RECOUP THE EXPERTS: COUPLING-AWARE MOE PRUNING UNDER DEPLOYED ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Expert-pruning criteria for sparse Mixture-of-Experts (MoE) models are computed under the routing of the original model, but after pruning each token is routed to its top-k experts among the remaining ones and their gates are renormalized. We propose RECOUP, a criterion that selects the set of experts whose outputs best reconstruct the layer outputs of the removed experts on the tokens they share. The objective has a closed form in the Gram matrix of gated expert outputs, and when expert outputs are orthogonal it reduces to a per-expert ranking that matches REAP up to the moment order. We also propose RECOUP$_{\text{deploy}}$, which computes the same statistics under the routing of the pruned model. On four MoE models, the form that performs better at low budgets follows the off-diagonal mass of the Gram matrix, which can be computed before pruning. On Gemma-4-26B, the model with the highest off-diagonal mass, RECOUP scores 59.2 on GSM8K when 25% of the experts are kept, against 15.6 for REAP. On Qwen3.6-35B, the model with the lowest, RECOUP$_{\text{deploy}}$ is 19.5 points above REAP on HumanEval+ when 12.5% are kept. These values are means over three disjoint calibration samples.

1 INTRODUCTION

In a pruned Mixture-of-Experts model (Shazeer et al., 2017; Fedus et al., 2022), the router selects each token's top-k experts among the remaining experts and renormalizes their gates. Tokens are therefore routed to experts they were not routed to before, and the gates of the remaining experts increase. Existing criteria, such as REAP's gate-weighted activation norm and related scores, are computed under the original routing and for each expert separately. They measure how much each expert contributes to a computation that no longer takes place after pruning.

A pruning criterion involves two choices: what is scored, and under which routing the statistics are computed. Importance criteria score each expert separately under the original routing. We propose Reconstruction of Coupled-expert Outputs under Pruning (**RECOUP**), which scores a kept set as a whole by how well it reconstructs the layer outputs of the removed experts from the kept experts that are co-routed with them. The optimal reconstruction has a closed form in the Gram matrix of gated expert outputs, whose off-diagonal entries are the co-routed coupling between experts that per-expert scores ignore (Section 3). The second choice is the routing under which the statistics are computed. The base form uses the original routing and requires one forward pass. The variant **RECOUP$_{\text{deploy}}$** computes the statistics under top-k re-selection and gate renormalization, that is, under the routing of the pruned model, and therefore minimizes the post-pruning output error directly without compensation coefficients.

The analysis indicates when the difference between the two matters. Within the family, the only free choice is the token weighting: the choice between summing and averaging over tokens is a choice of weighting, and selection is invariant to rescaling individual experts. When expert outputs are orthogonal, the Gram matrix is diagonal, the objective is modular, and RECOUP reduces to a per-expert norm ranking, which under expert-uniform weighting is REAP's ranking up to the moment order. RECOUP therefore generalizes REAP, and the two differ only when expert outputs are correlated. This can be checked before pruning with the *coupling ratio*, the ratio of off-diagonal to diagonal Gram mass, shown for each model in Figure 1.

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.

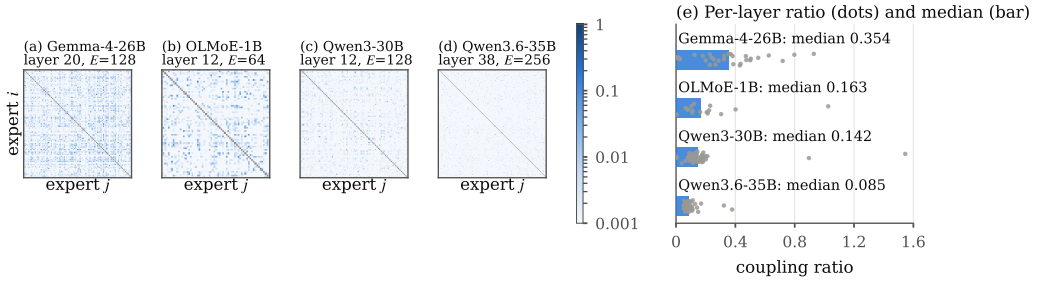


Figure 1: Coupling ratio of the four models. (a–d) Normalized Gram entries $|G_{ij}|/\sqrt{G_{ii}G_{jj}}$ between experts i and j of the layer whose ratio is the model's median, on a log colour scale; the diagonal, which equals one, is shown in grey. (e) Coupling ratio $\sum_{i\neq j}|G_{ij}|/\sum_i G_{ii}$ of each layer (dots) and its median over layers (bars), which we use as the model's coupling ratio. All statistics are computed on the 10^5-token REAP code mixture.

	Scales to E=256	Output coupling	Closed form	Post-pruning routing
Importance criteria: MAN, MSAN, MoNE, SE, REAP	✓	×	–	×
SHAPE (Zhang, 2026)	✓	×	×	×
Enumeration (Lu et al., 2024)	×	✓	×	✓
RECOUP (ours)	✓	✓	✓	×
RECOUP$_{\text{deploy}}$ (ours)	✓	✓	✓	✓

Table 1: Pruning criteria. –: no set objective.

We evaluate four models whose coupling ratios differ by more than a factor of four (Table 2). On Gemma-4-26B, which has the highest ratio and a parallel dense MLP, RECOUP delays the sharp drop in accuracy by one to two budget steps: at 25% it scores 59.2 on GSM8K against REAP's 15.6, and at 37.5% it scores 53.7 on HumanEval+ against 3.9. On Qwen3-30B, with an intermediate ratio, both forms are ahead of REAP on GSM8K at 25%, tied at 50% and above, and 4 to 6 GSM8K points behind at 37.5%. On Qwen3.6-35B, which has the lowest ratio and 256 experts, RECOUP$_{\text{deploy}}$ is ahead of REAP on both tasks at 12.5%, by 19.5 points on HumanEval+ and 30.0 on GSM8K. Because results at intermediate budgets depend on a small number of experts near the selection boundary, all of these values are means over three disjoint calibration samples (Tables 3 and 4).

2 Related Work

Importance criteria. Router-weighted expert activation pruning (REAP; Lasby et al., 2026) uses the gate-weighted activation norm averaged over each expert's tokens. The mean activation norm (MAN) and its squared variant (MSAN) omit the gate (Liu et al., 2026), MoNE (Zhang et al., 2026) combines routing frequency with output variance, and super-expert protection (SE; Su et al., 2026) retains experts with activation outliers. All of them score experts separately under the original routing (Table 1). Liu et al. (2026) express this family through frequency, gate and activation factors, with MAN and MSAN as the gate-free members, and state that the rerouting residual is ignored. SHAPE (Zhang, 2026) computes Shapley values over observed top-k coalitions. The interaction it models is routing co-occurrence rather than correlation in output space, and selection still assumes the original routing. REAP averages over tokens in order to retain rarely routed but consistent experts; our analysis of the orthogonal case shows when this choice is already optimal.

Reconstruction-based compression. Layer-output reconstruction is standard in weight pruning (Frantar & Alistarh, 2023). For experts, Lu et al. (2024) enumerate expert subsets per layer and keep the one with the smallest change in the layer output after pruning, which is an exhaustive

		Shared	Coupling ratio		Better form		Largest gain over REAP	
Model	E	path	Used	Domains	GSM8K	HumanEval+	Task, kept	Points
Gemma-4-26B	128	MLP	0.354	0.399–0.467	RECOUP	RECOUP	HumanEval+, 37.5%	+49.8
OLMoE-1B	64	none	0.163	0.156–0.158	–	–	–	–
Qwen3-30B	128	none	0.142	0.127–0.147	$\text{RECOUP}_{\text{deploy}}$	tie	GSM8K, 25%	+12.0
Qwen3.6-35B	256	expert	0.085	0.079–0.111	tie	$\text{RECOUP}_{\text{deploy}}$	GSM8K, 12.5%	+31.7

Table 2: Summary per model. Coupling ratio: on the REAP code mixture (used) and range over six single-domain calibration sets (Appendix B). Better form: the form of RECOUP with the higher mean at 25% and 12.5% in Tables 3 and 4 (tie: overlapping sample ranges; Qwen3.6-35B HumanEval+ from sample 1). Largest gain: of either form, mean over three calibration samples. OLMoE-1B: MC-8 only (Appendix L).

search over an objective of the kind $\text{RECOUP}_{\text{deploy}}$ evaluates. Enumeration is feasible with eight experts per layer, but not with 64 to 256 experts, and the closed form in Section 3 makes the objective tractable. Our contributions are the expert-level formulation, whose closed form shows that correlation between expert outputs is the relevant quantity, the analysis of the token weighting, and the objective under the routing of the pruned model ($\text{RECOUP}_{\text{deploy}}$).

Routing after compression. We treat re-selection as part of the computation that the criterion should evaluate. The closest analogue in weight pruning is SparseGPT (Frantar & Alistarh, 2023), which after each pruning step updates the remaining weights from calibration statistics, so that the next step is scored on the modified model. SparseGPT updates the weights, whereas $\text{RECOUP}_{\text{deploy}}$ updates the routing. Router calibration (Hyeon & Do, 2026) retrains the router by distillation after compression; it corrects the mismatch after selection, whereas we account for it during selection, and the two can be combined.

3 METHOD

3.1 THE RECOUP OBJECTIVE

A MoE layer outputs $y(x) = \sum_{e \in \text{top}k(x)} g_e(x) f_e(x)$ over its routed experts. Shared or parallel dense paths are not affected by expert pruning and cancel in every residual below, so they are excluded from $y(x)$. For a candidate set S, RECOUP computes the best linear combination of the kept experts under the original routing:

$$J(S) = \min_{\gamma} \sum_x \left\| y(x) - \sum_{i \in S \cap \text{top}k(x)} \gamma_i \, g_i(x) f_i(x) \right\|_2^2. \tag{1}$$

With $G[i,j] = \sum_x g_i g_j \langle f_i, f_j \rangle$ (nonzero only for co-routed pairs) and $h = G\mathbf{1}$, the residual is $J(S) = c - h_S^\top G_S^{-1} h_S$ with $c = \mathbf{1}^\top G \mathbf{1} = \sum_x \|y(x)\|^2$ (G includes a ridge of $10^{-6}\,\text{tr}(G)/E$), and greedy maximization of $V(S) = h_S^\top G_S^{-1} h_S$ with rank-one updates scales to E=256. V is not submodular in general, so greedy selection with swap refinement is a heuristic without a worst-case guarantee; Appendix H compares it with random-restart local search. Setting γ=$\mathbf{1}$ gives the score without compensation, so REAP-type criteria belong to this family.

3.2 TOKEN WEIGHTING AND THE ORTHOGONAL CASE

Lemma 1 (token weighting). For any token weights $w_x \geq 0$ the derivation is unchanged. The choice between summing and averaging in the literature (Lasby et al., 2026; Liu et al., 2026) is a choice of w (deployment weighting versus expert-uniform weighting), not of the aggregation rule. Selection is invariant to rescaling an individual expert (γ absorbs the scale), so within this family w is the only free choice. **Lemma 2 (orthogonal case).** If expert outputs are orthogonal, G is diagonal, $V(S) = \sum_{e \in S} G_{ee}$ is modular, greedy selection is optimal, and under expert-uniform weighting the ranking equals REAP's up to the moment order. RECOUP therefore contains REAP as a special case and differs from it only when the Gram matrix has off-diagonal mass.

3.3 SCORING UNDER THE ROUTING OF THE PRUNED MODEL

The linear objective above still assumes the original routing. RECOUP$_{\text{deploy}}$ evaluates the computation of the pruned model:

$$J^{\text{deploy}}(S) = \sum_x \left\| y(x) - \sum_{i \in \text{top}k(S,x)} \tilde{g}_i(x) f_i(x) \right\|_2^2, \tag{2}$$

where $\text{top}k(S, x)$ selects among the kept experts and $\tilde{g}$ are the renormalized gates. This is the inference computation itself, so there are no compensation coefficients γ. We do not rerun the router for every candidate S. For a candidate pool P of the $\min(2M, \frac{3}{4}E)$ experts ranked highest by REAP's saliency when at most half of the experts are kept (all E experts above that), we collect, in one pass under routing restricted to P with renormalized gates, the Gram matrix A of the reweighted expert outputs and their inner products b with y. For $S \subseteq P$ the objective is then evaluated as $c - 2\,\mathbf{1}^\top b_S + \mathbf{1}^\top A_S \mathbf{1}$, which equals Eq. 2 when $S = P$ and otherwise uses the routing of the pool as an approximation of the routing of S. Greedy selection with swap refinement is run inside P (Appendix I). The pool matters: widening it to all experts lowers the HumanEval+ score of RECOUP$_{\text{deploy}}$ on Qwen3.6-35B at 12.5% from 60.4 to 20.7 (Appendix I), and we did not tune its size.

3.4 THE COUPLING RATIO

The coupling ratio of layer ℓ is $\sum_{i \neq j} |G_\ell[i, j]| \,/\, \sum_i G_\ell[i, i]$ under sum weighting on the 10^5-token REAP code mixture of Section 4, and the ratio of a model is the median over its layers (Figure 1). It can be computed from calibration data before pruning, and it depends much more on the model than on the calibration domain: six single-domain calibration sets change a model's ratio by a factor of at most 1.4, while the models differ by a factor of about 4 to 5 (Table 2; Appendix B), so every calibration set gives the same ordering of the four models. No domain is consistently the most or least coupled: code is the lowest on Qwen3-30B, mathematics the highest on Qwen3-30B and Qwen3.6-35B, and tool calls the highest on Gemma-4-26B. On Qwen3-30B, C4 in twelve languages spans 0.104 (Hindi) to 0.137 (English), and two disjoint English C4 samples give 0.135 and 0.137. When the ratio is high, kept experts can partly reconstruct removed ones, and the reconstruction term of RECOUP is useful, especially at low budgets. When the ratio is low, the Gram matrix is close to diagonal, only the change in routing remains, and RECOUP$_{\text{deploy}}$, whose statistics are computed under the new routing, is the safer choice. With four models this gives an ordering rather than a threshold: on the models studied, the base form performed best at low budgets only at the highest ratio (0.354), and the deployed form was needed at 0.142 and below. We do not claim a threshold for other models. The evidence is observational: the parallel dense MLP of Gemma-4-26B varies together with the ratio, and we do not intervene on the architecture. The main tables report both forms under sum weighting, which we fix in advance because it weights tokens as they occur at deployment. Expert-uniform (mean) weighting is reported only as an ablation (Appendix C) and is never selected per setting.

4 EXPERIMENTAL SETUP

We use four models (Table 26, Figure 1; Muennighoff et al., 2024; Qwen Team, 2025; Gemma Team, 2026; Qwen Team, 2026) and refer to them by the short names in Table 26; results are ordered by coupling ratio. The budget is the percentage of routed experts kept in every layer, and we use seven budgets from 87.5% to 12.5%. Calibration data are matched to the task. For GSM8K we use the math calibration set (Tulu-3 personas-math); for HumanEval+ and the other code benchmarks we use the REAP code mixture, the calibration mixture released with REAP (evol-codealpaca, MoT code, math and science, xLAM and SWE-smith in equal shares, shuffled and packed). Each calibration set is divided into three disjoint samples of $102{,}400$ tokens (chunks 0–49, 100–149 and 200–249 of $2{,}048$ tokens), which we call calibration samples 1, 2 and 3. Results from a single sample use sample 1; repeated results report the mean $\pm$ half-range over the three samples, that is, the mean and half the difference between the largest and smallest value. REAP is recomputed on the same tokens for every task; its full released calibration set is compared in Appendix F. We evaluate GSM8K (8-shot, strict match; Cobbe et al., 2021) and HumanEval+ (greedy pass@1, EvalPlus; Chen et al., 2021; Liu et al., 2023) on exported pruned checkpoints, and the average accuracy on eight multiple-choice tasks (MC-8: ARC-challenge, ARC-easy, BoolQ, HellaSwag, MMLU, OpenBookQA, RTE

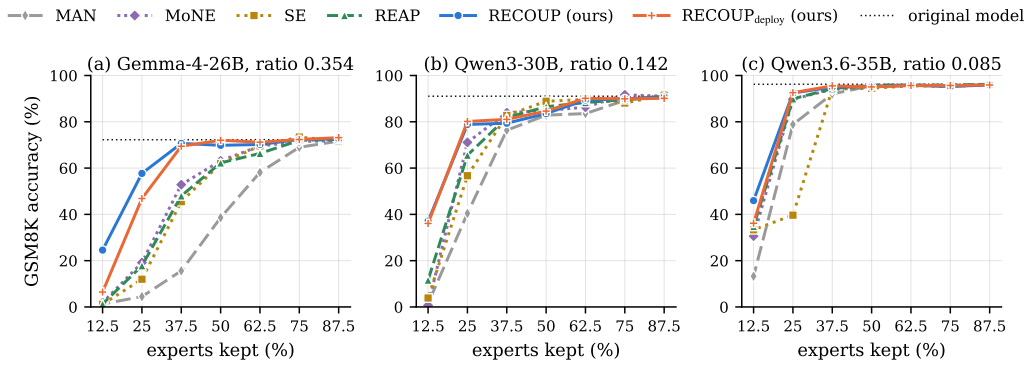


Figure 2: GSM8K accuracy against the percentage of experts kept, with the math calibration set and calibration sample 1; panels ordered by coupling ratio. The dotted line is the original model. Exact values are in Tables 20, 21 and 22.

and WinoGrande). We also evaluate MBPP+ (Austin et al., 2021), MATH-500 (Hendrycks et al., 2021b; Lightman et al., 2024) with Minerva 4-shot `math_verify`, LiveCodeBench v6 (Jain et al., 2025) with the official n=10 sampling, and BFCL v4 (Patil et al., 2025) on Qwen3-30B with the REAP code mixture, and LiveCodeBench on Gemma-4-26B (Section 5.3). We call RECOUP, the objective under the original routing, the base form, and RECOUP$_{\text{deploy}}$, the objective under the routing of the pruned model, the deployed form; both use sum weighting, and mean-weighted variants appear only in Appendix C. The baselines SE, MoNE and MAN are implemented from their sources (activation Eq. 6 for SE; Eqs. 4–7 of MoNE; the gate-free member of Liu et al., 2026 for MAN) and use the same tokens and protocol. MSAN is not evaluated. In all tables, bold marks the best criterion in each setting (all tied values are bold, and 0.0 is never bold).

5 Results

We report GSM8K and then HumanEval+ on the three models evaluated on generation, as means over three calibration samples, followed by secondary benchmarks, the variation across calibration samples and selection cost. Table 2 summarizes the results per model. At 12.5% on the REAP code mixture the two forms share 57% of the kept experts on Gemma-4-26B, 64% on Qwen3-30B, 87% on Qwen3.6-35B and 13% on OLMoE-1B.

5.1 GSM8K

Table 3 gives GSM8K over three calibration samples at the three budgets at which the criteria differ most, and Figure 2 shows all seven budgets on calibration sample 1 (values in Tables 20, 21 and 22). On Gemma-4-26B, where the coupling ratio is highest and Lemma 2 does not apply, RECOUP delays the sharp drop by one to two budget steps. At 25% no importance criterion (REAP, SE, MoNE or MAN) exceeds 18.2, and RECOUP is 43.6 points above REAP (59.2 ± 2.4 vs. 15.6 ± 2.4); at 12.5% the gap is 23.3 points on calibration sample 1 ($P < 10^{-4}$, paired permutation test). RECOUP$_{\text{deploy}}$ is equal or better at intermediate budgets (50–75%, where it matches the original model) but worse at the lowest budgets: with large off-diagonal mass, the Gram matrix under the original routing already describes how kept experts can reconstruct removed ones, and the base form performs best, as expected for a high-ratio model. A diagonal-only control, which sets the off-diagonal Gram entries to zero, is within the sample range of the base form on GSM8K at 50 and 37.5% (37.5% only on Qwen3.6-35B) and on Gemma-4-26B HumanEval+ (Table 5; Appendix D), so in these settings the gain over REAP comes from the summed diagonal rather than from the coupling term. The deployed form differs from the control on Qwen3-30B HumanEval+ at 25% (38.6 ± 14.0 vs. 15.0 ± 9.4). On Qwen3-30B, with an intermediate ratio, both forms show large gains at low budgets, are tied with REAP at 50% and above, and lose to REAP at 37.5% by 4 to 6 points with non-overlapping sample ranges (Table 3). At 25% RECOUP$_{\text{deploy}}$ is 12.0 points above REAP (80.5 ± 0.8 vs. 68.5 ± 2.4), and at 12.5% the two forms are 25.4 and 24.5 points above REAP

Model	Experts kept	GSM8K accuracy					
		MAN	MoNE	SE	REAP	RECOUP	RECOUP$_{\text{deploy}}$
Gemma-4-26B	original				72.3		
	50%	41.1 ±3.0	63.0 ±0.6	59.7 ±2.3	61.7 ±0.8	69.5 ±1.2	**71.1** ±1.3
	37.5%	16.2 ±2.1	52.3 ±0.8	43.6 ±3.7	48.1 ±1.0	**70.2** ±0.5	69.2 ±0.5
	25%	5.1 ±0.6	18.2 ±0.8	13.9 ±4.1	15.6 ±2.4	**59.2** ±2.4	46.5 ±2.2
Qwen3-30B	original				91.1		
	50%	82.3 ±1.7	86.8 ±1.8	86.6 ±2.2	86.5 ±0.2	86.2 ±2.1	**88.1** ±2.7
	37.5%	76.6 ±1.1	**84.6** ±0.7	83.7 ±0.8	84.1 ±1.9	80.1 ±0.7	78.1 ±3.8
	25%	43.1 ±2.1	72.2 ±1.7	58.9 ±3.1	68.5 ±2.4	78.7 ±0.4	**80.5** ±0.8
Qwen3.6-35B	original				96.2		
	37.5%	91.8 ±0.3	94.1 ±0.2	94.2 ±0.3	94.5 ±0.2	**95.3** ±0.3	95.2 ±0.3
	25%	80.6 ±1.4	88.5 ±2.2	45.9 ±7.3	88.4 ±2.2	92.7 ±0.3	**93.0** ±0.4
	12.5%	5.6 ±6.4	30.2 ±11.2	12.7 ±16.0	24.2 ±11.6	**55.9** ±9.5	54.2 ±18.1

Table 3: GSM8K accuracy, mean ± half-range over three calibration samples (math calibration set), at the three budgets where the criteria differ most.

on calibration sample 1 (37.0 and 36.1 vs. 11.6, $P < 10^{-4}$). The strongest baseline at 25%, MoNE (72.2 ± 1.7 over three samples, Appendix E), is 8 points below RECOUP$_{\text{deploy}}$ (80.5 ± 0.8) with non-overlapping ranges. On Qwen3.6-35B, which has the lowest ratio, the two forms are equivalent across calibration samples (92.7 ± 0.3 vs. 93.0 ± 0.4 at 25%, 55.9 ± 9.5 vs. 54.2 ± 18.1 at 12.5%, Table 3), both are above REAP at 25% (88.4 ± 2.2) and at 12.5% (24.2 ± 11.6) on the mean over three calibration samples, and the base form scores 45.9 against REAP's 34.8 on the first calibration sample. The result at 12.5% depends on the second and third samples: on the first calibration sample the deployed form is 1.4 points ahead (36.2 vs. 34.8, $P = 0.46$), and at this budget the samples of every criterion vary widely (MoNE 30.2 ± 11.2, SE 12.7 ± 16.0).

On Qwen3.6-35B we also observed the following. With the shared few-shot completion prompt and no chat template, twenty-one selections at 25% that differ in at most 0.66% of the experts, nineteen of which have GSM8K samples (Appendix M), scored between 55 and 83, and the difference followed a single expert, $(L9, E2)$. Selections containing it opened a reasoning block in 25 to 54% of responses, and the others in 63 to 94%; responses that open the block are usually cut off at the token limit. With the chat template and reasoning disabled, the nineteen selections with GSM8K samples score 92.0 to 93.2 (Appendix M). The expert determines which mode the model answers in rather than its arithmetic ability, and we therefore evaluate GSM8K on Qwen3.6-35B with reasoning disabled throughout.

5.2 HumanEval+

Table 4 gives HumanEval+ over three calibration samples with the REAP code mixture, and Table 23 gives all seven budgets on calibration sample 1. On Gemma-4-26B, RECOUP retains 53.7 at 37.5% and 39.0 at 25%, where no importance criterion exceeds 10.4 and 0.4. Its gaps over REAP are 49.8 and 39.0 points, with non-overlapping ranges, and the deployed form is behind the base form at these two budgets, as on GSM8K. On Qwen3-30B the results vary strongly with the calibration sample. A 10^5-token sample of the REAP mixture contains only nine to eleven documents, and at 25% every importance criterion drops sharply in at least one sample (REAP 45.7/18.3/21.3, MAN 18.9/0.0/3.0; Table 19). RECOUP$_{\text{deploy}}$ is above REAP in all three samples (55.5/27.4/32.9, mean difference 10.2), whereas the base form is tied with REAP on the mean, and at 37.5% REAP and MoNE are ahead of both forms. We therefore report the Qwen3-30B code result only as a result for the deployed form at 25%. A diagonal-only control (Appendix D) shows that this is the setting in which the statistics under the new routing differ from the Gram diagonal. On Qwen3.6-35B, RECOUP$_{\text{deploy}}$ is 2.4 points behind REAP at 37.5%, ahead at 25% within the sample ranges (85.2 vs. 82.3), and at 12.5% gives the largest difference on this model, 56.9 ± 3.7 against 37.4 for both REAP and MoNE, with non-overlapping ranges. The base form, evaluated on calibration sample 1 only, scores 17.1 at 12.5% against REAP's 37.8: the reconstruction term has little effect, as Lemma 2

Model	Experts kept	HumanEval+ pass@1					
		MAN	MoNE	SE	REAP	RECOUP	$\text{RECOUP}_{\text{deploy}}$
Gemma-4-26B	original				57.9		
	50%	17.7 ±2.1	37.0 ±6.1	29.5 ±10.1	32.9 ±2.4	53.9 ±4.0	**56.5** ±3.0
	37.5%	1.8 ±1.5	10.4 ±6.4	7.9 ±4.0	3.9 ±3.7	**53.7** ±4.9	46.1 ±7.9
	25%	0.0	0.4 ±0.6	0.2 ±0.3	0.0	**39.0** ±11.6	17.7 ±9.4
Qwen3-30B	original				87.8		
	37.5%	61.6 ±13.4	**76.8** ±4.6	58.1 ±10.7	**76.8** ±4.0	60.2 ±18.3	65.0 ±16.5
	25%	7.3 ±9.4	13.8 ±4.9	9.6 ±9.8	28.5 ±13.7	27.2 ±18.0	**38.6** ±14.0
	12.5%	0.2 ±0.3	0.2 ±0.3	0.0	0.0	**4.5** ±2.7	3.0 ±2.4
Qwen3.6-35B	original				90.2		
	37.5%	88.2 ±0.6	89.8 ±0.6	87.4 ±1.2	**90.2** ±0.9	–	87.8 ±1.8
	25%	78.5 ±1.5	81.3 ±4.6	76.8 ±3.7	82.3 ±2.4	–	**85.2** ±4.3
	12.5%	26.8 ±0.6	37.4 ±7.0	11.6 ±5.5	37.4 ±2.4	–	**56.9** ±3.7

Table 4: HumanEval+ pass@1, mean ± half-range over three calibration samples (REAP code mixture). –: sample 1 only (Table 23).

Model	Kept	REAP	Diagonal only	RECOUP	$\text{RECOUP}_{\text{deploy}}$
Gemma-4-26B	50%	61.7 ± 0.8	69.6 ± 1.7	69.5 ± 1.2	71.1 ± 1.3
	37.5%	48.1 ± 1.0	69.5 ± 1.3	70.2 ± 0.5	69.2 ± 0.5
	25%	15.6 ± 2.4	60.0 ± 1.3	59.2 ± 2.4	46.5 ± 2.2
Qwen3-30B	50%	86.5 ± 0.2	88.2 ± 0.8	86.2 ± 2.1	88.1 ± 2.7
	37.5%	84.1 ± 1.9	78.5 ± 1.9	80.1 ± 0.7	78.1 ± 3.8
	25%	68.5 ± 2.4	79.3 ± 1.8	78.7 ± 0.4	80.5 ± 0.8
Qwen3.6-35B	37.5%	94.5 ± 0.2	95.3 ± 0.1	95.3 ± 0.3	95.2 ± 0.3
	25%	88.4 ± 2.2	92.7 ± 0.5	92.7 ± 0.3	93.0 ± 0.4
	12.5%	24.2 ± 11.6	42.4 ± 8.9	55.9 ± 9.5	54.2 ± 18.1

Table 5: Diagonal-only control on GSM8K (math calibration set): the base objective with off-diagonal Gram entries set to zero. Mean ± half-range over three calibration samples.

predicts for nearly orthogonal experts. When the coupling ratio is low, the change in routing is the main remaining source of error, and accounting for it helps.

The choice of token weighting has little effect on the intermediate model above 12.5%: sum and mean weighting of $\text{RECOUP}_{\text{deploy}}$ are within 3 points at five or more of seven budgets on both tasks (Appendix C). Over the settings in Tables 3 and 4, the largest gap to the best criterion is 43.3 to 54.1 points for each importance criterion and 21.3 points for $\text{RECOUP}_{\text{deploy}}$, which is at most 11.8 points below the best on the two Qwen models; the larger gap on Gemma-4-26B is to the base form, which performs best there at the lowest budgets.

5.3 Secondary benchmarks

Figure 3 extends the comparison to MBPP+ and MATH-500. On Qwen3-30B the pattern is the same as on the main tasks: ties at high budgets and a gap that increases as the budget decreases, with the largest difference at 25%. There, MBPP+ is 55.3 for $\text{RECOUP}_{\text{deploy}}$ against REAP's 35.5 (19.8 points, $P < 10^{-4}$, paired permutation test over 378 problems; the best importance criterion reaches 40.0), and MATH-500 is 72.8 against REAP's 54.0 and MoNE's 62.6. At 12.5% MBPP+ is 9.3 and 8.5 for the two forms, while every importance criterion is at most 0.3 (Table 16). On LiveCodeBench and BFCL the gains are smaller or absent. On BFCL, where tool-call formatting is relatively robust to pruning, the criteria are within a few points of one another at 75–50%. On LiveCodeBench the deployed form is 9.2 points above REAP at 37.5% (38.1 vs. 28.9, MoNE 32.4), and at 25% every criterion fails on this model (best score 4.7). On Gemma-4-26B LiveCodeBench does separate them (Table 6). At 50% both forms retain about 17.5 pass@1 of the original model's 23.7 (18.0 and 17.3), while the best importance criterion, MoNE, retains 6.1 and REAP 4.5. At

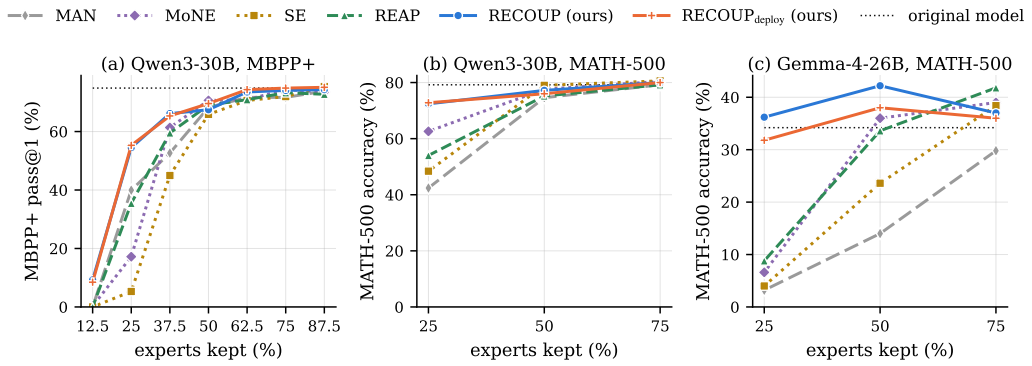


Figure 3: (a) MBPP+ on Qwen3-30B at seven budgets with the REAP code mixture; (b) MATH-500 on Qwen3-30B and (c) MATH-500 on Gemma-4-26B at three budgets with the math calibration set. Calibration sample 1; the dotted line is the original model. In (c) the ordering at 75% and 50% is within the range over calibration samples (Appendix F). Exact values are in Table 16, and LiveCodeBench v6 and BFCL v4 in Table 17.

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				23.7			
MAN	18.0	13.3	1.5	0.2	0.0	0.0	0.0
MoNE	20.7	21.5	15.5	6.1	0.4	0.0	0.0
SE	17.5	14.3	5.5	0.9	0.1	0.0	0.0
REAP	20.4	19.6	13.8	4.5	0.1	0.0	0.0
RECOUP (ours)	22.8	**21.6**	18.9	**18.0**	10.1	**3.2**	0.0
RECOUP$_{deploy}$ (ours)	**23.0**	21.3	**19.4**	17.3	**13.1**	0.2	0.0

Table 6: LiveCodeBench v6 pass@1 on Gemma-4-26B (REAP code mixture, sample 1, n=10). Prompt format in Appendix N.

37.5% the base form scores 10.1 and the deployed form 13.1, while every importance criterion is at most 0.4. At 25% only the base form is above one point (3.2), and at 12.5% every criterion scores zero.

On Gemma-4-26B, which has the highest coupling ratio, MATH-500 shows a larger separation (Figure 3c). At 25% RECOUP retains 36.2, while REAP drops to 8.8 and no importance criterion exceeds 8.8, a four-fold difference on a benchmark that neither objective used. MBPP+ could not be evaluated on Gemma-4-26B: under the EvalPlus protocol the model emits function bodies without signatures, so every criterion and the original model fail execution-based scoring for formatting reasons. We exclude it.

5.4 Variation across calibration samples

On GSM8K (Table 3), results vary across calibration samples by at most ±4.1 points on Gemma-4-26B and ±3.8 on Qwen3-30B; on Qwen3.6-35B above 12.5%, SE varies by up to ±7.3 and the other criteria by at most ±2.2. On HumanEval+ the variation is larger, up to ±18.3 on Qwen3-30B (Table 4), because each 10^5-token code sample holds few documents (Appendix E). A single calibration sample is therefore not reliable at these budgets: SE on Qwen3.6-35B at 25% ranges over 14.6 points between samples. Orderings within the range over calibration samples are treated as ties throughout. The low-budget results above hold in every sample, with two limits: the GSM8K result of RECOUP$_{deploy}$ on Qwen3.6-35B at 12.5% is above REAP's sample range by only 0.3 points, and on Qwen3-30B HumanEval+ at 25% only the deployed form is above REAP in every sample.

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				58.0			
REAP	58.4	54.6	47.2	40.3	35.7	34.4	34.4
RECOUP (ours)	57.4	52.1	47.5	42.0	36.9	35.3	35.3
$\text{RECOUP}_{\text{deploy}}$ (ours)	57.4	52.1	47.5	42.1	37.5	34.0	33.9

Table 7: MC-8 accuracy on OLMoE-1B, mean over three calibration samples. All differences are within 2.5 points.

	RECOUP	$\text{RECOUP}_{\text{deploy}}$
Statistics	1 pass, 10^5 tokens	1 pass per pool, $\approx$5 GPU-min
Gram matrix	0.5–20 MB	0.5–20 MB
Selection	$<$1 s, CPU	$<$1 s, CPU
Memory / time per layer	$O(E^2)$ / $O(ME^2)$	$O(E^2)$ / $O(ME^2)$

Table 8: Selection cost. Neither form needs training. Pool time measured on Qwen3.6-35B.

The fourth model, OLMoE-1B (E=64), is evaluated on multiple-choice tasks under the same protocol with three calibration samples. Differences stay within 2.5 points and sample ranges touch or overlap, so this model with fewer, larger experts shows no dependence on the coupling ratio; the base form is not below REAP from 50% down (Table 7). At 12.5% the two forms share only 13% of the kept experts on this model, yet their MC-8 accuracies differ by 1.4 points.

5.5 Selection cost

Table 8 gives the cost of both forms. Each rank-one update of G_S^{-1} costs $O(|S|^2)$ and every candidate is rescored at each step, which gives the per-layer complexity. On Qwen3.6-35B one pass with masked routing takes about five minutes, so the four pools that cover seven budgets require about twenty GPU-minutes. Neither form requires a backward pass or training, and both are cheaper than the export and evaluation steps that follow. For comparison, exhaustive enumeration (Lu et al., 2024) evaluates every kept set of a layer: $\binom{8}{4} = 70$ sets for eight experts at 50%, but about 10^{30} sets for E=128 at 25% and about 10^{41} for E=256 at 12.5%, which is why it is marked infeasible in Table 1. When the coupling ratio is high, RECOUP obtains its gains at low budgets with a single forward pass, and the additional cost of $\text{RECOUP}_{\text{deploy}}$ is a small constant.

6 Conclusion

RECOUP scores a kept set by how well it reconstructs the outputs of the removed experts, and $\text{RECOUP}_{\text{deploy}}$ computes the same statistics under the routing of the pruned model. On four models, the coupling ratio indicated before pruning which form performs best at low budgets. On the model with the highest ratio, Gemma-4-26B, the base form retains most of the original model's accuracy on GSM8K and HumanEval+ at budgets at which every importance criterion fails, and it separates the criteria on MATH-500 and LiveCodeBench as well. On the model with the lowest ratio, Qwen3.6-35B, the deployed form gives the largest gains at the smallest budget. On Qwen3-30B, with an intermediate ratio, both forms are ahead of REAP on GSM8K at 25%, the deployed form by more. Both forms require only calibration statistics and no training, so the coupling ratio can be computed first and the form chosen before pruning. Appendix A discusses the limitations of this study, and we release the kept sets, the raw evaluation outputs and the scripts that produce each table and figure.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, 2019.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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), 2022.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation. Zenodo, `https://doi.org/10.5281/zenodo.10256836`, 2023.

Gemma Team. Gemma 4. `https://huggingface.co/google/gemma-4-26b-a4b`, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021b.

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

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

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 *International Conference on Learning Representations*, 2026.

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*, 2022.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

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*, 2023.

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*, 2026.

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*, 2024.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. WizardCoder: Empowering code large language models with Evol-Instruct. In *International Conference on Learning Representations*, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of EMNLP*, 2018.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

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 *International Conference on Machine Learning*, 2025.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.6-35b-a3b. `https://huggingface.co/Qwen/Qwen3.6-35B-A3B`, 2026.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

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.

Zunhai Su, Qingyuan Li, Hao Zhang, Weihao Ye, Qibo Xue, Yulei Qian, Yuchen Xie, Ngai Wong, and Kehong Yuan. Unveiling super experts in mixture-of-experts large language models. In *International Conference on Learning Representations*, 2026.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Geng Zhang, Yuxuan Han, Yuxuan Lou, Yiqi Zhang, Wangbo Zhao, and Yang You. MoNE: Replacing redundant experts with lightweight novices for structured pruning of MoE. In *International Conference on Learning Representations*, 2026.

Yuhao Zhang. SHAPE: Coalition-aware expert pruning for sparse mixture-of-experts LLMs. *arXiv preprint arXiv:2606.09886*, 2026.

Reproducibility Statement

Every number in this paper is regenerated from artifacts we ship as supplementary material: the per-layer kept sets of every criterion, model and budget, the raw evaluation outputs, and the scripts that produce each table and figure from them. The models are public checkpoints (Table 26); the calibration sets, their sources, chunking and the three disjoint calibration samples are specified in Section 4 and Appendix N. The objective and its closed form are derived in Section 3, with proofs of the two lemmas in Appendix H; the candidate pool and swap refinement of $\text{RECOUP}_{\text{deploy}}$ are specified in Appendix I, and selection is deterministic given the calibration set. Baseline implementations, generation and multiple-choice protocols with their decoding settings, the contamination check between calibration sets and evaluation items, the compute used, and the run-to-run variation of the generation evaluations are given in Appendix N.

AI Use Statement

Large language models (Anthropic's Claude, accessed through the Claude Code agentic interface) were used in this work under the authors' direction, and the authors verified every result. They were used for retrieving and reading related work, each characterization of which the authors checked against the original paper; for writing and running the selection, analysis and evaluation code that produced the kept sets, statistics and tables; for proposing experiments and controls, among which the authors chose; for drafting sections of the paper and the appendix from the authors' outline and results, which the authors edited; for checking the derivations in Section 3 and Appendix H; and for creating and editing figures. No synthetic data were generated. The research questions, claims, experimental design and interpretation of results are the authors'. The authors take full responsibility for the content.

A Limitations

Observational evidence for the coupling ratio. The relation between the coupling ratio and the better form is observed across four architectures, and the parallel dense MLP varies together with the ratio. We do not intervene on the architecture, and four models give an ordering rather than a threshold (Section 3). A diagonal-only control (Appendix D) is within the sample range of the base form at 50 and 37.5% on GSM8K for all three models and on Gemma-4-26B HumanEval+, so at these budgets the base form's gain over REAP comes from the summed diagonal rather than from the coupling term; the deployed form differs from both on Qwen3-30B HumanEval+ at 25%. **Dependence on the candidate pool.** $\text{RECOUP}_{\text{deploy}}$ requires statistics under masked routing for each candidate pool, which is more expensive than saliency but bounded (Section 5.5), and it selects within a pool of at most $2M$ experts ranked by saliency. Widening the pool to all experts lowers its HumanEval+ score on Qwen3.6-35B at 12.5% from 60.4 to 20.7 (Appendix I), so the gains of the deployed form depend on the pool, whose size we did not tune. **Benchmark coverage.** MBPP+, MATH-500, LiveCodeBench and BFCL are evaluated on Qwen3-30B, and MATH-500 and LiveCodeBench on Gemma-4-26B. MBPP+ and BFCL cannot be evaluated on Gemma-4-26B because of output and tool-call formatting. All results are one-shot, with uniform budgets and without recovery training. On multiple-choice tasks the advantage is limited to the smallest budgets, and norm-based criteria are better above them (Appendix K). **Calibration-sample noise.** Three calibration samples bound the selection noise but do not estimate it. We report half-ranges and treat orderings within them as

ties. On code, the REAP mixture contains too few documents per 10^5-token sample for Qwen3-30B's low-budget HumanEval+ results to be consistent across calibration samples for any criterion; only the lead of $\text{RECOUP}_{\text{deploy}}$ over REAP at 25% holds in every sample.

B Coupling ratio per model and per calibration domain

Table 9 lists the per-model coupling ratio of Figure 1 with its range over layers. The ratio is $\sum_{i \neq j} |G_\ell[i,j]| / \sum_i G_\ell[i,i]$ per layer with sum weighting on the 10^5-token REAP code mixture, and the model value is the median over layers.

Model	Median	Min	Max
Gemma-4-26B	0.354	0.003	0.929
OLMoE-1B	0.163	0.004	1.026
Qwen3-30B	0.142	0.001	1.545
Qwen3.6-35B	0.085	0.052	0.378

Table 9: Coupling ratio over layers (Figure 1).

Dependence on the calibration domain. Figure 4 and Table 10 recompute the ratio, with the same definition and the same 10^5-token budget, on six single-domain calibration sets per model: English web text (C4), Korean web text (C4), creative writing (WritingPrompts), tool calls (xlam-function-calling), code (evol-codealpaca) and mathematics (Tulu-3 personas-math), and for Qwen3-30B additionally on C4 in twelve languages (Table 11). OLMoE-1B has three of the six sets. The ratio depends much more on the model than on the calibration set. Within a model the six domain medians differ by a factor of at most 1.4 (Qwen3.6-35B, 0.079 to 0.111) and by 1.2 or less on the other three, whereas across models the same domain differs by a factor of 3.9 to 5.4. No domain is consistently the most or least coupled: code is the lowest on Qwen3-30B and the second lowest on Gemma-4-26B, mathematics the highest on Qwen3-30B and Qwen3.6-35B, and tool calls the highest on Gemma-4-26B. The REAP-mixture values of Figure 1 (dashed lines) lie inside the domain band on the two Qwen models, below it on Gemma-4-26B (0.354 vs. 0.399–0.467) and just above it on OLMoE-1B (0.163 vs. 0.156–0.158). On Qwen3-30B, C4 in twelve languages spans 0.104 (Hindi) to 0.137 (English), a factor of 1.3, and two disjoint English C4 samples give 0.135 and 0.137. Any of these calibration sets therefore gives the same ordering of the four models.

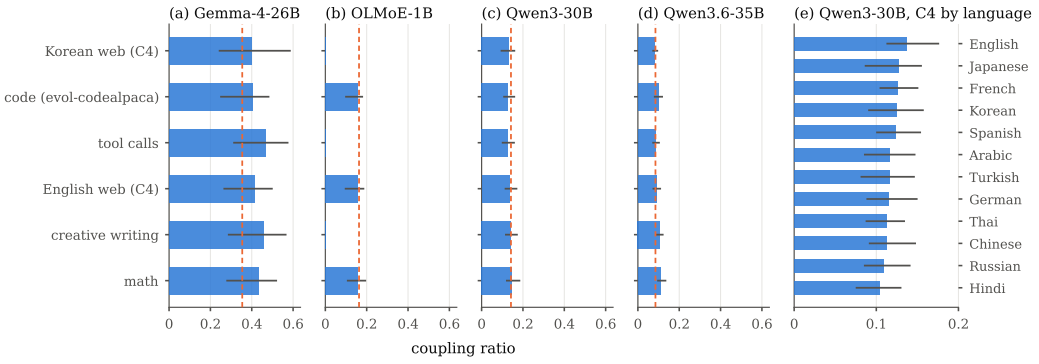


Figure 4: Coupling ratio per calibration set. (a–d) Median over layers (bars) and interquartile range over layers (lines) on six single-domain calibration sets of 10^5 tokens; OLMoE-1B has three of them. The dashed line is the model's value on the REAP code mixture (Figure 1). (e) Qwen3-30B on C4 in twelve languages.

Calibration set	Gemma-4-26B	OLMoE-1B	Qwen3-30B	Qwen3.6-35B
Korean web (C4)	0.399 (0.081–0.995)	–	0.132 (0.002–1.067)	0.079 (0.032–0.810)
Code (evol-codealpaca)	0.402 (0.092–0.723)	0.156 (0.004–0.995)	0.127 (0.001–2.114)	0.100 (0.055–0.629)
Tool calls	0.467 (0.137–0.920)	–	0.129 (0.002–1.703)	0.086 (0.047–0.514)
English web (C4)	0.415 (0.103–0.791)	0.156 (0.004–0.914)	0.135 (0.002–2.531)	0.091 (0.049–0.750)
Creative writing	0.460 (0.110–0.814)	–	0.143 (0.002–2.692)	0.107 (0.056–0.849)
Mathematics	0.434 (0.137–0.756)	0.158 (0.004–0.791)	0.147 (0.001–1.733)	0.111 (0.068–0.471)

Table 10: Coupling ratio by calibration set: median over layers (min–max). –: not measured.

Language	Coupling ratio	Language	Coupling ratio
Hindi	0.104 (0.002–1.359)	Arabic	0.117 (0.002–0.685)
Russian	0.109 (0.002–1.414)	Spanish	0.123 (0.002–1.979)
Chinese	0.112 (0.002–1.349)	Korean	0.125 (0.002–0.986)
Thai	0.113 (0.002–0.783)	French	0.126 (0.002–2.022)
German	0.115 (0.002–1.769)	Japanese	0.127 (0.002–0.679)
Turkish	0.116 (0.002–1.147)	English	0.137 (0.002–2.560)

Table 11: Coupling ratio of Qwen3-30B on C4 by language (Figure 4e): median over layers (min–max).

C TOKEN-WEIGHTING ABLATION

The main text uses sum weighting. Tables 12, 13 and 14 add expert-uniform (mean) weighting for $\text{RECOUP}_{\text{deploy}}$ under identical protocols. Consistent with Lemma 1, the choice has little effect where the coupling ratio is intermediate (Qwen3-30B, within 3 points at every budget from 87.5 to 25% except GSM8K at 37.5%, where mean leads by 5.1; at 12.5% sum leads by 17.2 on GSM8K and 4.3 on HumanEval+). On Qwen3.6-35B, which has the lowest ratio, the two weightings agree on GSM8K down to 25% (both 92.6); at 12.5% the first calibration sample gives sum 36.2 against mean 62.9, but the sum weighting's three samples there span 36 to 72, so we draw no conclusion at that budget. On Gemma-4-26B, which has the highest ratio, the two weightings stay within about two points on GSM8K down to 25% (46.9 vs. 44.6) and 12.5% (6.4 vs. 4.3), while on HumanEval+ sum weighting is better at 37.5% and 25% (43.3 vs. 32.3, 10.4 vs. 0.0).

GSM8K, math calibration set

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				72.3			
RECOUP (sum)	72.7	72.3	70.2	69.8	**70.7**	**57.7**	**24.6**
$\text{RECOUP}_{\text{deploy}}$ (mean)	72.2	**72.8**	**72.7**	**72.5**	69.9	44.6	4.3
$\text{RECOUP}_{\text{deploy}}$ (sum)	**73.2**	72.3	71.2	72.0	69.5	46.9	6.4

HumanEval+, REAP code mixture

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				57.9			
RECOUP (sum)	59.8	**59.8**	55.5	**57.3**	**48.8**	**25.0**	0.0
$\text{RECOUP}_{\text{deploy}}$ (mean)	54.9	58.5	**61.0**	52.4	32.3	0.0	0.0
$\text{RECOUP}_{\text{deploy}}$ (sum)	**61.6**	56.7	54.9	54.9	43.3	10.4	0.0

Table 12: Token weighting on Gemma-4-26B (Lemma 1), sample 1. Sum is used in the rest of the paper.

GSM8K, math calibration set

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				91.1			
RECOUP (sum)	**90.8**	**89.9**	89.4	83.5	79.4	78.8	**37.0**
$\text{RECOUP}_{\text{deploy}}$ (mean)	**90.8**	88.8	**90.3**	81.7	**86.1**	**80.2**	18.9
$\text{RECOUP}_{\text{deploy}}$ (sum)	90.1	**89.9**	90.1	**84.6**	81.0	80.1	36.1

HumanEval+, REAP code mixture

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				87.8			
RECOUP (sum)	**89.6**	86.0	86.0	83.5	**82.3**	48.8	1.8
$\text{RECOUP}_{\text{deploy}}$ (mean)	89.0	**89.0**	**87.2**	**86.6**	79.9	54.3	0.6
$\text{RECOUP}_{\text{deploy}}$ (sum)	88.4	86.6	85.4	86.0	81.1	**55.5**	**4.9**

Table 13: Token weighting on Qwen3-30B, as in Table 12.

GSM8K, math calibration set

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				96.2			
RECOUP (sum)	**96.0**	95.2	95.8	**95.2**	95.0	92.4	45.9
$\text{RECOUP}_{\text{deploy}}$ (mean)	95.6	**96.0**	**96.2**	**95.2**	94.6	**92.6**	**62.9**
$\text{RECOUP}_{\text{deploy}}$ (sum)	95.9	95.7	95.7	95.1	**95.5**	**92.6**	36.2

HumanEval+, REAP code mixture

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				90.2			
RECOUP (sum)	90.2	90.8	90.8	**92.7**	87.8	**88.4**	17.1
$\text{RECOUP}_{\text{deploy}}$ (mean)	**90.8**	**92.1**	**91.5**	87.8	**90.2**	87.2	40.9
$\text{RECOUP}_{\text{deploy}}$ (sum)	90.2	90.8	89.0	92.1	86.0	**88.4**	**60.4**

Table 14: Token weighting on Qwen3.6-35B, as in Table 12.

D DIAGONAL-ONLY CONTROL

Setting every off-diagonal entry of the Gram matrix to zero turns the base objective into a ranking by $G_{ee} = \sum_x g_e(x)^2 \|f_e(x)\|^2$ (Lemma 2), so the kept set is chosen by each expert's own summed energy without the coupling term. Tables 5 and 15 place this control next to REAP and the two RECOUP forms under the protocol with three calibration samples, on GSM8K for the three generative models and on HumanEval+ for Gemma-4-26B and Qwen3-30B. At 50 and 37.5% the control lies within the sample range of the base form on Gemma-4-26B (both tasks), on Qwen3-30B (GSM8K) and on Qwen3.6-35B (GSM8K), so the base form's margin over REAP at those budgets is carried by the diagonal, that is by summing each expert's energy over its tokens rather than averaging it, and the off-diagonal term does not separate from the diagonal within sample-to-sample noise there. The deployed form separates from both the control and the base form in every sample at 25% on Qwen3-30B HumanEval+ ($55.5/27.4/32.9$ against $21.9/3.0/20.1$ and $48.8/12.8/20.1$), the only setting in which the statistics under the new routing differ from the diagonal under the original routing. At 12.5% the sample ranges of every criterion overlap and we draw no conclusion.

Model	Kept	REAP	Diagonal only	RECOUP	$\text{RECOUP}_{\text{deploy}}$
Gemma-4-26B	50%	32.9 ± 2.4	49.8 ± 8.8	53.9 ± 4.0	56.5 ± 3.0
	37.5%	3.9 ± 3.7	57.1 ± 5.8	53.7 ± 4.9	46.1 ± 7.9
	25%	0.0	35.2 ± 13.7	39.0 ± 11.6	17.7 ± 9.4
	12.5%	0.0	2.0 ± 2.1	7.1 ± 5.5	0.0
Qwen3-30B	37.5%	76.8 ± 4.0	72.4 ± 16.2	60.2 ± 18.3	65.0 ± 16.5
	25%	28.5 ± 13.7	15.0 ± 9.4	27.2 ± 18.0	38.6 ± 14.0
	12.5%	0.0	4.3 ± 0.6	4.5 ± 2.7	3.0 ± 2.4

Table 15: Diagonal-only control on HumanEval+ (REAP code mixture), as in Table 5.

E Repeated calibration samples on HumanEval+

Table 19 gives the HumanEval+ result of each calibration sample behind the means in Table 4, on Qwen3-30B and, at 12.5%, on Qwen3.6-35B. On Qwen3.6-35B at 12.5%, $\text{RECOUP}_{\text{deploy}}$ (56.9 ± 3.7) is above the range of every baseline (SE 11.6 ± 5.5, MoNE 37.4 ± 7.0, MAN 26.8 ± 0.6, REAP 37.4 ± 2.4).

F Secondary benchmarks, exact values

On BFCL, items whose prompt exceeds the model's 40,960-token context are scored as failures. For the criteria we logged, this affects 40–54 of the 200 multi-turn long-context items, and the count rises as budgets shrink, itself a symptom of pruned models becoming more verbose.

The high-budget ordering in panel (c) of Figure 3 is not repeat-robust. Across three calibration samples the base form scores 39.2 ± 1.7 at 75% and 40.4 ± 1.5 at 50% on Gemma-4-26B MATH-500, and the deployed variant 38.1 ± 1.9 and 39.3 ± 1.4, overlapping ranges in both cases, so we read the $75/50$ cells as ties. The low-budget separation of the panel is unaffected.

Calibration data for HumanEval+. Every Qwen3-30B HumanEval+ row in this paper is selected on the same 102,400 tokens of the REAP mixture. Table 18 shows what the stream does to REAP: its 25% HumanEval+ moves from 17.7 under the full released mixture (2.1 million tokens in our count) to 45.7 on the 10^5-token sample of the same mixture and to 36.0 on the evol-codealpaca subset alone, a 28-point difference due to the calibration set alone, and its 37.5% cell moves between 57.9 and 82.9. The choice of calibration set therefore also affects saliency criteria.

G Exact values for Figure 2

Tables 20, 21 and 22 list the GSM8K values plotted in Figure 2 for Gemma-4-26B, Qwen3-30B and Qwen3.6-35B.

Table 23 lists HumanEval+ at all seven budgets on calibration sample 1.

H Proofs of the two lemmas

Quality of greedy selection. On Qwen3-30B at the 25% budget we compared greedy plus swap refinement against the best of ten random-restart swap local searches (best-improvement swaps to convergence) in eight evenly spaced layers. The two reach the same objective value to within 0.2% in every layer: greedy plus swap is at least as good in five layers and below the best restart by 0.2% in the other three (including layers 0 and 47), and swap refinement never changes the greedy solution's value by more than rounding. Within this search family greedy selection with swap refinement is therefore close to the best solution found; this is empirical evidence, not a guarantee.

Coupling ratio across layers. The coupling ratio also varies across layers of a single model. On Qwen3-30B, layers with a higher ratio show a larger reconstruction gain of the coupled objective

Qwen3-30B, MBPP+

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				74.9			
MAN	73.8	71.7	71.7	68.2	52.6	40.0	0.0
MoNE	73.5	74.3	70.9	**70.6**	61.4	17.2	0.3
SE	**75.4**	72.0	70.6	65.9	45.0	5.3	0.0
REAP	72.8	73.3	70.9	69.0	59.5	35.5	0.0
RECOUP (ours)	74.3	74.1	73.5	67.5	**66.1**	54.5	**9.3**
$\text{RECOUP}_{\text{deploy}}$ (ours)	75.1	**74.9**	**74.3**	69.6	65.3	**55.3**	8.5

Qwen3-30B, MATH-500

	Experts kept (%)		
Criterion	75	50	25
Original		79.2	
MAN	79.2	74.6	42.4
MoNE	80.4	77.6	62.6
SE	**80.6**	**79.0**	48.4
REAP	79.2	75.2	54.0
RECOUP (ours)	80.0	77.2	72.4
$\text{RECOUP}_{\text{deploy}}$ (ours)	80.0	76.0	**72.8**

Gemma-4-26B, MATH-500

	Experts kept (%)		
Criterion	75	50	25
Original		34.2	
MAN	29.8	14.0	3.2
MoNE	39.0	36.0	6.6
SE	38.4	23.6	4.0
REAP	**41.8**	33.6	8.8
RECOUP (ours)	37.0	**42.2**	**36.2**
$\text{RECOUP}_{\text{deploy}}$ (ours)	36.0	38.0	31.8

Table 16: Values of Figure 3, sample 1.

over the diagonal (REAP-type) selection (Spearman 0.46 over 48 layers), so the relation also holds within one architecture.

Token weighting. Write the objective with token weights $w_x \geq 0$: $J_w(S) = \min_\gamma \sum_x w_x \|y(x) - \sum_{i \in S \cap \text{top}k(x)} \gamma_i g_i(x) f_i(x)\|^2$. The normal equations are unchanged except that every accumulator carries w_x, so $G_w[i,j] = \sum_x w_x g_i g_j \langle f_i, f_j \rangle$ and $h_w = G_w \mathbf{1}$, giving $J_w(S) = c_w - h_{w,S}^\top G_{w,S}^{-1} h_{w,S}$ with the same algebraic form. Choosing $w_x = 1$ weights tokens as they occur at deployment (the sum weighting). Choosing w_x inversely proportional to the number of tokens routed to the expert weights experts uniformly (the mean measure). Selection is invariant to rescaling any expert's output, since γ absorbs the scale, so w is the only remaining degree of freedom. The sum/mean question in the literature (Lasby et al., 2026; Liu et al., 2026) is a choice of deployment weighting, not of aggregation rule.

Orthogonal case. If expert outputs are pairwise orthogonal on the calibration set, $\langle f_i, f_j \rangle = 0$ for $i \neq j$, so G is diagonal and $V(S) = h_S^\top G_S^{-1} h_S = \sum_{e \in S} G_{ee}$. A modular objective is maximized exactly by taking the M largest terms, so greedy is optimal and the selection reduces to ranking experts by $G_{ee} = \sum_x g_e^2 \|f_e\|^2$. Under expert-uniform weighting this ranking coincides with REAP's gate-weighted activation norm up to the moment order, so RECOUP contains REAP as a special case and differs from it only when the Gram matrix has off-diagonal mass.

LiveCodeBench v6

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				44.5			
MAN	42.9	41.1	41.1	38.5	11.7	3.5	0.0
MoNE	43.9	**44.1**	41.8	41.0	32.4	1.5	0.0
SE	43.5	42.7	41.1	34.6	12.5	0.4	0.0
REAP	44.0	43.7	42.5	40.5	28.9	1.8	0.0
RECOUP (ours)	44.8	43.8	**42.8**	40.0	36.2	**4.7**	0.0
$\text{RECOUP}_{\text{deploy}}$ (ours)	**45.2**	43.6	42.0	**41.6**	**38.1**	3.4	0.0

BFCL v4

	Experts kept (%)		
Criterion	75	50	25
MAN	**36.5**	30.8	13.0
MoNE	36.2	**34.0**	16.1
SE	34.9	30.4	12.1
REAP	35.7	32.8	14.7
RECOUP (ours)	35.2	32.1	15.3
$\text{RECOUP}_{\text{deploy}}$ (ours)	35.8	32.8	**17.9**

Table 17: LiveCodeBench v6 pass@1 and BFCL v4 accuracy on Qwen3-30B (REAP code mixture, sample 1). BFCL was not run on the original model.

			Experts kept (%)						
Criterion	Calibration data	Size	87.5	75	62.5	50	37.5	25	12.5
REAP	REAP mixture	2.1M	86.0	89.0	89.0	87.8	82.9	17.7	3.7
REAP	REAP mixture	0.1M	88.4	**88.4**	**87.2**	**87.8**	79.9	45.7	0.0
REAP	evol-codealpaca	0.1M	88.4	86.0	88.4	86.6	57.9	36.0	0.0
RECOUP	REAP mixture	0.1M	**89.6**	86.0	86.0	83.5	**82.3**	48.8	1.8
$\text{RECOUP}_{\text{deploy}}$	REAP mixture	0.1M	88.4	86.6	85.4	86.0	81.1	**55.5**	**4.9**

Table 18: HumanEval+ pass@1 on Qwen3-30B with different calibration data for REAP. 2.1M: the released REAP set. 0.1M REAP mixture: sample 1, used in the rest of the paper. Bold among the 0.1M REAP-mixture rows.

I $\text{RECOUP}_{\text{DEPLOY}}$ POOLS AND SWAP REFINEMENT

Evaluating the pruned model's computation requires statistics under masked routing, which depend on the kept set. We therefore use a candidate pool. For budget $M \leq E/2$ we take the pool of the $P = \min(2M, \frac{3}{4}E)$ experts ranked highest by REAP's saliency, and all E experts for larger M; collect re-selected, re-normalized routing statistics once per pool, and run greedy selection followed by swap refinement inside it. Pools are shared across budgets that map to the same P, so seven budgets need four passes. The pool is a restriction: an expert outside it cannot be selected even if re-selection would favor it. Increasing P increases compute proportionally, and in the one setting we tested it lowers the score. On Qwen3.6-35B at 12.5% HumanEval+, widening the pool from the default P=64 to 128, 192 and 256 (the full set) lowers $\text{RECOUP}_{\text{deploy}}$ from 60.4 to 45.7, 37.8 and 20.7, the last below REAP's 37.8. Statistics collected under a wide pool's re-selected routing no longer describe the routing the final M experts will run, so greedy selection drifts toward experts that look good only under the pool's routing. The saliency-ranked pool keeps the estimate close to the routing of the final kept set. We did not tune P per model.

Criterion	Sample 1	Sample 2	Sample 3	Mean ± half-range
Qwen3-30B, 37.5% kept				
MAN	43.9	70.7	70.1	61.6 ± 13.4
MoNE	81.7	76.2	72.6	76.8 ± 4.6
SE	47.6	68.9	57.9	58.1 ± 10.7
REAP	79.9	78.7	72.0	76.8 ± 4.0
RECOUP (ours)	82.3	52.4	45.7	60.2 ± 18.3
$\text{RECOUP}_{\text{deploy}}$ (ours)	81.1	65.8	48.2	65.0 ± 16.5
Qwen3-30B, 25% kept				
MAN	18.9	0.0	3.0	7.3 ± 9.4
MoNE	17.7	15.8	7.9	13.8 ± 4.9
SE	0.6	20.1	7.9	9.6 ± 9.8
REAP	45.7	18.3	21.3	28.5 ± 13.7
RECOUP (ours)	48.8	12.8	20.1	27.2 ± 18.0
$\text{RECOUP}_{\text{deploy}}$ (ours)	55.5	27.4	32.9	38.6 ± 14.0
Qwen3-30B, 12.5% kept				
MAN	0.0	0.6	0.0	0.2 ± 0.3
MoNE	0.0	0.6	0.0	0.2 ± 0.3
SE	0.0	0.0	0.0	0.0 ± 0.0
REAP	0.0	0.0	0.0	0.0 ± 0.0
RECOUP (ours)	1.8	4.3	7.3	4.5 ± 2.7
$\text{RECOUP}_{\text{deploy}}$ (ours)	4.9	4.3	0.0	3.0 ± 2.4
Qwen3.6-35B, 12.5% kept				
MAN	26.8	26.2	27.4	26.8 ± 0.6
MoNE	42.7	28.7	40.9	37.4 ± 7.0
SE	5.5	16.5	12.8	11.6 ± 5.5
REAP	37.8	34.8	39.6	37.4 ± 2.4
$\text{RECOUP}_{\text{deploy}}$ (ours)	60.4	53.0	57.3	56.9 ± 3.7

Table 19: HumanEval+ pass@1 per calibration sample (REAP code mixture). Each sample has nine to eleven documents, 65–80% of the tokens from SWE-smith. RECOUP on Qwen3.6-35B: sample 1 only.

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				72.3			
MAN	71.7	69.0	58.1	38.6	15.6	4.5	1.3
MoNE	71.9	71.5	69.5	63.1	52.8	19.2	1.8
SE	71.8	**73.4**	69.6	61.9	45.5	12.0	1.1
REAP	72.4	72.0	66.3	62.2	48.0	17.8	1.3
RECOUP (ours)	72.7	72.3	70.2	69.8	**70.7**	**57.7**	**24.6**
$\text{RECOUP}_{\text{deploy}}$ (ours)	**73.2**	72.3	**71.2**	**72.0**	69.5	46.9	6.4

Table 20: GSM8K accuracy on Gemma-4-26B (Figure 2a; math calibration set, sample 1).

J Layer-local statistics of expert $(L9, E2)$

For the selections of Appendix M, which differ in expert $(L9, E2)$ (expert 2 of layer 9), we asked whether any layer-local statistic predicts which of two nearly-identical selections scores higher. Across selections on Qwen3.6-35B at 25%, Spearman correlations between task score and layer-residual moments were $|\rho| \leq 0.3$ and unstable in sign: $\mathbb{E}\|r\|$ -0.14, $\mathbb{E}\|r\|^2$ $+0.29$, $\mathbb{E}\|r\|^4$ $+0.11$. The expert whose inclusion moved GSM8K by 19.2 points had a 0.31% routing share, median-range saliency, and maximal output correlation 0.04 with any other expert. None of these layer-local statistics distinguishes it from other experts.

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				91.1			
MAN	91.0	89.2	83.5	82.9	76.3	40.4	1.5
MoNE	91.0	**91.7**	86.3	84.9	**83.9**	71.1	0.7
SE	**91.5**	88.1	89.7	**88.9**	82.7	56.7	3.9
REAP	91.0	89.5	88.3	86.5	81.8	65.6	11.6
RECOUP (ours)	90.8	89.9	89.4	83.5	79.4	78.8	**37.0**
$\text{RECOUP}_{\text{deploy}}$ (ours)	90.1	89.9	**90.1**	84.6	81.0	**80.1**	36.1

Table 21: GSM8K accuracy on Qwen3-30B (Figure 2b; math calibration set, sample 1).

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Original				96.2			
MAN	95.9	95.5	**95.8**	95.6	92.0	78.8	13.3
MoNE	96.1	95.4	**95.8**	95.7	94.0	89.8	30.6
SE	95.8	95.4	95.5	94.5	94.1	39.6	33.4
REAP	**96.2**	**95.9**	**95.8**	**95.9**	94.3	89.9	34.8
RECOUP (ours)	96.0	95.2	**95.8**	95.2	95.0	92.4	**45.9**
$\text{RECOUP}_{\text{deploy}}$ (ours)	95.9	95.7	95.7	95.1	**95.5**	**92.6**	36.2

Table 22: GSM8K accuracy on Qwen3.6-35B (Figure 2c; math calibration set, sample 1, reasoning disabled).

K THE MULTIPLE-CHOICE EVALUATION

Table 24 reports the eight-task multiple-choice average (ARC-challenge and ARC-easy, Clark et al., 2018; BoolQ, Clark et al., 2019; HellaSwag, Zellers et al., 2019; MMLU, Hendrycks et al., 2021a; OpenBookQA, Mihaylov et al., 2018; RTE, Wang et al., 2019; WinoGrande, Sakaguchi et al., 2020) under one calibration for every row (the REAP calibration mixture, 10^5 tokens) and router-mask evaluation. The ordering depends on the budget: from 87.5% to 37.5% REAP leads by 2 to 5 points on Qwen3-30B (except at 50%, where the deployed form is 0.2 ahead) and by at most 2.2 points on Gemma-4-26B (tied at 50 and 37.5%), while at 12.5% both RECOUP variants exceed it on both models and at 25% the deployed variant ties it on Qwen3-30B (REAP keeps a one-point lead on Gemma-4-26B), the same low-budget regime in which the generative gains of Section 5 appear.

L OLMOE MULTIPLE-CHOICE REPEATS

Three-window means on multiple-choice tasks for OLMoE-1B (Section 5.4). At 62.5% and above the criteria are statistically indistinguishable, with spreads inside the sample half-ranges. From 50% down the base form sits 0.9 to 1.7 points above REAP in every sample mean, with ranges that touch rather than separate, and $\text{RECOUP}_{\text{deploy}}$ drops slightly below REAP at the two smallest budgets. OLMoE-1B's coupling ratio (0.163) is unexceptional. With 64 experts, each with a large share of the computation, re-selection has few alternatives, and we interpret the result as showing no dependence on the coupling ratio rather than a reversal.

M REASONING-MODE ENTRY ON QWEN3.6-35B

The comparison of Section 5.1 is measured on GSM8K (8-shot chain of thought, the few-shot completion prompt without a chat template, greedy decoding, $1{,}024$ new tokens) on Qwen3.6-35B at 25% of experts kept ($M{=}64$ of 256). Qwen3.6-35B is a hybrid reasoning model, and under this prompt the original model opens a `<think>` block spontaneously in 8% of responses. Many pruned selections do so far more often, and a response that enters the reasoning block is usually cut at the

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Gemma-4-26B (E=128, *ratio* 0.354)							
Original				57.9			
MAN	52.4	56.1	39.6	15.2	0.6	0.0	0.0
MoNE	**63.4**	59.8	50.6	29.9	7.3	0.0	0.0
SE	48.8	47.0	36.6	18.3	3.0	0.0	0.0
REAP	56.7	**61.0**	47.6	29.9	4.3	0.0	0.0
RECOUP (ours)	59.8	59.8	**55.5**	**57.3**	**48.8**	**25.0**	0.0
$\text{RECOUP}_{\text{deploy}}$ (ours)	61.6	56.7	54.9	54.9	43.3	10.4	0.0
Qwen3-30B (E=128, *ratio* 0.142)							
Original				87.8			
MAN	87.2	87.2	**87.2**	84.8	43.9	18.9	0.0
MoNE	87.2	87.8	**87.2**	85.4	81.7	17.7	0.0
SE	**90.8**	**90.8**	**87.2**	76.8	47.6	0.6	0.0
REAP	88.4	88.4	**87.2**	**87.8**	79.9	45.7	0.0
RECOUP (ours)	89.6	86.0	86.0	83.5	**82.3**	48.8	1.8
$\text{RECOUP}_{\text{deploy}}$ (ours)	88.4	86.6	85.4	86.0	81.1	**55.5**	**4.9**
Qwen3.6-35B (E=256, *ratio* 0.085)							
Original				90.2			
MAN	89.6	**92.1**	**91.5**	89.6	87.8	79.9	26.8
MoNE	88.4	90.2	90.8	90.2	90.2	82.3	42.7
SE	**91.5**	89.6	89.0	89.0	86.0	74.4	5.5
REAP	89.0	90.2	**91.5**	**92.7**	**90.8**	82.3	37.8
RECOUP (ours)	90.2	90.8	90.8	**92.7**	87.8	**88.4**	17.1
$\text{RECOUP}_{\text{deploy}}$ (ours)	90.2	90.8	89.0	92.1	86.0	**88.4**	**60.4**

Table 23: HumanEval+ pass@1 (REAP code mixture, sample 1). Three-sample means are in Table 4.

	Experts kept (%)						
Criterion	87.5	75	62.5	50	37.5	25	12.5
Qwen3-30B							
Original				68.2			
REAP	**65.7**	**62.5**	**55.4**	47.7	**43.0**	**36.5**	34.4
RECOUP (ours)	63.5	57.0	52.9	46.7	40.2	35.2	**34.9**
$\text{RECOUP}_{\text{deploy}}$ (ours)	63.7	57.5	53.0	**47.9**	40.5	**36.5**	34.7
Gemma-4-26B							
Original				69.2			
REAP	**66.4**	**62.0**	**55.3**	**48.8**	42.8	**39.3**	35.1
RECOUP (ours)	65.2	59.4	54.1	48.0	**42.9**	38.3	**36.0**
$\text{RECOUP}_{\text{deploy}}$ (ours)	65.4	59.8	54.7	**48.8**	42.8	38.4	35.8

Table 24: MC-8 accuracy (%) with the REAP code mixture (sample 1), computed by masking the router.

token limit or loses the answer format. Table 25 splits every response of the nineteen selections that carry GSM8K samples by whether it entered the reasoning block. Selections that contain expert $(L9, E2)$ enter it in 25 to 54% of responses and score 70 to 83; selections without it enter in 63 to 94% and score 55 to 64 (the mean-weighting endpoint, at 63% entry, scores 73.1); and the accuracy of the responses that did *not* enter the block is 90 to 98 for every selection. The expert therefore does not affect arithmetic ability. It determines whether the pruned model stays in the mode requested by the prompt, and the difference in score is due to leaving that mode. With the chat template

applied and reasoning explicitly disabled (last column) all nineteen selections score 92.0 to 93.2. Qwen3-30B and Gemma-4-26B never enter a reasoning block under the completion prompt, so every Qwen3.6-35B GSM8K and MATH-500 cell in this paper is measured with reasoning disabled while the other models keep the completion prompt.

	Has	<think>	GSM8K accuracy		
Selection	$(L9, E2)$	(%)	All	No <think>	Reasoning off
Endpoint 1	no	75	63.9	92.3	92.0
Endpoint 2	yes	25	83.2	92.4	92.5
Step 1	no	81	55.2	95.2	92.4
Step 2	no	82	55.3	94.4	92.3
Step 3	yes	41	77.6	92.0	93.0
Step 4	no	75	58.4	94.2	92.5
Step 5	yes	34	81.9	93.5	92.5
Step 6	yes	45	77.5	92.5	92.6
Step 7	no	94	55.8	97.2	92.7
Step 8	yes	36	80.1	93.2	93.2
Step 9	no	84	59.2	93.7	92.3
Step 10	no	82	55.1	96.4	92.3
Step 11	yes	45	78.1	92.2	92.7
Probe 1	no	86	55.3	98.4	92.7
Probe 2	no	82	57.2	95.7	92.3
Probe 3	no	82	59.6	94.9	92.2
Probe 4	no	90	55.6	95.3	93.0
$\text{RECOUP}_{\text{deploy}}$, sum	yes	54	69.6	92.1	92.6
$\text{RECOUP}_{\text{deploy}}$, mean	no	63	73.1	90.3	92.6

Table 25: Qwen3.6-35B at 25%: 19 selections that differ in at most 0.66% of the experts. Endpoints: the two most different selections; steps: bisection between them; probes: one expert added or removed; last two rows: $\text{RECOUP}_{\text{deploy}}$ with sum and mean weighting. <think>: share of responses that open a reasoning block under the completion prompt.

N PROTOCOL DETAILS

Model	Checkpoint	Experts	Top-k	Shared path	Params (total / active)
OLMoE-1B	OLMoE-1B-7B-0924	64	8	none	6.9B / 1.3B
Qwen3-30B	Qwen3-30B-A3B	128	8	none	30.5B / 3.3B
Gemma-4-26B	Gemma-4-26B-A4B	128	8	dense MLP	26.0B / 4.0B
Qwen3.6-35B	Qwen3.6-35B-A3B	256	8	shared expert	35.0B / 3.3B

Table 26: Models. Short names are used in the paper.

Models and calibration sets. Models are the public checkpoints `allenai/OLMoE-1B-7B-0924`, `Qwen/Qwen3-30B-A3B`, `google/gemma-4-26b-a4b` and `Qwen/Qwen3.6-35B-A3B`. Calibration sets are 2,048-token chunks in each model's tokenizer: math calibration set from `allenai/tulu-3-sft-personas-math` (Lambert et al., 2024) and REAP code mixture from the calibration mixture released with Lasby et al. (2026) (evol-codealpaca, `theblackcat102/evol-codealpaca-v1`, Luo et al., 2024; MoT code, math and science; xLAM; SWE-smith; 4,096 documents each, shuffled with seed 7 and packed), 50 chunks (102,400 tokens) per sample, with samples at chunks 0–49, 100–149 and 200–249.

Selection and baselines. Gram statistics use sum weighting; $\text{RECOUP}_{\text{deploy}}$ collects them under masked routing for the pool of Appendix I; greedy selection with swap refinement is deterministic given the calibration set. SE, MoNE and MAN are re-implemented from their papers (activation Eq. 6 for SE; Eq. 4–7 of MoNE; the gate-free instance of Liu et al., 2026 for MAN) and REAP is re-derived, all on identical tokens.

Generation and multiple choice. Pruned checkpoints are exported and evaluated with GSM8K 8-shot strict match and $1{,}024$ new tokens; EvalPlus greedy pass@1 with the chat template and thinking disabled (768 new tokens for HumanEval+, $2{,}048$ for MBPP+); MATH-500 Minerva 4-shot (Lewkowycz et al., 2022) with `math_verify`; LiveCodeBench v6 with the official $n{=}10$, temperature 0.2, top-p 0.95, $2{,}000$-token setting in each model family's chat format; and BFCL v4 (non-live, live, multi-turn) served with vLLM (Kwon et al., 2023) and a $40{,}960$-token context. The multiple-choice evaluation (Appendix K) uses `lm-eval` (Gao et al., 2023) with router masking on the unexported model. Quoted gaps are tested by two-sided paired permutation over per-problem outcomes. Gemma-4-26B ships no chat template in its repository or in the library, so we apply its turn format from a template file (`<|turn>`$\cdots$`<turn|>`, thinking disabled with an empty thought block), stop generation at the end-of-turn token or at a repeated empty code fence, and keep the standard EvalPlus extractor without repair. On LiveCodeBench, where the model otherwise repeats code blocks until the token limit, generation stops at the closing fence of the first block (kept in the output) and the standard extractor is applied unchanged.

Contamination. A 13-gram containment check finds no overlap between the calibration sets actually used and any GSM8K, MATH-500, HumanEval+ or MBPP+ item.

Evaluation noise. Repeating the HumanEval+ evaluation of one exported checkpoint three times (vLLM, greedy, bf16) at 25% on Qwen3-30B returned $48.8/47.0/47.6$ for RECOUP, $55.5/51.2/53.1$ for $\text{RECOUP}_{\text{deploy}}$ and $45.7/45.1/45.1$ for REAP; batched bf16 greedy decoding is not bit-reproducible, and we read HumanEval+ differences inside this spread as noise. GSM8K under the HF backend reproduced to the digit across three runs of each of two checkpoints.

Compute and artifacts. All runs use single A100 80GB GPUs. Kept sets, raw result files and the scripts that regenerate every table and figure ship with the submission.