion damage, RMS lowers reverse KL relative to the conditional mean in all four displayed model–budget settings, although the Qwen3.6-35B-A3B differences are small.

Reduction ordering depends on the token quantity. The grid reproduces the RCS ordering RMS $<$ mean $<$ sum from Section C.3 at both displayed budgets on both models, but this ordering does not hold for every quantity. Conditional RMS has the lowest KL in 15 of the 16 cells with all three reductions, including four cross-collection comparisons involving mean refill. The exception is refill at 50% removal on Qwen3.6-35B-A3B. On GLM-4.7-Flash, corpus sum yields lower KL than conditional mean for output magnitudes at both budgets (0.22309 versus 0.22852, and 0.89425 versus 1.01692) and for both residual quantities at 50% removal (0.91110 versus 0.95748, and 0.88086 versus 0.96456), reversing the middle two terms in four of the sixteen cells. On Qwen3.6-35B-A3B, mean yields lower KL than sum in all eight available comparisons, with sum yielding 1.48–2.55$\times$ the KL of mean. RMS beats sum in all sixteen matched comparisons, but the mean–sum ordering depends on the quantity and backbone.

For the four non-refill quantities in the core grid, RMS lowers reverse KL relative to mean in 14 of the 16 quantity–model–budget cells. Both exceptions are the routing factor without the consensus reference, $\frac{w_i}{1-w_i}\|f_i\|_2$, on Qwen3.6-35B-A3B, where RMS raises KL from 0.07882 to 0.09535 at 25% removal and from 0.48975 to 0.64640 at 50%. The benefit of consensus subtraction likewise depends on aggregation. With the routing factor held at w_i, subtracting the consensus lowers KL in all four settings under the mean but in three under RMS, raising it from 0.86866 to 0.92318 on GLM-4.7-Flash at 50% removal. With the leave-one-out factor, it lowers KL in all four settings under RMS but in three under the mean, raising it from 0.22922 to 0.23017 on GLM-4.7-Flash at 25% removal. Neither RMS nor consensus subtraction is uniformly beneficial, and their effects depend on the other scoring choices.

An aggregation-matched baseline also matters. In this collection, conditional RMS of the REAP token quantity $w_i\|f_i\|_2$ gives lower KL than RCS-LOO in both Qwen3.6-35B-A3B settings and slightly lower KL on GLM-4.7-Flash at 50%. Thus the mean-to-RMS benefit for fixed-support deletion damage does not imply that its residual reference is uniformly better under RMS. RCS-Refill under RMS has the lowest KL of all fourteen rows on GLM-4.7-Flash at 50% (0.85460) and beats the RMS baseline on GLM-4.7-Flash at both budgets. It trails that baseline on Qwen3.6-35B-A3B at both budgets (0.06874 versus 0.06869 and 0.44409 versus 0.43948), where RCS remains lowest.

With both quantities aggregated by the conditional mean, refill yields lower KL than REAP in all four settings (0.07386 versus 0.07469 and 0.44143 versus 0.45691 on Qwen3.6-35B-A3B, 0.22481 versus 0.22852 and 0.96456 versus 1.01692 on GLM-4.7-Flash). The mean RCS-LOO quantity still gives lower KL than mean refill on Qwen3.6-35B-A3B at 25% (0.07151). These contrasts use the mean-refill row from the separate main-sweep shards, not the identical batches used elsewhere in this table. The separate main-sweep comparison in Table 6 likewise favors RCS-LOO over RCS on GLM-4.7-Flash, with the reverse ordering on Qwen3.6-35B-A3B. The local identity fixes what a score measures, not its empirical ordering after joint pruning. These checkpoint-fidelity contrasts do not isolate the source of downstream benchmark gains.

C.6 Matched component contrasts at two removal budgets

The main component study uses four evaluation shards at 25% and 50% removal, with tokens pooled within each axis and the eight axes weighted equally. It remains separate from the collection in Appendix C.5. Each contrast fixes the other scoring choices, with both mean and RMS conditioned on routed tokens.

With routing weights and mean aggregation fixed, consensus subtraction lowers reverse KL by 1.5%/3.5% on Qwen3.6-35B-A3B and 0.2%/7.1% on GLM-4.7-Flash at 25%/50% removal. With residuals and LOO fixed, RMS lowers KL in all four settings, by 11.6%/9.0% on GLM-4.7-Flash and 0.05%/0.44% on Qwen3.6-35B-A3B. The small Qwen3.6-35B-A3B differences do not establish statistical separation. Moreover, RMS raises Qwen3.6-35B-A3B's excess perplexity by 16.5%/2.7% in this same contrast, whereas it lowers GLM-4.7-Flash's by 12.5%/13.0%. Relative to RMS without LOO, RCS-LOO lowers GLM-4.7-Flash's reverse KL but raises Qwen3.6-35B-A3B's at both budgets (Table 6). For refill damage, RMS changes reverse KL relative to the conditional mean by $-7.0\%/+0.6\%$ on Qwen3.6-35B-A3B and $-9.0\%/-11.4\%$ on GLM-4.7-Flash, and excess perplexity by $-8.5\%/+6.9\%$ and $-8.9\%/-13.2\%$. These contrasts compare checkpoint outcomes under alternative proxies without establishing an unqualified ordering of either the fidelity metrics or the proxies.

To examine transfer beyond calibration-covered task types, we also average the four covered and four held-out axes separately, giving axes equal weight within each group. Excess perplexity is $\exp(\Delta\mathrm{NLL}) - 1$. Across the two budgets, REAP's held-out reverse KL is 2.1–2.9$\times$ its in-domain reverse KL, and its held-out excess perplexity is 3.4–3.8$\times$ its in-domain excess perplexity. These grouped comparisons therefore probe a different fidelity regime from the all-axis average. The RCS-LOO score lowers Qwen3.6-35B-A3B's all-axis KL by 4.3%/1.2% relative to REAP, but its excess perplexity changes by $-0.7\%/+12.6\%$. GLM-4.7-Flash improves in both metrics, with KL reductions of 10.9%/14.3% and excess-perplexity reductions of 8.9%/17.1%. Relative to REAP, RAZOR lowers Qwen3.6-35B-A3B's KL by 7.9%/2.8% and changes its excess perplexity by $-9.5\%/+6.6\%$. Against RCS-LOO, refill lowers that model's excess perplexity by 8.9%/5.3%. On GLM-4.7-Flash, RAZOR lowers KL by 10.5%/16.0% and excess perplexity by 8.7%/19.3% relative to REAP.

C.7 SELECTION DIAGNOSTICS

Three probes examine different consequences of expert selection. Conditioning on keep-set disagreements relates selected events to output differences. Joint residual statistics measure interactions within deletion sets. Single-expert ablations test transfer from local scores to deployed interventions. None directly measures a reduction in redundancy among the final retained experts. The budgeted benchmark comparisons instead test preservation of task utility.

C.7.1 OUTPUT DIFFERENCES CONDITIONED ON SELECTION DISAGREEMENTS

We compare RCS-LOO and REAP at 50% removal on 16 batches spanning eight axes for each of Qwen3.6-35B-A3B and GLM-4.7-Flash. This collection is separate from the four-shard main sweep and is subject to the checkpoint-matching limits in Appendix B.5. It contains no RAZOR predictions, so this probe is not repeated for refill.

Each conditioning variable uses routed events from a specified layerwise keep-set disagreement. For routing concentration, eligible experts are retained by RCS-LOO but not by RCS. We select the event with the largest $1/(1 - w_i)$ across layers for each token and use its normalized routing entropy. For output consensus, eligible experts are retained by residual RMS without LOO but not by REAP. We select the event with the largest residual damage relative to the expert's calibration mean and use its $\kappa = \|c\|_2^2 / \sum_j w_j \|f_j\|_2^2$ and relative residual norm. For relative impact, eligible experts are retained by RCS-LOO but not by conditional-mean fixed-support scoring. We take the largest $\delta_i^{\mathrm{loo}} / \mathbb{E}_{\mathrm{cal}}[\delta_i^{\mathrm{loo}} \mid i \in S]$ across their routed events, or zero if none is observed. This ratio measures shift relative to an expert's mean damage, not event rarity or preservation of rare capabilities.

All diagnostics use model-wide quantiles on a common finite-token subset. Within each bin, we average RCS-LOO-minus-REAP reverse KL within batches and then equally across batches, with 5,000 paired-batch bootstrap resamples. The most concentrated routing, highest-consensus, and highest-relative-impact bins give Qwen3.6-35B-A3B and GLM-4.7-Flash differences of $-0.14/-0.33$, $-0.18/-0.28$, and $-0.25/-0.47$, respectively. These are associations conditional on selection disagreements. They neither isolate a component's causal effect nor imply the same ordering on unconditioned tokens.

C.7.2 INTERACTIONS WITHIN THE SELECTED DELETION SETS

The joint identity in Eq. 4 motivates a separate test of whether cross-expert residuals cancel or reinforce within score-selected deletion sets. We select sets with RCS-LOO at 25% and 50% removal and accumulate their interaction Gram matrices on the same calibration tokens, comprising 785,703 on Qwen3.6-35B-A3B and 785,632 on GLM-4.7-Flash. Define $u_i(x) = w_i(x)(f_i(x) - c(x))$ for routed experts and zero otherwise, and let $G_{ij} = \sum_x \langle u_i(x), u_j(x) \rangle$ within a layer. For deletion set $\mathcal{A}_\ell$, the interaction ratio is

$$\rho_\ell = \frac{\sum_{i,j \in \mathcal{A}_\ell} G_{ij}}{\sum_{i \in \mathcal{A}_\ell} G_{ii}}. \tag{23}$$

The common layer output scale cancels in the ratio. Values below one indicate net cancellation, while values above one indicate reinforcement. We compute one ratio per layer and report equal-layer quantiles over 40 Qwen3.6-35B-A3B and 46 GLM-4.7-Flash layers.

At 25% and 50% removal, median ratios are 1.080/0.921 on Qwen3.6-35B-A3B and 0.992/0.811 on GLM-4.7-Flash. At 50%, 29 of 40 Qwen3.6-35B-A3B layers and all 46 GLM-4.7-Flash layers show net cancellation. Sets selected by RAZOR from the same accumulators give median ratios of 1.077/0.925 on Qwen3.6-35B-A3B and 0.997/0.831 on GLM-4.7-Flash, with cancellation in 29 of 40 and 42 of 46 layers at 50%. Cancellation alone does not indicate better selection. The identity $\sum_i u_i(x) = 0$ constrains the cross terms, and the Gram matrices do not retain the removed routing mass $W_{\mathcal{A}}(x)$. They therefore lack the $(1 - W_{\mathcal{A}}(x))^{-2}$ factor needed to recover fixed-support deletion losses where $W_{\mathcal{A}}(x) < 1$. If no routing mass survives, that local counterfactual is undefined. The Gram ratios alone therefore neither evaluate complete deletion damage nor establish the quality of deployed selections.

C.7.3 SINGLE-EXPERT REMOVAL WITH DEPLOYED ROUTING

This probe tests whether local saliency transfers to end-to-end single-expert damage under deployed routing. With a fixed seed, we uniformly sample 32 experts and intervene on each in six layers spanning shallow, middle, and deep positions, giving 192 ablations per model. Each intervention masks one expert's router score to $-\infty$ and reselects and renormalizes the top-k from the remaining pool, potentially admitting previously inactive experts.

We compare the intervened model q with the unpruned model p on 16 shared batches, two per evaluation axis, using assistant-token reverse KL $D_{\mathrm{KL}}(q\|p)$. The resulting damage estimate pools 78,091 scored tokens on Qwen3.6-35B-A3B and 79,975 on GLM-4.7-Flash, weighting tokens rather than axes equally. For each scoring rule, we compute Spearman correlation over the 32 interventions within each layer and average the six correlations equally. The saliency statistics use the 785,703-token Qwen3.6-35B-A3B and 785,632-token GLM-4.7-Flash calibration collections.

Mean correlations for RCS-LOO, REAP, and EAN are 0.31, 0.34, and 0.31 on Qwen3.6-35B-A3B, and 0.19, 0.18, and 0.10 on GLM-4.7-Flash, respectively. Every Qwen3.6-35B-A3B correlation is positive, but several GLM-4.7-Flash values are near zero and EAN is negative in two layers. The refill score RAZOR gives 0.31 on Qwen3.6-35B-A3B and 0.18 on GLM-4.7-Flash. The point estimates show no consistent advantage across models. This probe tests transfer from conditional local output shifts to sample-wide prediction KL after router reselection. It neither tests the fixed-support identity nor evaluates jointly selected sets at a fixed budget. Better ordering of all single deletions is not necessary for better budgeted selection, but this probe also does not explain the benchmark gains.

D CALIBRATION STABILITY AND SENSITIVITY

These supplementary diagnostics concern total calibration budget and corpus configuration. We distinguish retained-set stability from output-quality comparisons. The first study varies total calibration tokens and measures overlap between selected expert sets. The second records reference-token PPL for 24 checkpoints across corpus and complementary-subset configurations. The budget study covers REAP and all three family members, because selection agreement needs only the archived per-expert accumulators. The corpus study uses the fixed-support conditional-RMS score RCS-LOO alone, because it requires one pruned checkpoint per configuration. Only the documented input-matched configurations support paired output comparisons.

D.1 CALIBRATION BUDGET AND SELECTION STABILITY

The Qwen3.6-35B-A3B collection contains 32 shards and 2,151,112 tokens. At each budget, we draw pairs of disjoint subsets with equal shard counts, merge statistics within each subset, and select experts. Agreement is the sum of layerwise retained-set intersection sizes divided by the total number of retained experts across scored layers. We report averages over four subset-pair draws from one fixed seed. Because every accumulator is a plain token sum, any shard subset is a valid pack for a smaller budget, so this study requires no model forward pass and is run for all four criteria on the same subsets. Subset agreement compares the two disjoint selections. Full-set agreement compares each selection with that from all 32 shards and averages the two values. Only the former measures agreement from disjoint calibration data, because the full-set reference includes both subsets.

Selection agreement generally increases with calibration budget for all four criteria (Figure 7). For RCS-LOO, disjoint-subset agreement rises from 85.72% to 93.39% at 50% removal as the budget grows from approximately 65,500 to 1.08 million tokens. At 25% removal, it rises from 92.82% to 96.58%. In the final budget step, disjoint-subset agreement increases by only 0.01 percentage points for RCS-LOO at 50% removal. At 25%, the corresponding changes across criteria range from -0.03 to $+0.07$ percentage points. These small changes across four draws do not establish whether agreement has plateaued.

Across eight calibration budgets, two removal ratios, and two agreement measures, REAP has the highest selection agreement and RCS the second highest in all 32 settings. RAZOR exceeds RCS-LOO in 31 settings. At the smallest budget and 50% removal, disjoint-subset agreement is 89.50% for REAP and 85.72% for RCS-LOO. At approximately 1.08 million tokens, their disjoint-subset gap narrows to 2.64 percentage points, alongside a narrowing full-set agreement gap. Even so, the residual criteria produce less stable retained sets than REAP in this collection. These comparisons do not identify the source of that sensitivity or establish how it affects fidelity and downstream performance.

These selection measurements do not identify a universally sufficient calibration budget. We report descriptive means over four subset-pair draws, without uncertainty intervals or a statistical test of the ordering.

D.2 CORPUS CONFIGURATIONS AND EVALUATION COMPARABILITY

The composition study compares full RAZORCAL, a tool-filtered subset, self-generated text, WikiText, and two complementary halves of the filtered subset at 25% and 50% removal. Filtering retains 1,790 of 2,048 examples by excluding 249 tool-calling and nine coding conversations with a tool-role message before the final assistant turn. The complementary halves combine alternating shards, $\{0, 2\}$ and $\{1, 3\}$. The filtered corpus retains RAZORCAL's mixed-source provenance and is not a human-authored control. Self-generated variants use the corresponding unpruned backbone to continue the prefix before the first assistant turn. Since later dialogue turns are not preserved, these variants change conversation structure as well as response content. For article-style language-modeling text, we use the raw WikiText-2 training split of curated Wikipedia articles (Merity et al., 2017).

Perplexity across configurations. We evaluate 24 RCS-LOO checkpoints, one per configuration, model and removal ratio. Each is scored on 128 sequences of 2,048 tokens, yielding 262,016 next-token targets. PPL exponentiates the mean reference-token NLL over all next-token positions, not only assistant turns. The unpruned references are 5.781 on Qwen3.6-35B-A3B and 3.563 on GLM-4.7-Flash. Comparisons across configurations are subject to the input-matching limits below.

Which inputs are matched? The Qwen3.6-35B-A3B full-RAZORCAL and self-generated runs share the original model's tokenization and evaluation rows. Their earlier conversation pool differs from the eight-axis fidelity pool. It contains the first 128 conversations in source order that fill the token window, each truncated to its first 2,048 tokens. On these matched inputs, self-generated calibration yields lower PPL than full RAZORCAL at both removal ratios (6.468 versus 7.544 at 25%, and 6.991 versus 9.028 at 50%). This contrast still changes calibration token budgets and conversation structure as well as response content.

Across the larger collection, matching reference and per-sequence NLL values for shared checkpoints do not establish that every configuration used identical input tokens. Self-generated PPL is lower

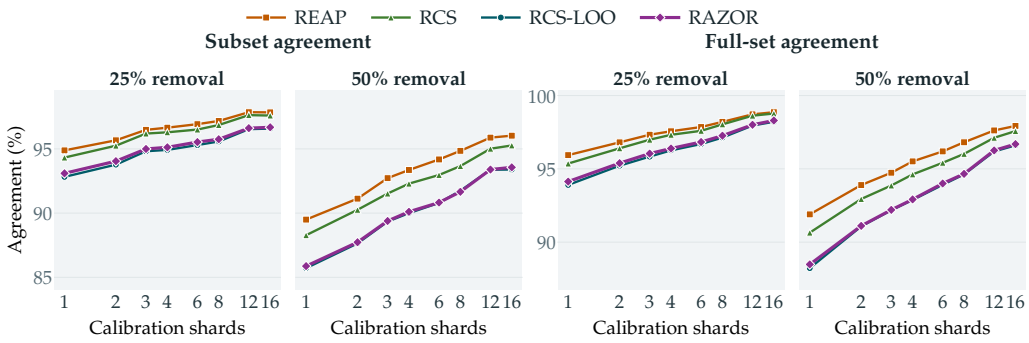


Figure 7: **Calibration budget and selection agreement on Qwen3.6-35B-A3B.** Subset agreement (left pair) compares two disjoint calibration subsets, while full-set agreement (right pair) compares each against the full 32-shard selection. Tokens per shard count are given in the text.

than tool-filtered PPL on Qwen3.6-35B-A3B and higher on GLM-4.7-Flash at both budgets, but this ordering cannot isolate a calibration-source effect. Similarly, Qwen3.6-35B-A3B's 50%-removal WikiText PPL (4.831) falls below the unpruned reference without establishing an improvement on matched inputs. The two complementary halves yield different outcomes, differing by 0.295 on Qwen3.6-35B-A3B at 25% and by 0.170 on GLM-4.7-Flash at 50%, but one split does not estimate population variance. These results compare configurations under limited input matching, not a controlled ranking of calibration sources or a robustness advantage over REAP.

E Output Fidelity and Response Diversity

The output analyses use distinct aggregation and sampling units. Per-axis fidelity resamples paired evaluation shards, domain-wise and routing-stratified analyses resample paired batches, and response diversity resamples questions. All reported 95% intervals are pointwise and unadjusted for multiple comparisons. They condition on the evaluated checkpoints and do not estimate variation across calibration draws or independent training runs. The protocols below specify how each estimator is recomputed within these sampling units.

E.1 Per-axis fidelity results

The main sweep pools four disjoint stratified shards, weighting tokens within axes and the eight axes equally. Table 6 reports its absolute and relative reverse KL at 25% and 50% removal. The RCS-LOO checkpoints have lower reverse KL than REAP on both models at both budgets, with larger relative reductions on GLM-4.7-Flash.

At 25% and 50% removal, the axis-wise comparison separates four calibration-covered axes (*math, code, instruction following, tool use*) from four held-out axes (*chat, creative, safety, SQL*). RCS-LOO has lower reverse-KL point estimates in 11 of 16 axis–ratio cells on Qwen3.6-35B-A3B and 13 of 16 on GLM-4.7-Flash. To distinguish favorable and adverse directions, we enumerate all $4^4 = 256$ ordered resamples of the four paired shards. In each resample we recompute the token-weighted KL of both methods within the axis and take $100(L_{\text{method}}/L_{\text{REAP}} - 1)$. The 2.5th and 97.5th percentiles give pointwise intervals, of which 23 lie below zero, four lie above zero, and five include zero. Per setting, the below/above/includes-zero counts are 5/1/2 for Qwen3.6-35B-A3B at both budgets, 6/1/1 for GLM-4.7-Flash at 25%, and 7/1/0 for GLM-4.7-Flash at 50%. The four adverse cells are creative writing on Qwen3.6-35B-A3B at both budgets, SQL on GLM-4.7-Flash at 25%, and mathematics on GLM-4.7-Flash at 50%. Thus the aggregate improvement coexists with identifiable axis-level regressions. The same procedure applied to RAZOR gives lower point estimates in 11 of 16 cells on Qwen3.6-35B-A3B and 14 of 16 on GLM-4.7-Flash, with 24 intervals below zero, three above zero, and five including zero (6/0/2 and 5/1/2 on Qwen3.6-35B-A3B at 25%/50%, 6/1/1 and 7/1/0 on GLM-4.7-Flash). Refill removes the adverse Qwen3.6-35B-A3B creative-writing interval at 25%, but the other three adverse cells remain. All comparisons share the same evaluation shards and ordered resamples within each axis. Their intervals are paired contrasts, not estimates of variation across

checkpoints or calibration draws. This view therefore decomposes the same predictions rather than replicating the experiment.

E.2 Domain-wise output fidelity

Figure 2 resolves fidelity by domain for four backbones at both removal ratios, using archived predictions collected by applying criterion-specific expert masks to each original model. These mask-based diagnostics are separate from the downstream benchmark runs. The figure compares REAP, RCS, RCS-LOO and RCS-Refill using reverse KL $D_{\mathrm{KL}}(q\|p)$ in nats over four calibration-covered and four held-out domains. Each domain estimate averages token means equally over six batches. ID/OOD summaries then give equal weight to domain means, and intervals use 10,000 paired-batch bootstrap draws, shared across criteria so every comparison stays paired on the same resampled batches.

For RCS-LOO versus REAP, lower-KL domain counts are 6/4/7/6 at 25% removal and 7/5/7/7 at 50%, in GLM-4.7-Flash, Qwen3.6-35B-A3B, DeepSeek-V4-Flash-0731, and Hy3 order, each out of eight domains. Qwen3.6-35B-A3B's held-out equal-domain mean is higher than REAP's by 7.3% at 25% and 2.3% at 50%, with relative-change intervals crossing zero at both budgets. By contrast, the other three models have lower held-out means at 50%. For RAZOR, the corresponding counts are 5/6/6/6 at 25% and 7/5/8/7 at 50%. Its Qwen3.6-35B-A3B held-out mean changes by -3.9% at 25% and $+0.2\%$ at 50% relative to REAP, again with intervals crossing zero, and the other three models again have lower held-out means at 50%.

Both removal ratios share reference predictions and token positions, with at most 1,024 valid positions per batch and no further subsampling. For RCS-LOO the 50% row uses the same predictions as Figure 8, but aggregates them by domain rather than routing decile. Neither view evaluates full generated trajectories.

Each spoke uses an independent KL scale spanning all four point estimates, padded by the median bootstrap half-width across criteria. Curves are comparable along a spoke, but distances are not comparable across spokes, panels, or removal ratios. Eight of 256 whiskers extend beyond the frame, with their full intervals recorded in the released plotted-value table alongside relative changes against REAP. Intermediate labels give mid-axis KL values.

E.3 Routing-stratified output fidelity

Figure 8 groups assistant tokens by routing concentration within each learned-router MoE layer. We rank tokens by negative normalized routing entropy $-H_r$ and form ten quantile groups with shared ID/OOD boundaries. A common finite-value mask pairs methods and metrics, including the accompanying entropy and PPL measurements. We exclude layer–batch–decile cells with fewer than 20 tokens, then average within-cell token means equally over valid layers and batches. This grouped estimator differs from the main sweep's equal-axis estimator. The horizontal axis runs from diffuse to concentrated routing.

Four calibration-covered and four held-out domains are evaluated separately. Within each ID/OOD group, 2,000 paired-batch bootstrap resamples preserve all ten deciles. The 25% and 50% curves share original-model routing profiles, token masks, quantile boundaries, and valid layer–batch–decile cells, but compute intervals separately.

The collector does not apply learned-router saliency scoring to DeepSeek-V4-Flash-0731's three fixed-hash-router layers, which are excluded from the curves. Across criteria, these layers use identical frozen keep sets based on occurrence counts in the token-to-expert table, not calibration-token frequencies. Remapping preserves k distinct routes per token.

At 50% removal, held-out reverse KL is 2.1–$3.0\times$ its corresponding in-distribution value. From the least to the most concentrated routing decile, RCS-LOO's relative advantage widens by $15.6/17.2$ percentage points on Qwen3.6-35B-A3B ID/OOD and by 4.3 points on GLM-4.7-Flash ID. Nominal paired-bootstrap intervals exclude zero for these three endpoint contrasts but cross zero for the other five, giving mixed evidence for a concentration-dependent advantage.

Aggregate point estimates favor RCS-LOO by 14.4% on GLM-4.7-Flash OOD, $15.4/10.3\%$ on Hy3 ID/OOD, and $7.5/8.9\%$ on DeepSeek-V4-Flash-0731 ID/OOD. Qwen3.6-35B-A3B's OOD

KL reduction is near zero (-1.5%, CI $[-5.3, +2.7]$). The same estimator applied to RAZOR, whose 50% predictions share the reference predictions and token mask, favors it by 15.6% on GLM-4.7-Flash OOD, $17.2/12.9\%$ on Hy3 ID/OOD, and $10.0/13.1\%$ on DeepSeek-V4-Flash-0731 ID/OOD, with Qwen3.6-35B-A3B OOD again nearly tied ($+0.5\%$, CI $[-3.7, +4.9]$). Relative to REAP, RCS-LOO's mean absolute entropy drift is lower by 15%, 22%, 11%, and 21% on GLM-4.7-Flash, Qwen3.6-35B-A3B, DeepSeek-V4-Flash-0731, and Hy3, respectively.

Adding the leave-one-out factor does not uniformly improve routing-stratified fidelity under conditional RMS. Averaged over deciles, RCS attains the largest KL reduction against REAP in three of the eight model–group panels. These are both Qwen3.6-35B-A3B panels ($+15.9\%$ ID and $+11.9\%$ OOD, against $+9.4\%$ and -0.4% for RAZOR) and DeepSeek-V4-Flash-0731 ID ($+11.4\%$ versus $+9.7\%$). RAZOR leads the remaining five panels by 1.2–4.3 percentage points over the better of RCS and RCS-LOO. The largest RCS–RCS-LOO gap occurs on Qwen3.6-35B-A3B OOD, where RCS-LOO and RAZOR both fall below REAP on average, while RCS improves on it. The gap narrows at the tenth decile, where the reductions are $+16.5\%$, $+11.8\%$, and $+12.4\%$ for RCS, RCS-LOO, and RAZOR, respectively. These equal-decile means at 50% removal do not establish a downstream ranking or replace the matched equal-axis estimates in Table 6. The RCS–RCS-LOO contrast shows that the gate factor's effect on fidelity depends on the backbone even with aggregation held fixed.

Entropy drift complements reverse KL by measuring changes in predictive dispersion. Predictive entropy itself is neither accuracy, calibration, nor response diversity, and neither higher nor lower entropy alone is preferable.

Reverse KL and perplexity deviation give different rankings. RAZOR lowers reverse KL relative to REAP in 73 of 80 deciles, with all seven exceptions in Qwen3.6-35B-A3B's held-out group. On perplexity deviation, however, it is farther from the original than REAP in $7/10$ ID deciles on Qwen3.6-35B-A3B and $9/10$ on DeepSeek-V4-Flash-0731, while remaining closer in all twenty GLM-4.7-Flash ID and OOD deciles. Across all groups it preserves PPL more closely in 47 of 80 deciles. These are the same paired point comparisons summarized in Appendix E.4, not separate experiments.

The incremental effect of refill depends on the budget and domain group. At 50% removal, RAZOR exceeds RCS-LOO by 1.0–4.3 percentage points in REAP-relative KL reduction, averaging deciles equally within each of the eight model–domain groups. Their marginal bands overlap throughout on Qwen3.6-35B-A3B. At 25% removal, the ordering varies by model. In the calibration-covered groups, RAZOR leads by 8.2 points on Qwen3.6-35B-A3B but trails RCS-LOO by 3.7 on GLM-4.7-Flash and by 1.5 on DeepSeek-V4-Flash-0731. The refill margin also does not increase consistently with concentration, as its linear slope at 50% ranges from -0.30 to $+0.06$ percentage points per decile across the eight groups. These profiles distinguish the two counterfactuals without establishing a concentration-dependent benefit from refill.

E.4 REFERENCE-TOKEN PERPLEXITY ON THE SAME ROUTING GROUPS

To distinguish distributional fidelity from reference-token likelihood, we compute PPL at 50% removal for all four backbones on the same predictions, finite-token mask, and layer–batch–decile cells as Figure 8, whose lower two rows plot the result. Using the reference-token losses in Appendix B.4, let $\Delta\ell(t) = \ell_q(t) - \ell_p(t)$. Grouped PPL is

$$\mathrm{PPL}_p = \exp\big(\mathcal{A}[\ell_p]\big), \qquad \mathrm{PPL}_q = \exp\big(\mathcal{A}[\ell_p + \Delta\ell]\big), \tag{24}$$

where $\mathcal{A}$ is the routing-stratified average over valid token cells, layers, and batches. Exponentiation follows NLL aggregation rather than preceding it. The result is neither mean per-token perplexity nor exponentiated predictive entropy, and need not equal whole-corpus token-weighted PPL. We measure preservation by the absolute deviation $|\mathrm{PPL}_q - \mathrm{PPL}_p|$, so "closer" means a smaller deviation than REAP on the same group, not simply a lower PPL.

The bootstrap pairs the original, REAP, and RCS-LOO predictions on the same retained positions. Within each ID/OOD group, shared whole-batch indices preserve all ten deciles. Recomputing and exponentiating each resampled NLL average gives marginal intervals and paired PPL differences.

The two fidelity views disagree in many cells. Across 80 point estimates from four models, two ID/OOD groups, and ten deciles, RCS-LOO has lower reverse KL than REAP in 72 but PPL closer to the original in only 47. Both measures favor RCS-LOO throughout GLM-4.7-Flash's groups. On

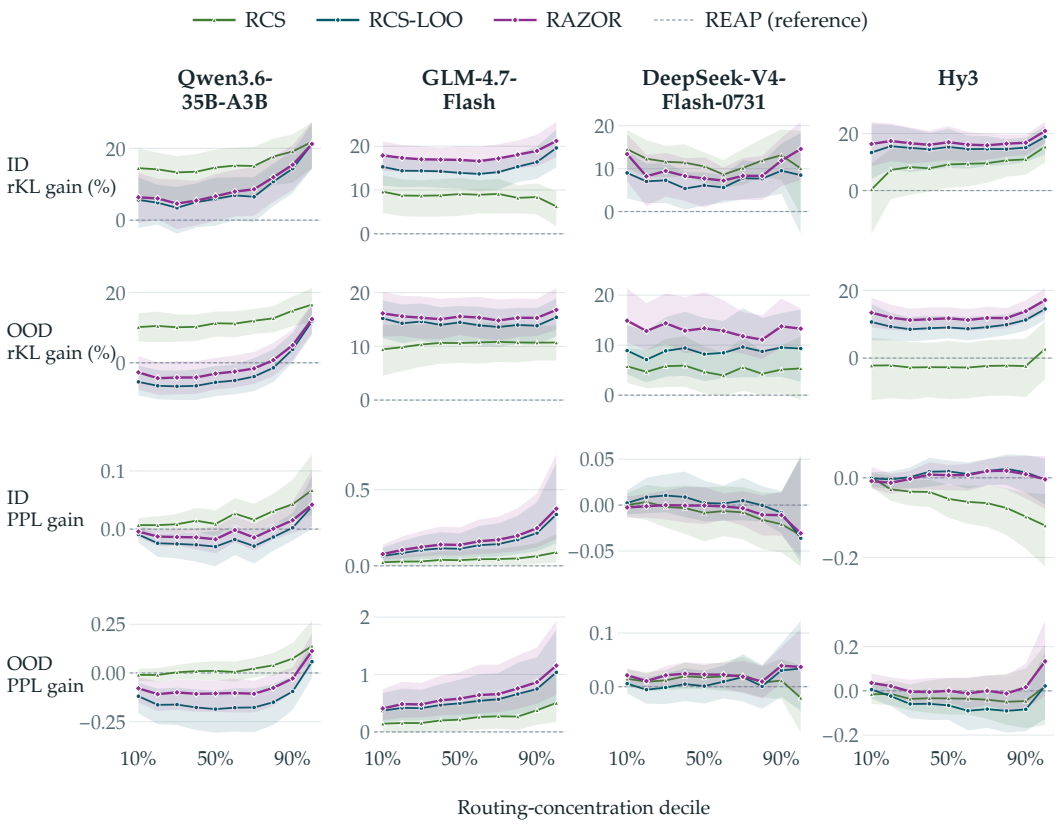


Figure 8: **Routing-stratified fidelity at 50% removal.** All three scored criteria use conditional RMS, so the curves differ only in the token quantity. The RCS arm was rescored for this figure on the same batches as the archive, with matching row order and domain labels and exactly matching stored original-model NLL and predictive entropy. All panels compare against REAP at the same decile, with a dashed zero line and higher values indicating better fidelity. The top two rows show reverse-KL reduction in percent. The bottom two show REAP's $|\mathrm{PPL}_q - \mathrm{PPL}_p|$ minus the criterion's, in PPL points (Equation 24), so positive values mean closer to the original. Differences avoid unstable ratios where REAP's deviation approaches zero (0.0016 on Hy3 ID). Bands are pointwise 95% paired batch-bootstrap intervals, and panels use independent vertical scales.

Qwen3.6-35B-A3B, however, PPL is closer in only two ID and one OOD decile despite lower reverse KL in every ID decile. DeepSeek-V4-Flash-0731's ID NLL changes are negative, so lower PPL need not mean smaller deviation from the original. On the same cells, RAZOR has lower reverse KL than REAP in 73 of 80 and closer PPL in 47. Better distributional fidelity therefore does not imply uniformly better preservation of reference-token likelihood.

E.5 RESPONSE DIVERSITY

Criterion-family comparison. The left panel of Figure 3 compares REAP, RCS, RCS-LOO, and RAZOR on Qwen3.6-35B-A3B using Distinct-n for $n = 2, 3, 4$. For each benchmark, all four criteria at both budgets and the unpruned reference use 16 fixed response positions per question, drawn from the first 256 positions for AIME and 32 for each coding task. These common pools are limited by RCS's generation counts. The 25% RCS-LOO AIME run also contains regenerated responses under the same recorded decoding protocol. This comparison therefore differs from the archived analysis below in both sampling pools and some response draws, so their percentages are not directly comparable.

Within each benchmark, values are $100(\bar{M}/\bar{M}_{\mathrm{Original}} - 1)$ for question-averaged metrics. AIME 2026, HumanEval+, and LiveCodeBench receive equal weight, so AIME contributes one third of the

weight on 30 of 369 questions. For pointwise 95% paired question-bootstrap intervals, we average the resampled relative changes across benchmarks before taking percentiles. These intervals condition on the 16 fixed responses rather than measure variation over new generations. Of twelve intervals per budget (four criteria and three orders), all twelve exclude zero at 50% and nine do so at 25%. They describe changes from Original, not the RCS-LOO–REAP difference. In per-benchmark point estimates, RCS-LOO exceeds REAP in all nine benchmark–order cells at 50% and six at 25%. Across the six benchmark-averaged budget–order cells, its median advantage over RCS is 4.02 percentage points in relative change from Original. Refill lowers Distinct-n relative to RCS-LOO in five of these six cells.

Archived collection and sampling. The archived repeated-generation collection is separate from the nine-task benchmark. It contains 30 AIME 2026 (Mathematical Association of America, n.d.), 219 Minerva (Lewkowycz et al., 2022), 164 HumanEval+ (Liu et al., 2023), and 175 LiveCodeBench (Jain et al., 2025) questions. Minerva uses 219 of OCWCourses' 272 undergraduate STEM problems from MIT OpenCourseWare, with the source reports recording the 53 omitted problems as ungradable. The 175 LiveCodeBench task IDs are preserved, but no immutable release identifier is recorded for this diagnostic collection, so it should not be equated with the main benchmark.

Qwen3.6-35B-A3B has 27 generation runs covering nine variants, comprising Original plus EAN, Frequency, REAP, and RCS-LOO at 25% and 50% removal. GLM-4.7-Flash adds Original, REAP, and RCS-LOO at both ratios, bringing the total to 14 variants and 42 runs. Unlike the three-benchmark Qwen3.6-35B-A3B comparison in Figure 3, the archived analysis includes Minerva and GLM-4.7-Flash.

Mathematics has 2,048 generations per question, HumanEval+ has 512, and LiveCodeBench has 256. Generation uses temperature 0.7, top-$p = 0.95$, top-$k = 20$, repetition penalty 1.0, an 8,192-token output cap, thinking disabled, and seed 42. These settings differ from benchmark decoding in Appendix B.3.

We sample 16 response positions per question without replacement using a question-specific seed derived from 20260910, reusing positions across methods. The archived analysis uses 131,712 responses, including 84,672 from Qwen3.6-35B-A3B across four tasks. Of these, 53,136 cover AIME 2026, HumanEval+, and LiveCodeBench. These counts exclude unsampled generations. Some upstream reports and generations have since changed. The archived results retain per-question measurements and sampled responses, but their original raw inputs cannot all be revalidated at the recorded locations.

Metric definitions. Metrics are computed per question and then averaged over questions. We adapt the distinct-n approach of Li et al. (2016) to model tokens, using $n \in \{2, 4\}$ in the archived analysis and $n \in \{2, 3, 4\}$ in Figure 3. Distinct-n divides the number of unique token n-grams pooled across 16 responses by their total occurrences, without crossing response boundaries. Jaccard distance averages $1 - |G_i \cap G_j|/|G_i \cup G_j|$ over all 120 response pairs, where G_i and G_j are response-level token 4-gram sets. Both lexical measures use each model family's own tokenizer.

Embedding distance averages pairwise cosine distances using the fixed all-MiniLM-L6-v2 encoder.[1] Our full-response extension splits text into non-overlapping chunks of at most 254 WordPieces, mean-pools token representations, and normalizes each chunk vector. We then normalize their length-weighted mean. This retains content beyond the first chunk but not chunk order. The resulting distances do not measure execution behavior or mathematical-strategy equivalence.

Normalization and uncertainty. For each method, family, and task, the archived comparison reports $100(\bar{M}/\bar{M}_{\text{Original}} - 1)$ using question means $\bar{M}$, where Original denotes the unpruned MoE. Higher values indicate more diversity. We recompute this ratio over 2,000 paired question-bootstrap draws and report pointwise percentile 95% intervals, conditional on the 16 fixed responses and evaluated checkpoints. They exclude new-response and checkpoint-level variation.

Length control and interpretation. The four archived metrics come from an exploratory 15-metric analysis and are correlated, so they are not independent evidence. We assess length sensitivity

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

by repeating the analysis on exactly the first 128 tokens of each response. Within a family, we retain only questions eligible for every available method at both pruning ratios. In AIME 2026, Minerva, HumanEval+, and LiveCodeBench order, the resulting question counts are 30/219/134/143 for Qwen3.6-35B-A3B and 29/176/82/142 for GLM-4.7-Flash.

At 50% removal, the four-task mean Distinct-4 gap between RCS-LOO and REAP changes from $+4.30\%$ on full responses to -0.37% under prefix control on Qwen3.6-35B-A3B. The corresponding gaps on GLM-4.7-Flash are $+10.35\%$ and $+16.50\%$. These means weight task-relative changes against REAP equally, unlike the Original-normalized diversity panel of Figure 3. The control changes both length and question eligibility, so Qwen3.6-35B-A3B's reversal is not a pure length effect. Neither lexical nor embedding diversity alone establishes correctness, token efficiency, or more distinct correct solutions.

E.6 Response delimiters, formatting, and termination

Protocol and operational definitions. This analysis follows the repeated-generation protocol in Appendix E.5, using temperature 0.7, top-$p = 0.95$, top-$k = 20$, an 8,192-token cap, thinking disabled, and seed 42. The tables and the three response-form panels of Figure 3 use 256 shared sample positions per question, giving 44,800 completions per variant on LiveCodeBench. These measurements characterize response form, not the benchmark accuracy of Table 1.

A *stray* `</think>` is a literal occurrence of that string in a completion. In this tokenizer it is a single vocabulary entry that is not flagged as a special token, so ordinary detokenization leaves it in the text. The chat template already closes the thinking block inside the prompt, so any further occurrence is emitted by the model and is not evidence about latent reasoning. We avoid the word *leakage* because it denotes benchmark contamination in evaluation, which these counts do not measure. *No fence* denotes the absence of a line-delimited code fence, while *prefix* denotes a nonempty prefix before such a fence. Both are shares of all completions, and neither is a semantic code or narration classifier. Unfenced responses can contain code, which the evaluator may extract through its bare-code fallback. *Length-limited* uses the decoder's finish reason.

Response lengths in this table count Unicode characters, whereas Figure 3 counts tokens under the backbone's own tokenizer, matching the units of its Distinct-n panel. Characters per token fall from 3.44 on the unpruned model to 3.04 for REAP at 50% removal, so the two units are not interchangeable. The character counts here understate that checkpoint's growth in generated tokens. Intervals resample questions to retain dependence among completions sharing a prompt.

Sample-pool sensitivity. Rare-event estimates depend on the number of sampled completions. REAP's stray-delimiter rate is 8.93 per 10,000 with 32 completions per question versus 13.39 with 256, although the criteria retain the same point-estimate ordering at 50% removal under either pool. Both criteria in the main-text panel use the full 44,800-completion pool, whose rates and intervals are reported below.

Budget-dependent changes. We measure ten RCS-LOO checkpoints from 0% to 75% removal on each backbone. Table 11 shows the quarter budgets and reports each rank correlation over all ten checkpoints. Only RCS-LOO covers every measured budget in this collection, while Table 10 compares other criteria at shared budgets. On Qwen3.6-35B-A3B, the unfenced and length-limited shares each have Spearman $\rho = +0.82$ with removal budget ($p = 0.007$), and the rate of stray `</think>` has $\rho = +0.75$ ($p = 0.020$). GLM-4.7-Flash gives $\rho = +0.93$ ($p < 0.001$), $+0.58$ ($p = 0.099$), and $+0.89$ ($p = 0.001$), respectively. These descriptive associations are not stepwise monotonic. Among the undisplayed budgets, Qwen3.6-35B-A3B's length-limited share falls at 40% and GLM-4.7-Flash's stray-delimiter rate peaks at 60%. The displayed values also show Qwen3.6-35B-A3B's stray rate higher at 50% than at 75%.

Median response length has $\rho = +0.43$ ($p = 0.24$) on Qwen3.6-35B-A3B and $+0.17$ ($p = 0.67$) on GLM-4.7-Flash. The prefix share has $\rho = -0.28$ ($p = 0.46$) and -0.88 ($p = 0.002$), respectively. The rank tests do not establish an association for every measure, but nonsignificance does not establish a flat response. All tests are exploratory, use ten checkpoints on shared questions, and report two-sided p-values uncorrected for multiple comparisons. They do not estimate independent-run variability. At

Table 10: **Response form on LiveCodeBench by criterion and budget, Qwen3.6-35B-A3B.** Rates are percentages except stray </think> per 10,000. Lengths are in kchar. Brackets give pointwise 95% question-bootstrap intervals. GLM-4.7-Flash is measured but not printed here because that collection covers only REAP and RCS-LOO, with no RAZOR row for comparison. Its RCS-LOO ladder appears in Table 11.

Criterion	Cut	Stray </think>/10k	No fence	Prefix	Len-lim.	Med. kchar
Qwen3.6-35B-A3B, 44,800 completions per checkpoint						
Original	–	3.6 [1.6, 6.0]	7.1	61.4	16.0	8.3
EAN	25%	0.4 [0.0, 1.1]	8.7	52.3	19.5	9.7
Frequency	25%	2.0 [0.7, 3.6]	5.2	60.5	15.1	9.2
REAP	25%	1.6 [0.4, 3.1]	8.4	55.1	19.1	10.3
RCS-LOO	25%	0.2 [0.0, 0.7]	8.4	52.3	17.0	8.6
RAZOR	25%	0.4 [0.0, 1.1]	7.7	54.7	16.3	8.4
EAN	50%	4.2 [2.2, 6.9]	18.3	40.2	33.9	9.6
Frequency	50%	16.1 [10.9, 22.3]	2.9	30.4	13.0	6.7
REAP	50%	13.4 [7.6, 20.1]	11.5	66.3	24.2	12.0
RCS-LOO	50%	68.5 [22.8, 133.7]	10.2	16.6	25.1	7.3
RAZOR	50%	8.5 [4.2, 13.6]	8.2	19.6	25.8	7.8

Table 11: **The RCS-LOO budget ladder on LiveCodeBench.** Spearman correlations use all 10 checkpoints, not only the four shown. Units follow Table 10.

Quantity	0%	25%	50%	75%	ρ	p
Qwen3.6-35B-A3B						
Stray </think>/10k	3.6	0.2	68.5	42.0	+0.75	0.020
No fence (%)	7.1	8.4	10.2	25.8	+0.82	0.007
Prefix (%)	61.4	52.3	16.6	43.6	−0.28	0.460
Length-limited (%)	16.0	17.0	25.1	42.0	+0.82	0.007
Median kchar	8.3	8.6	7.3	8.4	+0.43	0.244
GLM-4.7-Flash						
Stray </think>/10k	0.4	0.0	2.9	70.5	+0.89	0.001
No fence (%)	0.2	0.5	1.7	17.0	+0.93	<0.001
Prefix (%)	66.5	68.9	58.3	14.3	−0.88	0.002
Length-limited (%)	17.5	19.6	18.6	78.7	+0.58	0.099
Median kchar	3.8	4.0	3.0	30.4	+0.17	0.668

75% removal, Qwen3.6-35B-A3B has 25.8% unfenced and 42.0% length-limited completions, while GLM-4.7-Flash has 78.7% length-limited completions and a median length of 30,444 characters.

Refill and delimiter emissions on Qwen3.6-35B-A3B. At 25% removal, every criterion's tag-rate point estimate is at or below the unpruned Qwen3.6-35B-A3B rate of 3.6 per 10,000, with RCS-LOO lowest at 0.2 [0.0, 0.7] (Table 10). At 50%, RCS-LOO reaches 68.5 [22.8, 133.7], versus REAP's 13.4 [7.6, 20.1]. The refill checkpoint reduces the observed rate to 8.5 [4.2, 13.6], below REAP's point estimate but with overlapping marginal intervals.

The RCS-LOO–REAP gap persists within baseline-length strata. In the fourth quintile of the unpruned length distribution, rates are 100.7 and 14.2 per 10,000, with RAZOR at 14.4. Matching question and sample position against Original gives 307 newly occurring tag events for RCS-LOO, 60 for REAP, and 38 for RAZOR. These checks preserve the comparison within shared prompts and baseline-length groups without ruling out effects of changed generated lengths or termination. On the same collection, GLM-4.7-Flash's corresponding RCS-LOO rate is 2.9 [1.1, 4.9], as shown at

50% removal in Table 11. REAP, which is not included in that table, has a rate of 1.6 $[0.4, 2.7]$. The magnitude of the Qwen3.6-35B-A3B gap therefore does not carry over. Across variants, 98–100% of tag-containing completions also contain code output, but this does not establish answer correctness or preserved utility.

Length shifts are small at 25% and mixed at 50%. At every 25% checkpoint, the median paired length change relative to Original, matched by question and sample position, is at most 17 characters in magnitude. At 50%, REAP lengthens responses by a median 542 characters (67.3% of completions longer), RCS-LOO by 126 and RAZOR by 114, while Frequency shortens them by 129. The share of completions with a prefix before the code fence falls to 16.6% for RCS-LOO and 19.6% for RAZOR but rises to 66.3% for REAP, against 61.4% unpruned.

Task and backbone dependence. HumanEval+, the shortest task on both backbones, is the only task nearly free of visible tags throughout, with one tag-containing completion out of 671,744. On the other tasks, the two backbones differ. Qwen3.6-35B-A3B emits tags on LiveCodeBench but almost nowhere else, with one AIME 2026 completion out of 84,480 and none of 616,704 on Minerva. GLM-4.7-Flash instead shows this behavior on mathematics. At 50% removal, 2.90% $[1.25, 5.18]$ of its AIME 2026 completions carry a tag under REAP and 7.25% $[4.11, 11.22]$ under RCS-LOO. The corresponding rates are 0.67% and 0.54% on Minerva, against at most 0.03% on LiveCodeBench.

Task-level median length alone does not order these rates. Qwen3.6-35B-A3B's AIME 2026 responses are longer than its LiveCodeBench responses (10,862 versus 8,332 characters) yet contain almost no tags. This comparison does not isolate length from other task differences. GLM-4.7-Flash has no refill checkpoint in this collection, leaving the effect of refill on its generation behavior untested.

Matched examples. For one LiveCodeBench question with sample position and seed fixed, the 50% checkpoints produce 1,894 and 1,897 characters under Frequency and RAZOR (both opening with code), 4,228 under REAP, 4,875 under EAN, and 6,923 under RCS-LOO, versus 3,423 unpruned. Only RCS-LOO emits a closing thinking tag, preceded by 6,233 characters ending with "`Let's code accordingly.`" and followed by program text. This describes the visible response sequence, not a latent reasoning process. The median pre-tag length among this checkpoint's tag-containing completions is 12,507 characters.

A separate RCS-LOO 70% completion reaches the cap after 21,966 characters without a code fence, contributing to both the unfenced and length-limited counts. The two examples were selected by median preamble length within their respective categories. The tag-containing example terminates normally. They illustrate the measured categories rather than establish comparative answer quality.

F EXTENDED RELATED WORK

MoE compression methods differ in their removal unit, importance signal, and selection or adaptation procedure. These distinctions separate the local scoring question studied here from broader changes to the expert pool and router.

F.1 WHOLE-EXPERT SALIENCY AND CONDITIONAL AGGREGATION

Routing, activation, and parameter statistics offer inexpensive rankings. SEER-MoE uses activation counts or routing-probability soft counts with layerwise or global removal (Muzio et al., 2024). AIMER instead ranks experts without calibration data using a normalized ℓ_1/ℓ_2 concentration statistic over their weights (Liu et al., 2026a). Jaiswal et al. (2025) compare several expert-dropping criteria, including activation norms. These independent scores do not reconstruct the routed mixture after deletion.

REAP, our closest magnitude-based comparator, averages $w_i\|f_i\|_2$ over tokens selecting expert i (Lasby et al., 2026). Its analysis includes promoted substitution and survivor renormalization, but its saliency uses only the removed expert's weighted output norm. A unified formulation similarly derives an exact rerouting residual before omitting it from practical one-pass scores (Liu et al., 2026b). Consensus residuals instead retain the removed expert's direction relative to the mixture, while RCS-Refill also includes the promoted expert's weighted residual.

Token-level scoring and cross-token aggregation remain separate choices. RAZOR uses conditional RMS rather than REAP's conditional mean, while the component comparisons hold the token quantity fixed when varying aggregation (Appendix C). Neither conditioning nor RMS follows from the local deletion identity, and no aggregation rule is uniformly preferred by the available results.

F.2 EXPERT RELATIONS AND CANDIDATE-SET SELECTION

Relational information is not unique to consensus-residual scoring. STUN clusters router-derived behavior before representative selection or selective reconstruction and unstructured pruning (Lee et al., 2025). HC-SMoE instead clusters average expert outputs and merges each group (Chen et al., 2025). SHAPE uses routing co-occurrence in a coalition-utility proxy (Zhang, 2026). ConMoE combines routing-conditioned contribution with nearest-expert parameter distance before prototype reassignment (Yao et al., 2026). MAESTRO derives importance from cross-layer routing transitions before layerwise selection (Goel et al., 2026), with its main results using attention-only LoRA recovery while experts and routers remain frozen.

Our scores instead evaluate output change under specified deletion counterfactuals at a fixed layer input. The reference is the routed mixture, not a cluster representative or a routing graph. This distinction concerns the local score, not a claim that independent expert ranking captures all interactions. Residuals can reinforce or cancel under joint removal (Appendix A.3), so exact single-deletion quantities do not yield an exact set-selection algorithm.

Other approaches evaluate candidate sets directly. Lu et al. (2024) select retained combinations by layer-output reconstruction, EEP searches pruning and merging configurations (Liu et al., 2024), and MoE-I^2 evaluates layerwise deletion sets before jointly choosing candidates across short layer blocks (Yang et al., 2024). We instead retain the highest-scoring experts under a fixed layerwise budget. Our local single-deletion identities neither solve combinatorial selection nor establish superiority to subset search.

F.3 MERGING, ROUTER ADAPTATION, AND BUDGET ALLOCATION

Expert merging changes the surviving functions, rather than only selecting which original experts remain. MC-SMoE groups experts around high-usage representatives using router-logit similarity, aligns neurons, and averages expert parameters before low-rank and structured compression (Li et al., 2024). REAM protects high-saliency experts as centroids and merges similar non-centroid experts into them (Jha et al., 2026). Neither operation is the router refill modeled by RCS-Refill, which promotes an existing unselected expert without averaging its parameters with the removed expert.

Router KD adapts only router parameters after compression, distilling the uncompressed model's next-token distribution while freezing the experts and other backbone parameters (Hyeon & Do, 2026). It is a recovery procedure, not an expert-ranking criterion. Budget allocation is another distinct choice. GRAPE compares residual redundancy across layers to allocate non-uniform budgets, merging similar expert pairs within the selected layer (Zhang et al., 2026). MC-SMoE also permits adaptive layerwise merging ratios (Li et al., 2024). These examples show that merging, adaptation, and budget allocation are overlapping design choices rather than mutually exclusive method families. Our experiments use uniform layerwise removal ratios, leave retained expert weights unchanged, and perform no recovery training. Combining residual-based scores with merging, router adaptation, or non-uniform allocation is outside the present comparison, so our results do not establish their compatibility or joint benefit.

F.4 FINER-GRAINED COMPRESSION AND COMPARISON SCOPE

MoE-Pruner sparsifies weights within experts using weight magnitudes and router-weighted input activation statistics (Xie et al., 2024). HEAPr treats feed-forward intermediate neurons as atomic expert units and uses output-space second-order information to estimate their importance (Li et al., 2026). Both operate at a finer granularity than removing entire routed experts. An atomic expert in this terminology is not one of the router's original selectable experts.

These approaches share an interest in preserving model behavior, but their removal units and compression procedures differ from ours. Their reported sparsity levels are therefore not matched

whole-expert budgets. Within our fixed pruning protocol, we compare the consensus-residual scores RCS, RCS-LOO, and RCS-Refill, with RAZOR denoting RCS-Refill under conditional RMS.