# NO EXPERT REPLACES ANOTHER: PRUNED EXPERTS' OUTPUTS ARE NOT RECOVERED BY THE REMAINING ONES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Expert pruning, merging and replacement in sparse Mixture-of-Experts (MoE) models are commonly motivated by expert redundancy, inferred from output similarity on shared inputs, from the small loss change after zeroing a single expert, or from replacing an expert by its mean output. None of these measurements tests whether the remaining experts can reproduce a pruned expert's output on the tokens routed to it, which is where pruning changes the layer output. We measure this on four MoE models. The remaining experts, combined linearly, recover only 2 to $28\%$ of the pruned output energy, whereas a noisy copy of the pruned expert recovers $98\%$. Pruning still works at moderate budgets because this energy is concentrated in few experts: keeping half of them removes only $4.5$ to $18\%$ of it, and within a budget the energy a criterion removes predicts its perplexity. Three consequences follow. Pruning scores should be summed over an expert's routed tokens rather than averaged, which lowers perplexity in $15$ of $18$ settings on the three larger models. Modeling interactions between experts changes the selection little. Calibration data should match the deployment domain rather than be large: 2K matched tokens outperform two million mismatched ones.

## 1 INTRODUCTION

Sparse Mixture-of-Experts (MoE) models (Shazeer et al., 2017; Fedus et al., 2022) route each token to a small subset of experts, so most parameters are inactive for any given token. Compression methods reduce this cost by deleting experts (Lu et al., 2024; Lasby et al., 2026), merging them (Li et al., 2024; Chen et al., 2025) or replacing them with cheaper modules (Zhang et al., 2026). These methods are usually motivated by redundancy among experts. Prior work reports that experts in a layer have similar representations (Zhang et al., 2025), that a token's secondary expert can be replaced by a similar primary expert at small loss (Wu et al., 2026), that an expert can be replaced by its mean output (Zhang et al., 2026), and that zeroing a single expert in an early layer changes the loss very little (Engmann et al., 2026). The last result has been interpreted as evidence that pruning succeeds because experts are redundant.

None of these measurements evaluates the remaining experts on the tokens that were routed to the pruned expert. When expert $j$ is removed, the layer output changes only on the tokens routed to $j$, by $j$'s gated output. Whether the removal is cheap therefore depends on whether the remaining experts can reproduce that output on those tokens. Similarity on a shared batch does not test this, because the router assigns each token to the experts with the largest response, so most tokens in a shared batch are not routed to either expert being compared. The small loss change after zeroing a single expert does not test it either, because a small perturbation of one layer can be absorbed downstream even if no other expert reproduces the removed output.

We measure this directly (Section 3). For each expert removed by a pruning criterion, we compute the fraction of its deletion energy, the squared error its removal causes, that the remaining experts can reconstruct on the tokens routed to it. We consider the best single remaining expert with an optimal scale, a joint linear fit of all remaining experts with an intercept, and a constant vector. Coefficients are fitted on one half of the calibration set and evaluated on the other. Figure 1a shows the result for four public models with $64$ to $256$ experts per layer. A joint fit of all remaining experts



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.

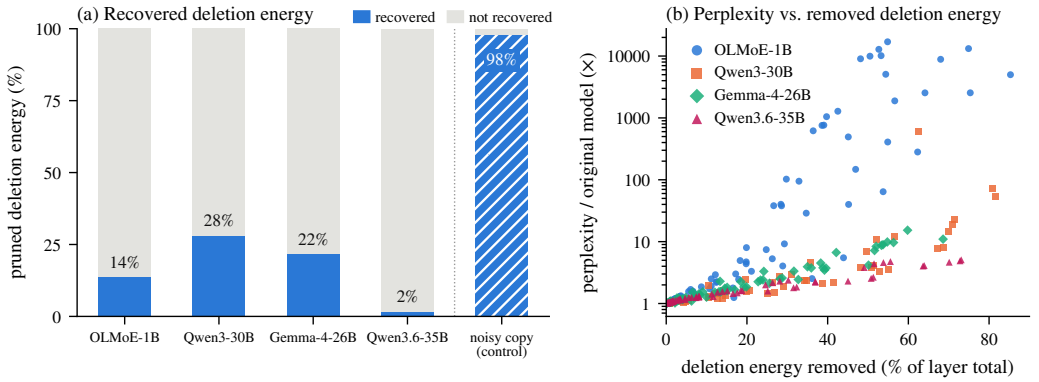


Figure 1: (a) Total deletion energy (Eq. 1) of the experts pruned at a 25% budget, split into the part that a linear fit of all remaining experts plus a constant recovers on the pruned experts' routed tokens, and the part it does not recover. Fits use one half of the C4 calibration set and are evaluated on the other half. The hatched bar is a control: a copy of the pruned expert with 10% relative weight noise (median over OLMoE-1B and Qwen3-30B). (b) WikiText perplexity relative to the original model against the deletion energy removed by the kept set (mean over layers), for nine criteria at seven budgets with C4 calibration. Spearman correlations per model: 0.94 (OLMoE-1B), 0.97 (Qwen3-30B), 0.99 (Gemma-4-26B) and 0.99 (Qwen3.6-35B). Figure 4 shows each criterion separately.

recovers 2 to 28% of the deletion energy removed by pruning, and the remaining 72 to 98% is not recovered. A noisy copy of the pruned expert, used as a control, recovers 98%. The best single remaining expert recovers less than the joint fit (Table 2).

Pruning nevertheless works at moderate budgets, because deletion energy is concentrated in a minority of experts. Figure 1b plots, for every kept set we evaluated, the perplexity against the fraction of deletion energy removed. Keeping the half of the experts with the largest deletion energy removes 4.5 to 18% of it, and on the three larger models the perplexity increase at this budget is small. We call the rule that keeps the top-$M$ experts by deletion energy maximum energy preservation (MEP). Within a fixed budget, the energy removed by a criterion predicts its perplexity (Section 4), which suggests that criteria differ mainly in how much energy they remove. This leads to three consequences, which Sections 4 to 6 test.

- **Pruning scores should be summed over an expert's routed tokens.** REAP, MAN and MSAN average the score over routed tokens, which is appropriate only if the tokens of frequently routed experts can be reconstructed by the remaining experts. Summing instead lowers the perplexity of REAP and MAN in 15 of 18 settings on the three models with 128 or more experts, by up to 4.3× (Tables 3 and 17).

- **Modeling interactions between experts changes the selection little.** Selection with the full matrix of pairwise interactions keeps 81 to 100% of the experts kept by MEP (Table 21).

- **Calibration data should match the evaluation domain; its size matters less.** For budgets down to 25%, 2,048 tokens of web text matched to the WikiText evaluation give perplexity within a few points of $10^5$ tokens, and at 25% they give 2.8 to 5.4 times lower perplexity than two million tokens of a mismatched mixture (Figure 2 and Table 6).

MEP is not a new criterion. It is the $b$=0, $\alpha$=$\beta$=2 member of the criterion family of Liu et al. (2026b), who recommend averaging ($b$=1) for task-agnostic pruning. Our contributions are the recoverability measurement, which explains when summation is preferable, the test of interaction-aware selection, and the calibration result. All results concern one-shot pruning without retraining. Merging with retraining and recovery training modify the remaining experts and are not covered by our measurement.

| | Reconstruction by other experts | Several experts removed |
|---|---|---|
| **Experts have similar outputs** (Zhang et al., 2025)<br>Evidence: CKA above 0.7 on a shared batch (Mixtral-8x7B).<br>→ *The most similar partner has CKA 0.27–0.55, but the best partner recovers only 0.13–7.1% of the deletion energy on the routed tokens of the pruned expert.* | × | × |
| **A single expert can be removed at little loss** (Engmann et al., 2026)<br>Evidence: small loss change; no observational routing statistic predicts it.<br>→ *The deletion energy predicts it (Spearman 0.57 vs. 0.33–0.42), and removing half a layer costs 1.36× the sum of single removals.* | × | × |
| **An expert can be replaced by its mean output** (Zhang et al., 2026)<br>Evidence: light replacement modules keep accuracy.<br>→ *A constant vector recovers 3–24% of the deletion energy on the routed tokens of the pruned expert.* | × | ✓ |
| **Similar experts can be merged** (Li et al., 2024; Chen et al., 2025)<br>Evidence: parameters shared within groups of similar experts.<br>→ *All remaining experts jointly recover 2–28% on the routed tokens of the pruned expert.* | × | ✓ |
| **No expert replaces another** (ours)<br>Evidence: the fraction of each pruned expert's deletion energy that remaining experts recover on its routed tokens (Tables 2 and 9). | ✓ | ✓ |

Table 1: Redundancy claims in prior work and our measurement on the tokens routed to the pruned expert (blue; values from Tables 2, 9, 12 and 14).

## 2 Redundancy Measurements in Prior Work

A MoE layer routes token $x$ to the top-$k$ of its $E$ experts and outputs $\sum_{e \in \text{top}k(x)} g_e(x) f_e(x)$, where $g_e(x)$ is the router gate and $f_e(x)$ the expert output. Let $X_j$ be the set of tokens routed to expert $j$, $n_j = |X_j|$, and $d_j(x) = g_j(x) f_j(x)$ the contribution of $j$ to the output on $x$. Removing $j$ subtracts $d_j(x)$ from the layer output for every $x \in X_j$ and leaves the output unchanged elsewhere. Table 1 compares four redundancy claims with the corresponding measurement on $X_j$.

**Output similarity.** Zhang et al. (2025) compute centered kernel alignment (CKA) between the outputs of two experts on a shared batch and merge experts with high CKA. CKA is invariant to rotation and isotropic scaling of the outputs, whereas the residual stream is not, so a high CKA does not imply that one expert can replace another. On our models, the most similar partner of an expert has a CKA of 0.27 to 0.55 (layer medians). On the tokens routed to the expert, however, its best partner among all experts of the layer recovers a median of only 0.13 to 7.1% of its deletion energy (layer medians over all experts, not only pruned ones), and CKA and the recovered fraction have a Spearman correlation of $+0.01$ to $+0.41$ across pairs (Table 12). Lo et al. (2025) compare expert weights after neuron alignment, and Chaudhari et al. (2026) show that the top-gated expert approximates the output of the full ensemble. Neither measures whether one expert can reproduce another's output.

**Single-expert ablation.** Engmann et al. (2026) zero one expert at a time, find that none of four observational routing statistics predicts the resulting loss change, and conclude that early-layer redundancy makes most selection criteria interchangeable. We use their checkpoint and statistics with a simpler design that zeroes each expert on all tokens (Appendix C). The deletion energy defined in Section 3, which was not among the statistics they tested, predicts the loss change better than all four of their statistics in 11 of 16 layers (Table 5). Removing the half of a layer with the smallest deletion energy costs a median of 0.54 times as much as removing a random half, and 1.36 times the sum of the corresponding single-expert losses, so the effects of removing several experts add up rather than

cancel. Wu et al. (2026) replace a token's secondary expert at small loss; the secondary slots have the smallest gates, whereas pruning removes an expert from all slots, including the primary one.

**Replacement and merging.** MoNE (Zhang et al., 2026) replaces an expert with a light module built around its mean output, and LightMoE (Hao et al., 2026) replaces experts with light modules whose capability is restored by fine-tuning. A constant vector is one of the reconstructions we evaluate. Merging (Li et al., 2024; Chen et al., 2025) shares parameters within groups of similar experts, and Lasby et al. (2026) argue that the resulting error grows with the functional distinctness of the experts. The joint fit in Section 3 quantifies this distinctness.

**Pruning criteria.** One-shot expert pruning scores each expert and keeps the top $M$ per layer. Existing scores include router traffic (SEER-MoE, Muzio et al., 2024), the mean or mean-squared activation norm (MAN and MSAN, Liu et al., 2026b), the gate-weighted output norm (REAP, Lasby et al., 2026; EASY-EP, Dong et al., 2025), routing frequency combined with output variance (MoNE), and protection of the experts with activation outliers that Su et al. (2026) identify as super experts (SE). Liu et al. (2026b) write the norm-based scores as $S_j(b, \alpha, \beta) = n_j^{-b} \sum_{x \in X_j} g_j(x)^\alpha \|f_j(x)\|^\beta$, where $b$=0 sums over routed tokens and $b$=1 averages. They choose averaging for task-agnostic pruning because it makes the kept set more stable across calibration corpora, note that task-specific pruning benefits from routing frequency, and include the $b$=0, $\alpha$=$\beta$=2 member in their task-specific comparison. REAP averages in order to retain rarely routed experts, and EASY-EP, a domain-specific method, sums a weighted gate-output norm over tokens. Neither choice is based on a measurement of whether the remaining experts can reconstruct the pruned expert's output. AIMER (Liu et al., 2026a) avoids calibration data by scoring experts from their weights.

## 3 Recoverability of Pruned Experts

**Deletion energy.** Removing expert $j$ causes the squared error

$$w_j = \sum_{x \in X_j} \|d_j(x)\|^2 = \sum_{x \in X_j} g_j(x)^2 \|f_j(x)\|^2 \tag{1}$$

summed over its routed tokens. Since $w_j$ is also the energy of $j$'s gated output on these tokens, we call it the deletion energy of $j$. Pruning keeps $M < E$ experts per layer, masks the others in the router and renormalizes the gates of the remaining experts. All evaluations below use this mechanism, and all statistics are computed per layer.

**Recovered fraction.** We evaluate a reconstruction only on $X_j$, weighting each token as in Eq. 1: for two output sequences $u, v$ let $\langle u, v \rangle_j = \sum_{x \in X_j} g_j(x)^2 \langle u(x), v(x) \rangle$. A single remaining expert $i$ with the optimal scale recovers the fraction

$$\rho_{ij}^2 = \frac{\langle f_j, f_i \rangle_j^2}{\|f_j\|_j^2 \, \|f_i\|_j^2} \tag{2}$$

of $w_j$, which is 0 if $f_i$ is orthogonal to $f_j$ on $X_j$ and 1 if $f_i$ is proportional to $f_j$. Since several remaining experts can share the tokens of a pruned expert, we also regress $f_j$ on the whole kept set $S$,

$$R_j^2(S) = 1 - \min_{\gamma} \frac{\|f_j - \sum_{i \in S} \gamma_i f_i\|_j^2}{\|f_j\|_j^2}, \tag{3}$$

which is the largest fraction that any fixed linear combination of the remaining experts can recover. With up to 64 regressors an in-sample fit overestimates this fraction for rarely routed experts, so $\gamma$ is fitted on one half of the calibration set and evaluated on the other, in both directions. We also fit the kept set with an additional intercept vector, and a constant vector alone, which corresponds to replacement by a mean output. Weighting the fraction of each pruned expert by $w_j$ gives the fraction of the total pruned deletion energy that is recovered.

**Controls.** As negative controls we use the experts of another layer, the experts of the same layer with shuffled token order, and a random direction. As positive controls we use a copy of the pruned expert with Gaussian weight noise, alone and added to the pool of remaining experts, and a synthetic target formed as a random linear combination of eight kept experts (Appendix B).

| Model | Recovered by remaining experts | | Not recovered |
|---|---|---|---|
| | best single expert | all, jointly | |
| OLMoE-1B | 4.3% | 13.6% | **86.4%** |
| Qwen3-30B | 17.2% | 27.9% | **72.1%** |
| Gemma-4-26B | 2.2% | 21.6% | **78.4%** |
| Qwen3.6-35B | 0.7% | 1.6% | **98.4%** |
| *Noisy copy of the pruned expert (control)* | *98.0%* | *98.0%* | *2.0%* |

Table 2: Share of the total deletion energy of the experts pruned at a 25% budget that remaining experts recover on the pruned experts' routed tokens, and the share that is not recovered (C4 calibration set, all layers; fits use one half of the calibration set and are evaluated on the other). "All, jointly": a linear fit of all remaining experts plus a constant. The control row replaces the remaining experts by a copy of the pruned expert with 10% relative weight noise (median over experts, OLMoE-1B and Qwen3-30B). Details in Table 9.

**Models, criteria and evaluation.** The four models differ in the number of experts and in the shared path: OLMoE-1B (64 experts, top-8), Qwen3-30B (128, top-8), Gemma-4-26B (128, top-8, with an always-active dense MLP) and Qwen3.6-35B (256, top-8, with a gated shared expert). We refer to the three models with 128 or more experts as fine-grained. Statistics are accumulated on $10^5$ tokens of English C4 web text (Raffel et al., 2020) per model in 2,048-token chunks, and the two halves used for cross-fitting contain 25 chunks each. Each calibration set is used as three disjoint samples of $10^5$ tokens (chunks 0–49, 100–149 and 200–249); results from a single sample use the first. The budget is the percentage of routed experts kept in every layer. The criteria are MAN, MSAN, MoNE, SE, REAP, the traffic criterion of the first stage of SEER-MoE, the summed variants of MAN and REAP, and MEP. All are computed from the same tokens, at budgets from 87.5% to 12.5% of the experts, uniformly across layers. We report WikiText-2 perplexity (Merity et al., 2017) and the average accuracy on eight multiple-choice tasks (MC-8) with lm-eval-harness (Gao et al., 2023), and GSM8K (Cobbe et al., 2021) and HumanEval+ (Chen et al., 2021; Liu et al., 2023) with task-matched calibration. Appendix A gives the remaining details.

**Results.** Table 2 summarizes the measurement: 72 to 98% of the deletion energy removed by pruning is not recovered by any fixed combination of the remaining experts, whereas a noisy copy of the pruned expert leaves only 2%. Table 9 gives the details. The best single remaining expert recovers a median of 0.02 to 2.1% of a pruned expert's deletion energy and 0.7 to 17% of the total pruned deletion energy. A joint fit of the kept set recovers 1.4 to 23% without and 1.6 to 28% with an intercept. A constant vector alone recovers as much as the linear fit or more on every model, five times more on Gemma-4 (20.6 versus 4.0%). The recoverable part is therefore mostly a constant offset rather than a token-dependent function. On Qwen3.6 the fit with an intercept is below the constant alone (1.6 versus 3.4%) because the additional coefficients overfit the fitting half. The larger fraction on Qwen3-30B comes largely from its last layer, where the linear fit recovers 55% and a constant alone 51% (Appendix B). Expressed as a correlation, the largest linear fraction, 23%, corresponds to 0.48 between the pruned expert's output and its best reconstruction.

The controls bound these values from both sides. A copy of the expert with 10% weight noise recovers 98%; on OLMoE-1B, adding it to the pool of all other experts raises the median joint $R^2$ in the two layers tested from 1.1–1.4% to 98% (Table 11), and the synthetic eight-expert target is recovered with a median $R^2$ of 100%. The estimator therefore detects a single or distributed substitute when one exists. On Qwen3-30B and Qwen3.6, the best single expert of the same layer recovers 5 to 14 times more than with shuffled tokens, although both values are small (0.07 to 0.36% versus 0.01 to 0.04%). On Gemma-4, the same-layer value of up to 6.2% is almost matched by the shuffled-token control (5.6%), which again indicates a shared offset rather than a shared function (Table 10).

**Concentration of expert output on routed tokens.** For each expert $j$, we compute the fraction of its total output energy that falls on $X_j$ and divide it by the mean of the same fraction for the other experts, which removes differences in overall output scale. This ratio exceeds one for 89 to 97% of the experts, with a median of about 2 (Table 13). Each expert produces about twice its average

| | | Deletion energy kept (%) | | | | WikiText perplexity | | | |
|---|---|---|---|---|---|---|---|---|---|
| Criterion | Score | OLMoE-1B | Qwen3-30B | Gemma-4-26B | Qwen3.6-35B | OLMoE-1B | Qwen3-30B | Gemma-4-26B | Qwen3.6-35B |
| Original | | 100.0 | 100.0 | 100.0 | 100.0 | 8.9 | 10.9 | 8.7 | 8.1 |
| MAN | mean $\|f\|$ | 43.4 | 31.3 | 60.4 | 48.7 | 16770 | 86.7 | 32.5 | 21.7 |
| MSAN | mean $\|f\|^2$ | 57.5 | 32.8 | 61.2 | 49.2 | 11398 | 84.3 | 32.9 | 20.8 |
| MoNE | frequency, var. | 60.9 | 48.6 | 64.1 | 63.1 | 6791 | 44.3 | 32.8 | 18.0 |
| SE | activation outliers | 54.9 | 52.0 | 60.8 | 55.0 | 4383 | 42.2 | 30.2 | 18.8 |
| REAP | mean $g\|f\|$ | 61.4 | 49.2 | 65.0 | 63.1 | 6724 | 42.0 | 34.2 | 18.5 |
| Traffic | token count | 24.7 | 64.2 | 49.9 | 69.8 | 22632 | 51.0 | 36.3 | 19.4 |
| MAN-sum | sum $\|f\|$ | 35.9 | 68.9 | 61.7 | 71.9 | 22601 | 31.6 | 39.3 | 18.7 |
| REAP-sum | sum $g\|f\|$ | 60.3 | 71.9 | 68.4 | 73.6 | 9312 | 29.2 | 29.1 | 17.7 |
| MEP | sum $g^2\|f\|^2$ | 63.6 | **73.8** | **75.8** | **75.4** | 5503 | **23.9** | **28.8** | **16.3** |

Table 3: Pruning criteria at a 25% budget: the score each criterion assigns to an expert ($g$ is the gate and $\|f\|$ the output norm on a routed token), the deletion energy kept by its kept set (mean over layers, C4 calibration set), and WikiText perplexity. Bold: most energy kept and lowest perplexity per model. MEP keeps the most energy by construction; the other criteria's perplexities largely follow the energy they keep. All criteria have collapsed on OLMoE-1B at this budget (grey). Table 17 gives 50%, 12.5% and MC-8 (MoNE and SE not included).

output on the tokens routed to it, while the other experts do not. A similarity computed over a shared batch averages over this structure.

## 4 Pruning Loss and Pruned Deletion Energy

If the remaining experts recover little, the loss of a kept set should depend mainly on the deletion energy of the experts it removes. Figure 1b shows this relation for all kept sets. The Spearman correlation between removed energy and perplexity is 0.94 to 0.99 per model, but this pooled value is inflated because a smaller budget increases both quantities. Within a single budget, the ranking of the nine criteria by removed energy agrees with their ranking by perplexity at 0.45 to 0.98, except on OLMoE at 12.5%, where all criteria have collapsed (Table 15). The agreement is weaker for MC-8: 0.22 to 0.93 on Qwen3-30B, lower below 50% on Gemma-4, and negative at 75% and below on Qwen3.6, where MEP also has lower MC-8 accuracy than the averaged criteria (Section 7). On Qwen3-30B we also measured the layer error directly: the removed energy ranks kept sets by their layer output error with a Spearman correlation of 0.91, 0.86 and 0.67 at 50, 25 and 12.5%, and the layer error ranks them by KL divergence from the original model at 0.92 to 0.99 (Appendix D).

**Retained energy at moderate budgets.** The half of the experts with the largest $w_j$ holds 82 to 95.5% of a layer's deletion energy. MEP at 50% therefore removes 4.5 to 18% of it, and perplexity increases from 10.9 to 12.5 on Qwen3-30B and from 8.1 to 9.9 on Qwen3.6. REAP at 50% removes 8 to 20% and has higher perplexity on the three fine-grained models (17.2, 15.4 and 10.7, compared with 12.5, 13.3 and 9.9 for MEP). The small loss at this budget is explained by the low energy of the removed experts, not by reconstruction from the remaining ones.

**Summation versus averaging.** The kept set with the smallest removed deletion energy at budget $M$ consists of the top-$M$ experts by $w_j$, which is MEP, the $b$=0, $\alpha$=$\beta$=2 member of the family of Liu et al. (2026b). REAP, MAN and MSAN divide by $n_j$ instead. Averaging is justified only if the tokens of frequently routed experts can be reconstructed by the remaining experts, which Section 3 does not support. Table 3 shows the effect at 25%. On the three fine-grained models, the summed variants of REAP and MAN retain more energy than the averaged versions, and they have lower perplexity in five of these six cases; the exception is MAN on Gemma-4. Changing only the aggregation of REAP and MAN lowers perplexity in 15 of 18 settings on these models, by up to 4.3×, and raises MC-8 in all 12 settings on Qwen3-30B and Gemma-4, but lowers it in all 6 on Qwen3.6 (Table 17). Interpolating between the two, $w_j/n_j^b$, the best $b$ under matched calibration is 0 or 0.25 and $b$=1 is worst (Table 18). On the layer objective, which involves no evaluation noise, the averaged criteria leave 1.13 to 11.3 times the reconstruction error of the best set found by greedy search, the summed norm-based criteria 1.11 to 3.9 times and MEP 1.00 to 1.13 times (Table 20).

OLMoE is an exception that shows when the token count matters. Its routing is close to uniform (median Gini coefficient of routed-token counts 0.19, compared with 0.56 to 0.62 for the other

| Model | Kept | MAN | MoNE | SE | REAP | MEP |
|---|---|---|---|---|---|---|
| Gemma-4-26B | Original | | | 72.3 | | |
| | 50% | 41.1 (38.6–44.7) | 63.0 (62.4–63.5) | 59.7 (57.3–61.9) | 61.7 (60.6–62.2) | **69.6** (68.4–71.7) |
| | 37.5% | 16.2 (14.4–18.6) | 52.3 (51.3–52.8) | 43.6 (39.0–46.4) | 48.1 (47.1–49.1) | **69.5** (67.8–70.5) |
| | 25% | 5.1 (4.5–5.6) | 18.2 (17.6–19.2) | 13.9 (10.8–18.9) | 15.6 (13.1–17.8) | **60.0** (58.8–61.3) |
| Qwen3-30B | Original | | | 91.1 | | |
| | 50% | 82.3 (80.4–83.8) | 86.8 (84.9–88.5) | 86.6 (84.4–88.9) | 86.5 (86.2–86.7) | **88.2** (87.6–89.2) |
| | 37.5% | 76.6 (75.6–77.9) | **84.6** (83.9–85.3) | 83.7 (82.7–84.2) | 84.1 (81.8–85.6) | 78.5 (76.4–80.2) |
| | 25% | 43.1 (40.4–44.6) | 72.2 (71.0–74.5) | 58.9 (56.7–62.9) | 68.5 (65.6–70.4) | **79.3** (77.3–80.9) |
| Qwen3.6-35B | Original | | | 96.2 | | |
| | 37.5% | 91.8 (91.5–92.0) | 94.1 (93.9–94.3) | 94.2 (93.9–94.5) | 94.5 (94.3–94.7) | **95.3** (95.2–95.4) |
| | 25% | 80.6 (78.8–81.7) | 88.5 (85.7–90.0) | 45.9 (39.6–54.2) | 88.4 (85.5–89.9) | **92.7** (92.2–93.2) |
| | 12.5% | 5.6 (0.4–13.3) | 30.2 (18.7–41.1) | 12.7 (1.3–33.4) | 24.2 (11.6–34.8) | **42.4** (32.9–50.8) |

Table 4: GSM8K accuracy (1,319 problems) with the math calibration set: mean over three disjoint calibration samples, with the lowest and highest sample in parentheses. Bold: highest mean in the row. Qwen3.6-35B has twice as many experts and is shown at budgets one step lower.

| | Statistic | Median Spearman with loss change of zeroing | Layers where it is best (of 16) |
|---|---|---|---|
| Engmann et al. (2026) | utilization | $+0.33$ | 2 |
| | activation norm | $+0.34$ | 0 |
| | mean gate | $+0.42$ | 1 |
| | activation std | $+0.36$ | 0 |
| Ours | per-token mean $\bar{w}_j$ | $+0.52$ | 4 |
| | deletion energy $w_j$ | $+\mathbf{0.57}$ | **10** |

Table 5: Single-expert ablation on OLMoE-1B-7B-0924 with the checkpoint and statistics of Engmann et al. (2026): Spearman correlation over the $64$ experts of a layer between each statistic and the loss change caused by zeroing that expert, median over the $16$ layers, and the number of layers in which the statistic has the highest correlation. Removing the half of a layer with the lowest $w_j$ costs a median $0.54$ times a random half and $1.36$ times the sum of the single removals. Per-layer values in Table 14.

models), the count explains $2\%$ of the variance of $\log w_j$ compared with $56$ to $62\%$ elsewhere, and summation increases perplexity ($71$ versus $39$ for REAP at $50\%$; Table 17 and Figure 6). The count is informative when routing is uneven, consistent with the observation of Liu et al. (2026b) that task-specific pruning benefits from routing frequency.

**Low budgets and recovery training.** Since the loss follows the removed energy, every criterion collapses once the budget forces it to remove a large fraction of the energy, and criteria differ in the budget at which this happens. Table 4 shows this on GSM8K with task-matched calibration, averaged over three calibration samples. MEP has the highest mean in eight of nine settings, and the difference is largest where the other criteria collapse: on Gemma-4-26B at $25\%$, MEP keeps $60.0$ of the original $72.3$, while no other criterion exceeds $18.2$. The exception is Qwen3-30B at $37.5\%$, where MoNE, SE and REAP are $5$ to $6$ points higher with non-overlapping ranges. Table 26 gives, for GSM8K and HumanEval+, the smallest budget at which each criterion keeps $90\%$ of the original model's score. Recovery distillation ($500$ steps) improves every criterion by $2$ to $7$ points of multiple-choice accuracy on Qwen3-30B, and MEP remains best at every budget (Table 31). Retraining is required to recover the removed output.

## 5 Interaction-Aware Selection

If experts were redundant, selection that accounts for their interactions should differ from selection by individual scores. We test this with the layer reconstruction error $J(S)$ after the best fixed linear compensation by the kept experts. $J(S)$ depends on the full Gram matrix of gated expert outputs,

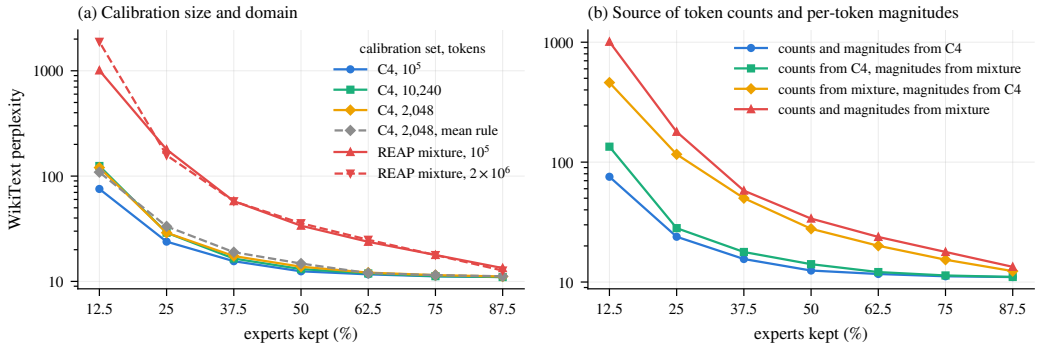


Figure 2: WikiText perplexity of MEP on Qwen3-30B with different calibration data (Gemma-4-26B and Qwen3.6-35B in Figure 7). (a) C4 calibration sets of $10^5$, 10,240 and 2,048 tokens, the mean rule ($w_j/n_j$) on 2,048 tokens (grey, dashed), and REAP's calibration mixture with $10^5$ and $2 \times 10^6$ tokens (red). (b) MEP scores an expert by its routed-token count times its mean per-token magnitude $g^2\|f\|^2$; the two factors are taken from different calibration sets.

which includes all pairwise interactions among removed experts and between removed and kept experts (Appendix F). We select by greedy minimization of $J(S)$, and with a second variant that computes the same statistics under the re-selected routing of the pruned model.

Selection with the full matrix keeps the same experts as MEP in 97 to 100% of the slots at 50%, 89 to 99% at 25% and 81 to 96% at 12.5% (Table 21). On GSM8K and HumanEval+ with three calibration samples, the ranges of both variants overlap those of MEP in all but two settings (Tables 23 and 24). On Gemma-4 GSM8K at 25% the re-routed variant is worse (44 to 49 versus 59 to 61), and on Qwen3-30B HumanEval+ at 25% it is better (27 to 56 versus 3 to 22), in a setting where all criteria vary strongly across calibration samples. Interactions are present and positive: at 50% they add 3 to 25% to the summed deletion energy of the removed experts (Table 22). They are too small to change the selection in most cases, which is consistent with a recovered fraction of at most 28%.

## 6 Calibration Domain and Calibration Size

If an expert's contribution is confined to the tokens routed to it, the cost of removing it depends on which tokens occur at deployment, and calibration data affect the selection through the routing they induce. Two predictions follow. Calibration data from a different domain should mislead MEP regardless of their size, and a small amount of data from the right domain should suffice, because routed-token counts converge quickly.

Figure 2 and Table 6 confirm both predictions. With REAP's released calibration mixture of code, math, science and agentic data, which does not match WikiText, MEP reaches a perplexity of 179 at 25% on Qwen3-30B, compared with 79 for REAP (Table 25). Increasing the mixture twentyfold to $2 \times 10^6$ tokens gives 158. A single 2,048-token chunk of C4 gives 29.0, compared with 23.9 for $10^5$ C4 tokens. Gemma-4 behaves in the same way (29.0 with 2,048 C4 tokens, compared with 95 and 81 with the two mixture sizes; Table 25 and Figure 7). On the same 2,048 tokens, the averaged rule has higher perplexity than MEP at five of seven budgets on Qwen3-30B and at all seven on Gemma-4.

The effect of the domain is almost entirely due to the routed-token counts. Using the per-token magnitudes from the mixture and the counts from C4 reduces perplexity at 25% from 179 to 28.1 on Qwen3-30B, from 95 to 32 on Gemma-4 and from 45 to 21 on Qwen3.6. The reverse combination gives 116, 61 and 28 (Table 6). A grid over twelve languages shows the same effect of calibration language, with a mismatch of up to 919 times the matched perplexity, and an equal-share mixture of the calibration domains stays within 2.1 times of the best calibration in every column of Figure 8a. The penalty of REAP's mixture comes from its composition: a $10^5$-token sample contains nine to eleven documents, mostly repository-level code. These results confirm the calibration sensitivity noted by Liu et al. (2026b); MEP is more sensitive to it than the averaged criteria. Instead of

| Calibration set | Tokens | Qwen3-30B | Gemma-4-26B | Qwen3.6-35B |
|---|---|---|---|---|
| Original model | | 10.9 | 8.7 | 8.1 |
| C4 (matched) | $10^5$ | 23.9 | 28.8 | 16.3 |
| C4 (matched) | 2,048 | 29.0 | 29.0 | 19.6 |
| C4, mean rule ($w_j/n_j$) | 2,048 | 33.3 | 31.5 | 20.2 |
| REAP mixture (mismatched) | $10^5$ | 179.0 | 95.1 | 45.4 |
| REAP mixture (mismatched) | $2 \times 10^6$ | 157.8 | 80.7 | n/a |
| counts from C4, magnitudes from mixture | $10^5$ | 28.1 | 31.9 | 20.6 |
| counts from mixture, magnitudes from C4 | $10^5$ | 116.2 | 61.1 | 27.9 |

Table 6: WikiText perplexity of MEP at a $25\%$ budget with different calibration sets. The last two rows take the routed-token counts and the per-token magnitudes of the score from different calibration sets. n/a: not evaluated.

averaging, or of avoiding calibration as AIMER does, we recommend calibrating on a few thousand tokens from the deployment domain.

## 7 Discussion

**Scope of the measurement.** The measurement does not show that experts are unrelated: a joint fit recovers up to $28\%$ of the pruned deletion energy on Qwen3-30B, mostly as a constant offset. It does not contradict the small effect of removing a single early-layer expert reported by Engmann et al. (2026); it shows that this effect is not due to reconstruction by other experts and does not extend to removing half of a layer. MEP is not the best criterion on every metric. On Gemma-4 MC-8 the traffic criterion is $5$ to $7$ points higher at $50$ and $25\%$. On Qwen3.6 MC-8 MEP is below every averaged criterion and MoNE at every budget, and below SE at $25\%$; compared with REAP at $25\%$ it is $15.7$ points lower on MMLU and $9.1$ and $6.8$ points lower on the two ARC tasks, and higher on HellaSwag and WinoGrande (Table 19). Knowledge-recall tasks may depend on rarely routed experts, which summation ranks low. On generation MEP is worse on Qwen3-30B GSM8K at $37.5\%$, where REAP's three samples ($81.8$ to $85.6$) lie above those of MEP ($76.4$ to $80.2$) and those of MoNE ($83.9$ to $85.3$) and SE ($82.7$ to $84.2$), and on Qwen3-30B HumanEval+ at $25\%$, where all criteria depend strongly on the calibration sample. The advantages of summation require calibration matched to the evaluation: with REAP's mixture, evaluated on WikiText and MC-8, the advantage of the summed criteria disappears on every model (Figure 9).

**Limitations.** We measure recovery by a fixed linear or affine combination of the remaining experts. The pruned model re-selects its top-$k$ experts and renormalizes their gates, which is a token-dependent combination; we do not measure its recovery directly and only observe in Figure 1b that its loss follows the removed energy. The recovered fraction is measured at a $25\%$ budget, and the positive controls are exact or near-exact substitutes; partial redundancy, for example in upcycled models, is not tested. Budgets are uniform across layers, all results are one-shot, and we do not evaluate Mixtral-8x7B, on which the CKA and weight-similarity results were reported. The calibration result specifies the domain to calibrate on; when the deployment domain is unknown, the equal-share mixture is the only option we evaluate.

## 8 Conclusion

Expert redundancy in MoE models has been inferred from similarity on shared inputs, from replacing secondary experts and from single-expert ablations. Measured on the tokens routed to the pruned expert, the fraction of its output that the remaining experts can reconstruct is small on four models. Pruning works at moderate budgets because deletion energy is concentrated in few experts, and within a budget the removed energy predicts perplexity. Accordingly, pruning scores should be summed over routed tokens rather than averaged, interaction-aware selection changes the kept set little, and calibration data should come from the deployment domain rather than be large.

## REFERENCES


Marmik Chaudhari, Idhant Gulati, Nishkal Hundia, Pranav Karra, and Shivam Raval. MoE Lens: An expert is all you need. *arXiv preprint arXiv:2603.05806*, 2026.

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 *International Conference on Machine Learning*, 2025.

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.

Zican Dong, Han Peng, Peiyu Liu, Wayne Xin Zhao, Dong Wu, Feng Xiao, and Zhifeng Wang. Domain-specific pruning of large mixture-of-experts models with few-shot demonstrations. *arXiv preprint arXiv:2504.06792*, 2025.

Leonard Engmann, Christian Medeiros Adriano, and Holger Giese. From observation to intervention: A causal audit of expert importance in mixture-of-experts models. *arXiv preprint arXiv:2606.10703*, 2026.

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.

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.

Jiawei Hao, Zhiwei Hao, Jianyuan Guo, Li Shen, Yong Luo, Han Hu, and Dan Zeng. LightMoE: Reducing mixture-of-experts redundancy through expert replacing. *arXiv preprint arXiv:2603.12645*, 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*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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.

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 *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, 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.

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.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. A closer look into mixture-of-experts in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4427–4447, 2025.

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.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

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.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 2020.

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.

Juntong Wu, Jialiang Cheng, Fuyu Lv, Ou Dan, and Li Yuan. SERE: Similarity-based expert re-routing for efficient batch decoding in MoE models. In *International Conference on Learning Representations*, 2026.

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.

Zeliang Zhang, Xiaodong Liu, Hao Cheng, Chenliang Xu, and Jianfeng Gao. Diversifying the expert knowledge for task-agnostic pruning in sparse mixture-of-experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

## 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, the accumulated per-expert Gram matrices behind Eqs. 2–3, and the scripts that produce each table and figure from them. The models are public checkpoints (Table 8); the calibration sets, their sources, chunking and the three disjoint calibration samples are specified in Appendix A. The statistics of Eqs. 1–3, the cross-fitting protocol, the null controls and the co-routed Gram are defined in Section 3 and Appendices A, B and F; selection is deterministic given those statistics. Likelihood and generation protocols, including the three model-specific deviations we had to make and why, are stated in Appendix A, together with the contamination check, the evaluation-noise measurement and the compute used.

## AI Use Statement

Large language models (Anthropic's Claude, accessed through the Claude Code agentic interface) were used throughout this work under the authors' direction and with the authors' verification of every result. Specifically: for retrieval and discovery of related work, including the reading of prior papers whose measurements Section 2 characterizes, each of which the authors verified against the original text; for research execution, including writing and running the analysis and evaluation code that produced the statistics, kept sets and tables reported here, and for proposing experiments and controls that the authors selected among; for drafting sections of the paper and the appendix from the authors' outline and results, which the authors edited; for checking the closed forms in Appendix F; and for figure creation and editing. No synthetic data were generated. The research questions, the claims, the experimental design decisions and the interpretation of results are the authors'. The authors take full responsibility for the content.

## A Protocol

Table 7 collects the symbols and Table 8 the models.

| Symbol | Where | Meaning |
|---|---|---|
| $E$, $k$, $M$ | Sec. 2 | experts per layer, experts routed per token, experts kept per layer; the budget is $M/E$ |
| $g_j(x)$, $f_j(x)$ | Sec. 2 | router gate and output of expert $j$ on token $x$ |
| $X_j$, $n_j$ | Sec. 2 | tokens routed to $j$ in the calibration set, and their number |
| $d_j(x) = g_j(x)f_j(x)$ | Sec. 2 | contribution of $j$ to the layer output on $x$, removed when $j$ is pruned |
| $w_j$ | Eq. 1 | deletion energy, $\sum_{x\in X_j} \|d_j(x)\|^2$: the squared error caused by removing $j$ |
| $\langle u, v\rangle_j$ | Eq. 2 | inner product over $j$'s tokens, each weighted by $g_j(x)^2$ |
| $\rho^2_{ij}$ | Eq. 2 | fraction of $w_j$ recovered by one remaining expert $i$ with optimal scale |
| $R^2_j(S)$ | Eq. 3 | fraction of $w_j$ recovered by a joint linear fit of the kept set $S$ |
| $S$, $P$ | Eq. 3 | kept and deleted set of a layer |
| $G_{ij}$ | App. F | Gram matrix over co-routed tokens, $\sum_x g_i g_j \langle f_i, f_j\rangle$, with $G_{jj} = w_j$ |
| $J(S)$ | Sec. 5 | layer error after the best fixed linear compensation by the kept experts |
| $b$, $\alpha$, $\beta$ | Sec. 2 | the family $n_j^{-b} \sum_{x\in X_j} g_j^\alpha \|f_j\|^\beta$ of Liu et al. (2026b); $b$=0 sums, $b$=1 averages |
| MEP | Sec. 1 | maximum energy preservation: keep the top-$M$ experts by $w_j$, the $b$=0, $\alpha$=$\beta$=2 member |

Table 7: Notation.

| Model | Routed experts | Shared or dense path | Top-$k$ | Layers | Parameters total | Parameters active |
|---|---|---|---|---|---|---|
| OLMoE-1B-7B | 64 | none | 8 | 16 | 6.9B | 1.3B |
| Qwen3-30B-A3B | 128 | none | 8 | 48 | 30.5B | 3.3B |
| Gemma-4-26B-A4B | 128 | dense MLP* | 8 | 30 | 26.0B | 4.0B |
| Qwen3.6-35B-A3B | 256 | 1 expert | 8 | 40 | 35.0B | 3.3B |

Table 8: Models. Routed experts, shared or dense path, experts per token, layers, and total and active parameters. *Gemma-4-26B has an always-active dense MLP in parallel with the routed experts, about three times the width of an expert; Qwen3.6-35B has one gated shared expert of expert width. In figures and tables we use the short names OLMoE-1B, Qwen3-30B, Gemma-4-26B and Qwen3.6-35B.

**Models and calibration sets.** The 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` (Muennighoff et al., 2024; Qwen Team, 2025; Gemma Team, 2026; Qwen Team, 2026). Calibration sets are packed into 2,048-token chunks in each model's tokenizer. The general calibration set is English C4 (`allenai/c4`). The code calibration set is the calibration mixture released with REAP, referred to as REAP's mixture: evol-codealpaca (Luo et al., 2024) and the MoT code, math and science, xLAM and SWE-smith sets as released with REAP, 4,096 documents each, shuffled with seed 7. It is task-matched for HumanEval+ and mismatched for WikiText and MC-8. The math calibration set is tulu-3 personas-math (Lambert et al., 2024). The $10^5$-token setting uses 50 chunks. The first calibration sample is chunks 0–49, and the two repeat samples are chunks 100–149 and 200–249.

**Statistics.** The deletion energy of Eq. 1 is accumulated in one forward pass over the calibration set and ranked per layer, so every kept set is a function of the calibration set alone. The inner products behind Eqs. 2 and 3 and the co-routed Gram $G$ are accumulated in separate passes that evaluate every expert on every routed token, with a per-expert cap of 131,072 tokens that is never reached at $10^5$ tokens. Cross-fitting uses chunks 0–24 and 25–49 as halves, fits with a ridge of $10^{-6}$ times the mean diagonal, scores on the other half and averages the two directions. Moving the ridge from $10^{-6}$ to $10^{-2}$ does not change the fraction of the total pruned deletion energy recovered by the linear joint fit (22.9% on Qwen3-30B and 1.4% on Qwen3.6 in that run, not tabulated).

**Likelihood evaluation.** The base model is loaded once and the router mask is swapped per selection, with top-$k$ re-selection and gate re-normalization active. WikiText-2 word-level perplexity and the eight multiple-choice tasks of MC-8 (ARC-challenge and ARC-easy, Clark et al., 2018; BoolQ, Clark et al., 2019; HellaSwag, Zellers et al., 2019; MMLU, Hendrycks et al., 2021; OpenBookQA, Mihaylov et al., 2018; RTE, Wang et al., 2019; WinoGrande, Sakaguchi et al., 2020; zero-shot accuracy) use `lm-eval` defaults; Qwen3.6 uses a $4{,}096$-token sample. The Kullback-Leibler (KL) divergence of Figure 5 is measured on chunks 50–57 of each model's C4 calibration set against the original model's top-$64$ next-token distribution plus a tail term, under the same mask.

**Generation.** Pruned checkpoints are exported and evaluated at published settings: GSM8K with `lm-eval`, 8-shot, strict match, greedy, $1{,}024$ new tokens; HumanEval+ with EvalPlus, greedy, chat template, thinking disabled, $768$ new tokens. GSM8K is calibrated on the math calibration set and HumanEval+ on the code calibration set. Gemma-4 ships no chat template, so its turn format is applied from a template file to every criterion and to the original model alike. Qwen3.6 is a hybrid reasoning model whose pruned selections open a reasoning block that is cut at the token limit, so its GSM8K uses the chat template with reasoning disabled.

**Recovery, noise, contamination and compute.** Recovery distillation matches the original model's logits for $500$ steps on FineWeb (Penedo et al., 2024) (batch $4$ of $1{,}024$-token sequences, about 2M tokens, learning rate $5 \times 10^{-5}$), training the router and rank-16 low-rank adapters (LoRA, Hu et al., 2022) on experts and attention. Repeating one HumanEval+ evaluation three times returned $45.7/45.1/45.1$, $36.0/33.5/40.2$ and $55.5/55.5/56.1$ for three selections, so differences inside that spread are noise; GSM8K reproduced to the digit. A 13-gram check finds no overlap between the calibration sets and any GSM8K or HumanEval+ item. All runs use single A100 80GB GPUs. One selection takes $18$ to $130$ seconds per model and one recoverability pass $5$ to $46$ minutes. Kept sets, raw results and the scripts that regenerate every table and figure ship with the submission.

## B Recoverability: Distributions and Controls

**Distributions.** Figure 3 shows the per-expert values behind the medians of Table 9. The mass at the left edge of the joint panel is experts whose out-of-sample $R^2$ is at or below zero. Fitted and scored on the same tokens, the kept set reaches medians of only $0.5$, $0.4$, $6.7$ and $0.2\%$ on the four models, as small as the cross-fitted $0.5$, $0.2$, $6.5$ and $0.01\%$. Cross-fitting therefore does not remove a substitute that an in-sample fit would find, and the values at the edge correspond to fits that do not generalize.

| | Best single expert ($\rho^2$) | | All remaining experts ($R^2$) | | Constant vector |
|---|---|---|---|---|---|
| Model | median expert | total | linear | linear + constant | total |
| OLMoE-1B | 0.20% | 4.3% | 7.6% | 13.6% | 11.1% |
| Qwen3-30B | 0.09% | 17.2% | 23.0% | 27.9% | 23.9% |
| Gemma-4-26B | 2.11% | 2.2% | 4.0% | 21.6% | 20.6% |
| Qwen3.6-35B | 0.02% | 0.7% | 1.4% | 1.6% | 3.4% |

Table 9: Details of Table 2. Fraction of the deletion energy of the experts pruned at a $25\%$ budget that remaining experts recover on the pruned experts' routed tokens (C4 calibration set, all layers; fits use one half of the calibration set and are evaluated on the other). "Median expert": median over pruned experts; "total": fraction of the summed deletion energy of all pruned experts.

**Intercept and the last layer of Qwen3-30B.** In Table 9 a constant vector recovers as much as the linear fit or more. On Qwen3-30B the recovered fraction is concentrated in the last layer, where the linear fit recovers $55\%$, the fit with an intercept $62\%$ and a constant alone $51\%$; the expert outputs in this layer are dominated by a shared offset.

**Null controls.** The tables of this appendix use different token samples (Table 10: sixteen chunks per layer; Table 11: two or three layers; Table 12: $1{,}024$ to $8{,}192$ tokens), so the same statistic can differ between them by a few tenths of a percent. Table 10 compares the best expert of the same

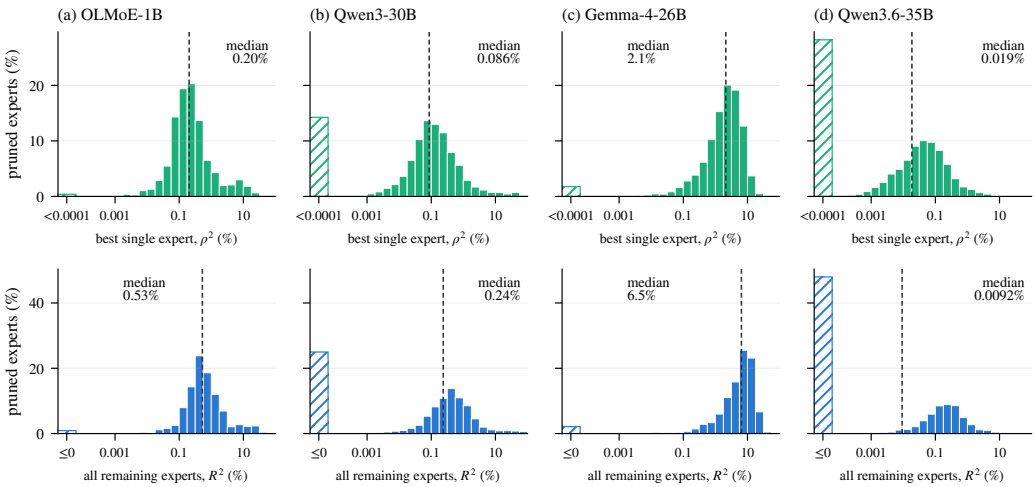


Figure 3: Recovered fraction for each expert pruned at a 25% budget (C4 calibration set, all layers), as a percentage of its deletion energy. Top: best single remaining expert ($\rho^2$, Eq. 2). Bottom: all remaining experts ($R^2$, Eq. 3), fitted on one half of the calibration set and evaluated on the other. The hatched bar at the left collects values below 0.0001% (top) or at or below zero (bottom; 1, 25, 2 and 48% of pruned experts). Dashed line: median.

layer with reconstructions that carry no information about the pruned expert's output on its routed tokens. On Qwen3-30B and Qwen3.6 the same-layer value is 5 to 14 times the shuffled-token value and within 1.2× of the other-layer value in one layer. On Gemma-4 the shuffled-token value almost matches the same-layer value, which indicates that the similarity among Gemma-4's experts is a token-independent offset.

| Model | Layer | Same layer | Other layer | Same layer, shuffled tokens |
|---|---|---|---|---|
| Qwen3-30B | 4 | 0.36% | 0.09% | 0.04% |
| | 12 | 0.13% | 0.09% | 0.01% |
| | 20 | 0.22% | 0.10% | 0.02% |
| | 28 | 0.17% | 0.14% | 0.02% |
| | 36 | 0.15% | 0.04% | 0.03% |
| Gemma-4-26B | 2 | 1.52% | 0.01% | 0.86% |
| | 8 | 6.23% | 0.02% | 5.60% |
| | 14 | 5.50% | 0.01% | 4.84% |
| | 20 | 2.28% | 0.18% | 1.56% |
| Qwen3.6-35B | 4 | 0.14% | 0.05% | 0.01% |
| | 12 | 0.07% | 0.05% | 0.01% |
| | 20 | 0.10% | 0.04% | 0.01% |
| | 28 | 0.10% | 0.02% | 0.01% |

Table 10: Median over experts of the best single-expert $\rho^2$ on an expert's routed tokens, with candidates from the same layer, from a layer eight blocks deeper (evaluated on the same tokens), and from the same layer with token order shuffled. A fixed random direction gives 0.001 to 0.004% in every layer. Sixteen C4 chunks, all experts of the layer.

**Positive controls.** Table 11 shows that both estimators approach one when a substitute exists. A copy of the expert with 10% relative weight noise recovers 98% of its deletion energy, whereas the best other expert recovers 0.15 to 0.41%. Added to the pool of all other experts, the same copy raises the cross-fitted joint $R^2$ from about 1% to 98%, so cross-fitting with 64 regressors retains a real substitute. A synthetic target built as a random combination of eight kept experts is recovered with a median cross-fitted $R^2$ of 100% on every model.

| | | Noisy copy of the expert, $\rho^2$ | | | | Best other | All other experts, $R^2$ | |
|---|---|---|---|---|---|---|---|---|
| Model | Layer | 1% noise | 5% | 10% | 30% | expert, $\rho^2$ | without copy | with 10% copy |
| OLMoE-1B | 4 | 100.0% | 99.5% | 98.0% | 83.6% | 0.23% | 1.1% | 98.0% |
| | 12 | 100.0% | 99.5% | 98.1% | 84.5% | 0.35% | 1.4% | 98.1% |
| Qwen3-30B | 4 | 100.0% | 99.6% | 98.3% | 86.2% | 0.41% | n/a | n/a |
| | 20 | 100.0% | 99.5% | 98.0% | 83.4% | 0.21% | n/a | n/a |
| | 36 | 100.0% | 99.4% | 97.8% | 82.4% | 0.15% | n/a | n/a |

Table 11: Positive controls. Median $\rho^2$ on an expert's routed tokens for a copy of the expert with Gaussian weight noise of the given relative size, and for the best other expert. Right: median $R^2$ of all other experts of the layer, without and with the $10\%$-noise copy (fitted on one half of the tokens, evaluated on the other; OLMoE-1B only, n/a otherwise). A synthetic target formed as a random combination of eight kept experts is recovered with median $R^2$ of $100.0\%$ on all four models.

**Representational similarity on the same tokens.** Table 12 computes, for every pair of experts in a layer, the linear CKA of their outputs and $\rho^2$ on the routed tokens. CKA is in the range that similarity studies interpret as redundancy, $\rho^2$ is close to zero, and the two are weakly correlated across pairs.

| | | Linear CKA (all tokens) | | $\rho^2$ (routed tokens) | | |
|---|---|---|---|---|---|---|
| Model | Layer | median pair | best per expert | median pair | best per expert | Spearman(CKA, $\rho^2$) |
| OLMoE-1B | 2 | 0.20 | 0.27 | 0.016% | 0.24% | +0.31 |
| | 6 | 0.20 | 0.28 | 0.018% | 0.27% | +0.26 |
| | 10 | 0.23 | 0.32 | 0.007% | 0.36% | +0.28 |
| | 14 | 0.25 | 0.38 | 0.103% | 1.22% | +0.22 |
| Qwen3-30B | 4 | 0.26 | 0.33 | 0.008% | 0.51% | +0.14 |
| | 20 | 0.37 | 0.46 | 0.008% | 0.30% | +0.14 |
| | 36 | 0.26 | 0.37 | 0.005% | 0.21% | +0.02 |
| Gemma-4-26B | 2 | 0.19 | 0.27 | 0.284% | 1.68% | +0.29 |
| | 8 | 0.21 | 0.31 | 3.399% | 7.08% | +0.41 |
| | 14 | 0.28 | 0.36 | 1.937% | 5.62% | +0.08 |
| | 20 | 0.21 | 0.40 | 0.590% | 2.50% | +0.19 |
| Qwen3.6-35B | 4 | 0.35 | 0.41 | 0.003% | 0.21% | +0.01 |
| | 12 | 0.33 | 0.41 | 0.003% | 0.13% | +0.12 |
| | 20 | 0.41 | 0.55 | 0.004% | 0.15% | +0.31 |
| | 28 | 0.34 | 0.47 | 0.004% | 0.20% | +0.06 |

Table 12: CKA and recovered fraction for the same expert pairs, $1{,}024$ to $8{,}192$ C4 tokens per layer. Linear CKA is computed on all tokens (the measure used by Zhang et al. (2025)); $\rho^2$ (Eq. 2) on the routed tokens of each expert. For each measure: median over pairs and median over experts of the best partner. Last column: Spearman correlation between CKA and $\rho^2$ over pairs. Experts with fewer than eight routed tokens are omitted.

**Concentration of expert output.** Table 13 gives the ratio used in the last paragraph of Section 3: the fraction of an expert's output energy on its routed tokens, divided by the mean of the same fraction for the other experts. The ratio correlates weakly with the recovered fraction on three models and moderately on Qwen3.6 ($+0.46$), so it does not identify which experts can be replaced.

## C Single-Expert Ablation

Table 14 uses the checkpoint and the four statistics of Engmann et al. (2026), OLMoE-1B-7B-0924, with a simpler design than theirs, which zeroes the highest- and lowest-ranked expert at sampled token positions. Each expert of each layer is zeroed in turn with routing and gates unchanged, and the mean change in next-token loss over $16{,}376$ next-token targets ($8$ held-out C4 chunks) is recorded. Their four routing statistics and our two are accumulated on the same tokens and correlated with the

| Model | Ratio $> 1$ | Ratio 25th pct. | Ratio median | Ratio 75th pct. | Spearman with best $\rho^2$ |
|---|---|---|---|---|---|
| OLMoE-1B | 96% | 1.87 | 2.21 | 2.72 | −0.20 |
| Qwen3-30B | 96% | 1.60 | 2.02 | 2.71 | +0.17 |
| Gemma-4-26B | 89% | 1.36 | 1.94 | 2.71 | +0.04 |
| Qwen3.6-35B | 97% | 1.58 | 2.15 | 2.91 | +0.46 |

Table 13: Output energy of an expert on its routed tokens relative to the other experts, after dividing by each expert's total output energy. Fraction of experts with a ratio above one, quartiles of the ratio, and its Spearman correlation with the best $\rho^2$ (C4 calibration set, all layers).

loss change across the layer's $64$ experts. The deletion energy has the highest median correlation, and it is above all four of their statistics in $11$ of $16$ layers. The right block tests deletion at pruning scale: zeroing the half of a layer with the smallest $w_j$ costs a median $0.54$ times a random half (the median of the per-layer ratios), and $1.36$ times the sum of its members' single deletions.

| Layer | Statistics of Engmann et al. (2026) Util. | Act. norm | Gate | Act. std | Ours $\bar{w}_j$ | $w_j$ | Zeroing half the layer lowest $w_j$ | random | ratio |
|---|---|---|---|---|---|---|---|---|---|
| 0 | +0.26 | +0.19 | +0.35 | +0.10 | +0.41 | **+0.45** | 0.88 | 1.66 | 2.98 |
| 1 | +0.09 | +0.03 | +0.09 | +0.42 | **+0.43** | +0.32 | 2.13 | 2.94 | 1.55 |
| 2 | +0.29 | +0.23 | +0.38 | +0.32 | **+0.44** | **+0.44** | 1.12 | 3.98 | 2.01 |
| 3 | +0.25 | +0.34 | **+0.52** | +0.30 | +0.51 | +0.50 | 1.41 | 2.30 | 2.34 |
| 4 | +0.35 | +0.45 | +0.50 | +0.33 | **+0.58** | +0.57 | 0.81 | 2.72 | 3.05 |
| 5 | +0.01 | +0.07 | +0.30 | +0.27 | **+0.41** | +0.20 | 1.20 | 1.65 | 1.32 |
| 6 | **+0.45** | +0.12 | +0.34 | +0.14 | +0.27 | +0.36 | 1.29 | 2.20 | 1.02 |
| 7 | +0.45 | +0.35 | +0.44 | +0.14 | +0.62 | **+0.64** | 1.03 | 2.36 | 1.41 |
| 8 | +0.28 | +0.16 | +0.45 | +0.28 | +0.42 | **+0.46** | 1.12 | 1.75 | 1.29 |
| 9 | +0.31 | +0.35 | +0.50 | +0.42 | +0.54 | **+0.62** | 1.61 | 2.88 | 1.45 |
| 10 | +0.23 | +0.28 | +0.43 | +0.39 | +0.53 | **+0.57** | 1.93 | 4.33 | 1.17 |
| 11 | +0.65 | +0.54 | +0.40 | +0.52 | +0.58 | **+0.71** | 2.08 | 4.21 | 1.03 |
| 12 | +0.59 | +0.54 | +0.17 | +0.47 | +0.56 | **+0.72** | 3.03 | 4.83 | 1.08 |
| 13 | +0.72 | +0.65 | +0.44 | +0.77 | +0.74 | **+0.82** | 3.23 | 6.50 | 1.08 |
| 14 | +0.50 | +0.35 | +0.56 | +0.51 | +0.65 | **+0.76** | 3.14 | 7.71 | 1.23 |
| 15 | **+0.79** | +0.58 | +0.34 | +0.44 | +0.46 | +0.72 | 5.63 | 8.39 | 1.43 |
| median | +0.33 | +0.34 | +0.42 | +0.36 | +0.52 | +0.57 | 1.51 | 2.91 | 1.36 |

Table 14: Single-expert ablation on OLMoE-1B-7B-0924 with the checkpoint and statistics of Engmann et al. (2026). Left: per layer, Spearman correlation over the $64$ experts between each statistic and the loss change caused by zeroing the expert (utilization, activation norm, mean gate and activation standard deviation from Engmann et al. (2026); our per-token mean $\bar{w}_j$ and deletion energy $w_j$); bold, the highest in the row. Right: loss change (in $10^{-2}$ nats/token) when zeroing the half of the layer with the lowest $w_j$ or a random half, and the ratio of the former to the sum of the corresponding single-expert changes.

## D Pruned Deletion Energy, Layer Error and Perplexity

Figure 4 separates the criteria of Figure 1b. On the three fine-grained models the curves of different criteria lie close together, so a criterion's perplexity depends mainly on how much energy it removes. On OLMoE, REAP-sum and MEP have higher perplexity than the averaged criteria at equal removed energy, consistent with the OLMoE exception in Section 4. Table 15 gives the within-budget correlations quoted in Section 4. The removed energy orders the criteria by perplexity at almost every budget; for MC-8 it does so on Qwen3-30B but not on Gemma-4 at low budgets or on Qwen3.6. On Qwen3-30B we also measure the intermediate quantities. On six layers ($4$, $12$, $20$, $28$, $36$, $44$) and $16{,}384$ C4 tokens that no calibration used, we apply each kept set with the deployed mechanism and measure the relative squared error of the layer output. Table 16 ranks $100$ random

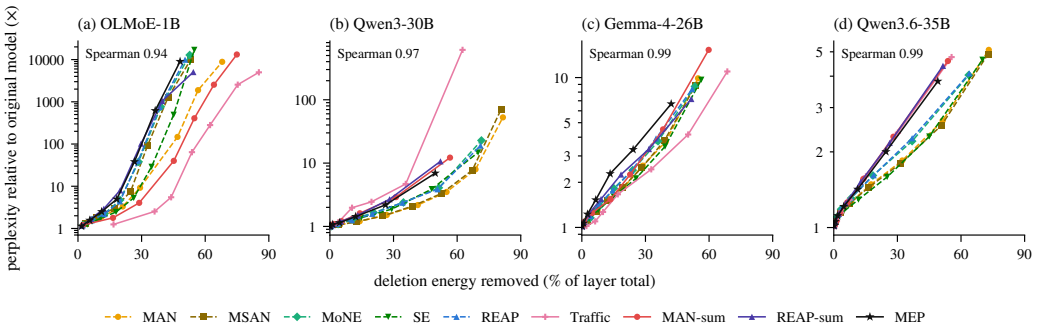


Figure 4: Figure 1b for each criterion. Each line connects the seven budgets of one criterion (C4 calibration set). Spearman correlations are computed over all criteria and budgets of a model. Vertical axes differ between panels.

| | | Experts kept (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Metric | 87.5 | 75 | 62.5 | 50 | 37.5 | 25 | 12.5 |
| OLMoE-1B | perplexity | +0.85 | +0.87 | +0.70 | +0.88 | +0.67 | +0.75 | −0.13 |
| | MC-8 accuracy | +0.30 | +0.25 | +0.87 | +0.67 | +0.82 | +0.03 | +0.40 |
| Qwen3-30B | perplexity | +0.63 | +0.97 | +0.98 | +0.90 | +0.90 | +0.88 | +0.73 |
| | MC-8 accuracy | +0.22 | +0.80 | +0.92 | +0.63 | +0.72 | +0.93 | +0.67 |
| Gemma-4-26B | perplexity | +0.87 | +0.88 | +0.87 | +0.90 | +0.63 | +0.45 | +0.97 |
| | MC-8 accuracy | +0.85 | +0.68 | +0.63 | +0.40 | +0.12 | −0.10 | −0.17 |
| Qwen3.6-35B | perplexity | +0.87 | +0.90 | +0.87 | +0.93 | +0.90 | +0.80 | +0.70 |
| | MC-8 accuracy | +0.55 | −0.08 | −0.62 | −0.73 | −0.52 | −0.58 | −0.47 |

Table 15: Spearman correlation over the nine criteria, within one budget, between the deletion energy removed by the kept set and WikiText perplexity, and between the removed energy and the drop in MC-8 accuracy (positive: more removed energy, lower accuracy). C4 calibration set.

kept sets and the criterion sets at each budget by several candidate objectives and correlates each with the measured error. Figure 5 continues from layer error to divergence from the original model and from divergence to perplexity.

| | Experts kept (%) | | |
|---|---|---|---|
| Fraction of the layer total that a kept set removes | 50 | 25 | 12.5 |
| MAN score removed | 0.08 | −0.03 | −0.23 |
| MoNE score removed | 0.48 | 0.39 | 0.25 |
| REAP score removed | 0.43 | 0.27 | 0.05 |
| deletion energy removed (calibration tokens) | 0.91 | 0.86 | 0.67 |
| deletion energy removed (held-out tokens) | 0.98 | 0.89 | 0.69 |

Table 16: Spearman correlation within each budget between the fraction of an objective that a kept set removes and the measured relative error of the layer output, over 100 random and all criterion kept sets (Qwen3-30B, six layers, held-out tokens).

## E Summation and Averaging on All Models

Table 17 gives the summed and averaged criteria on all four models at three budgets, with perplexity and MC-8. It is the source of the counts in Section 4: changing REAP and MAN from mean to sum lowers perplexity in 15 of 18 cells on the three fine-grained models and in 1 of 6 on OLMoE.

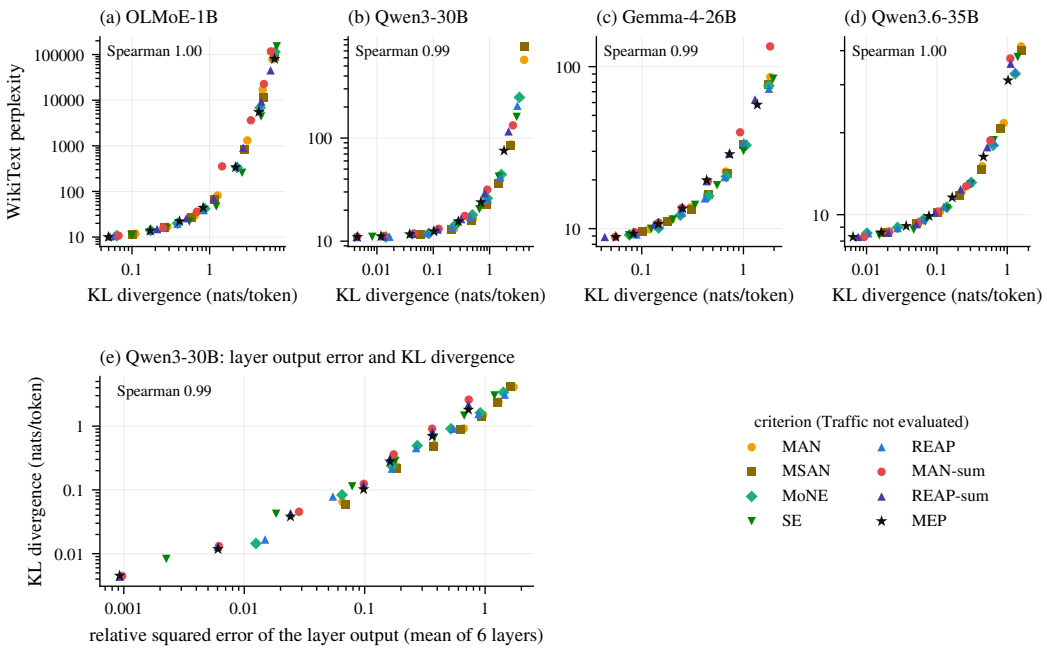


Figure 5: (a–d) KL divergence of each pruned model from the original model on held-out C4 tokens against WikiText perplexity, one point per criterion and budget. Printed Spearman values are computed over all budgets; within a budget they are 0.57 to 1.00 (OLMoE-1B), 0.69 to 1.00 (Qwen3-30B), 0.57 to 0.98 (Gemma-4-26B) and 0.77 to 0.94 (Qwen3.6-35B). (e) Qwen3-30B: mean relative error of the layer output over six layers against the same KL divergence; Spearman 0.92 to 0.99 within each budget. Traffic was not evaluated.

**Interpolation between sum and mean.** Table 18 varies the exponent $b$ of the count in $w_j/n_j^b$. Under calibration matched to the evaluation domain the best $b$ is 0 or 0.25 and the mean is worst. With REAP's mixture, which does not match WikiText, the order reverses and the mean is best, consistent with Section 6.

**Role of the token count.** Figure 6 relates the OLMoE exception to routing imbalance. The deletion energy factors as $w_j = n_j \bar{w}_j$ with $\bar{w}_j$ the per-token mean. The count explains 56 to 62% of the variance of $\log w_j$ on the three fine-grained models but 2% on OLMoE, whose routing is close to uniform. Within every model, the layers where summing retains the most extra energy are the layers with the most uneven traffic (Spearman 0.40 to 0.75).

**Per-task accuracy on Qwen3.6.** Table 19 splits Qwen3.6's MC-8 by task. The difference to REAP is concentrated on MMLU and the two ARC tasks, and MEP is higher on HellaSwag and WinoGrande.

## F Interaction-Aware Selection

**Closed forms.** Let $G^{(j)}_{ik} = \langle f_i, f_k \rangle_j$ be the Gram matrix of all experts on $j$'s tokens and $a_S = G^{(j)}_{Sj}$. The joint share of Eq. 3 is $R^2_j(S) = a_S^\top (G^{(j)}_{SS})^{-1} a_S / w_j$, with a ridge of $10^{-6}$ times the mean diagonal. The co-routed Gram of the layer is $G_{ij} = \sum_x g_i(x) g_j(x) \langle f_i(x), f_j(x) \rangle$, nonzero only for experts routed together, with $G_{jj} = w_j$. Deleting $P = [E] \setminus S$ removes $d_P = \sum_{j \in P} d_j$, and the best fixed linear compensation by the remaining experts leaves $J(S) = \mathbf{1}^\top G \mathbf{1} - h_S^\top G_{SS}^{-1} h_S$ with $h = G\mathbf{1}$. When the off-diagonal of $G$ vanishes, $J$ is modular and minimized by keeping the top-$M$ experts by $w_j$, so MEP is the diagonal limit of this objective. Greedy selection with swap refinement reaches the same value of $J$ as the best of ten random-restart swap searches to within

| | WikiText perplexity | | | MC-8 accuracy (%) | | |
|---|---|---|---|---|---|---|
| Experts kept (%) | 50 | 25 | 12.5 | 50 | 25 | 12.5 |
| *OLMoE-1B*, original model: perplexity 8.9, MC-8 58.0 | | | | | | |
| MAN | 81.9 | 16770 | 78731 | 40.1 | 32.1 | 31.5 |
| MSAN | 66.1 | 11398 | 90138 | 39.8 | 31.8 | 32.1 |
| REAP | 39.0 | 6724 | 88454 | 40.7 | 32.5 | 31.6 |
| Traffic | 570.3 | 22632 | 44415 | 39.8 | 32.4 | 31.6 |
| MAN-sum | 355.4 | 22601 | 117376 | 38.9 | 32.1 | 31.6 |
| REAP-sum | 71.2 | 9312 | 45105 | 40.9 | 31.3 | 31.6 |
| MEP | 44.0 | 5503 | 80283 | 41.5 | 32.2 | 32.0 |
| *Qwen3-30B*, original model: perplexity 10.9, MC-8 68.2 | | | | | | |
| MAN | 23.7 | 86.7 | 573.5 | 51.8 | 38.2 | 33.6 |
| MSAN | 22.7 | 84.3 | 774.2 | 52.9 | 38.2 | 33.0 |
| REAP | 17.2 | 42.0 | 205.7 | 59.1 | 40.2 | 34.9 |
| Traffic | 21.5 | 51.0 | 6679 | 51.7 | 39.9 | 31.3 |
| MAN-sum | 13.2 | 31.6 | 132.8 | 63.6 | 45.1 | 36.6 |
| REAP-sum | 13.0 | 29.2 | 116.1 | 63.5 | 46.3 | 36.9 |
| MEP | 12.5 | 23.9 | 75.5 | 63.0 | 47.1 | 35.7 |
| *Gemma-4-26B*, original model: perplexity 8.7, MC-8 69.2 | | | | | | |
| MAN | 16.2 | 32.5 | 86.2 | 50.4 | 38.6 | 36.3 |
| MSAN | 16.1 | 32.9 | 77.7 | 50.1 | 37.9 | 34.9 |
| REAP | 15.4 | 34.2 | 73.0 | 50.8 | 38.6 | 36.3 |
| Traffic | 14.5 | 36.3 | 95.7 | 61.8 | 47.3 | 39.4 |
| MAN-sum | 13.4 | 39.3 | 133.4 | 61.8 | 44.9 | 37.4 |
| REAP-sum | 13.5 | 29.1 | 62.7 | 58.9 | 45.6 | 38.1 |
| MEP | 13.3 | 28.8 | 58.2 | 54.8 | 42.3 | 36.6 |
| *Qwen3.6-35B*, original model: perplexity 8.1, MC-8 69.2 | | | | | | |
| MAN | 12.1 | 21.7 | 41.5 | 69.4 | 57.3 | 42.7 |
| MSAN | 11.8 | 20.8 | 39.9 | 69.1 | 57.4 | 43.2 |
| REAP | 10.7 | 18.5 | 33.4 | 68.2 | 57.6 | 43.5 |
| Traffic | 10.5 | 19.4 | 38.8 | 66.5 | 52.3 | 41.2 |
| MAN-sum | 10.3 | 18.7 | 37.5 | 66.5 | 53.1 | 41.4 |
| REAP-sum | 10.2 | 17.7 | 35.8 | 66.5 | 52.6 | 42.4 |
| MEP | 9.9 | 16.3 | 31.1 | 67.7 | 53.6 | 42.5 |

Table 17: WikiText perplexity and MC-8 accuracy of the averaged criteria, their summed variants and MEP on all four models (C4 calibration set). The values of the original model are given in each block header. MoNE and SE are not included.

| | | 25% of experts kept | | | | | 12.5% of experts kept | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Calibration | $b$=0 | 0.25 | 0.5 | 0.75 | 1 | $b$=0 | 0.25 | 0.5 | 0.75 | 1 |
| Qwen3-30B | C4 | 23.9 | **23.1** | 23.5 | 27.7 | 50.1 | 75.5 | **63.1** | 71.1 | 98.5 | 269.8 |
| | REAP mixture | 179.0 | 154.9 | 118.5 | 94.2 | **76.1** | 1008 | 864.7 | 737.0 | 474.7 | **347.5** |
| Gemma-4-26B | C4 | **28.8** | 29.6 | 30.2 | 30.6 | 34.5 | **58.2** | 59.6 | 59.4 | 69.5 | 79.3 |
| | REAP mixture | 95.1 | 83.3 | 72.2 | 64.1 | **53.0** | 169.5 | 162.6 | 137.0 | 121.4 | **97.9** |

Table 18: WikiText perplexity of the rule $w_j/n_j^b$ from the sum ($b$=0, MEP) to the mean ($b$=1), with C4 calibration (matched to the evaluation) and with REAP's mixture (mismatched). Bold: best $b$ within each budget and row.

$0.2\%$ in every layer checked on Qwen3-30B at $25\%$; we use it as the reference in Table 20, which is the evaluation-free comparison quoted in Section 4.

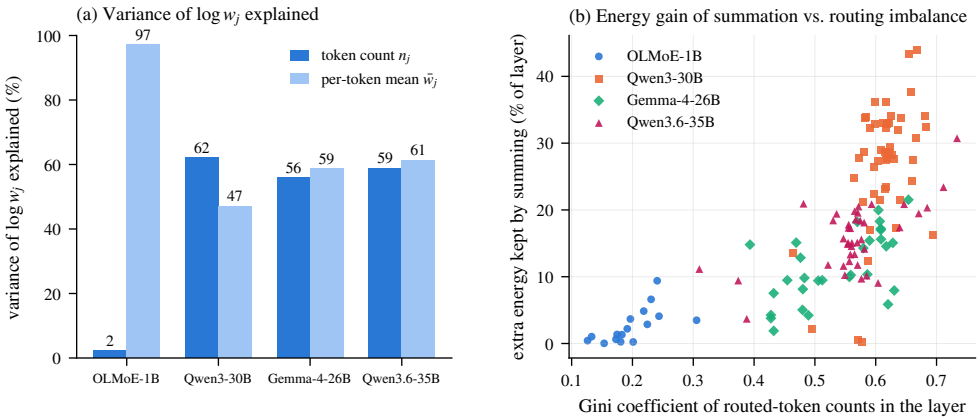


Figure 6: (a) Fraction of the variance of $\log w_j$ explained by $\log n_j$ alone and by $\log \bar{w}_j$ alone, where $w_j = n_j \bar{w}_j$ (C4 calibration set, all layers). (b) Per layer, the Gini coefficient of routed-token counts against the additional deletion energy kept by summing instead of averaging at a 25% budget, as a percentage of the layer total.

| 50% kept | ARC-c | ARC-e | BoolQ | HellaSwag | MMLU | OBQA | RTE | WinoGrande | MC-8 |
|---|---|---|---|---|---|---|---|---|---|
| MAN | **61.1** | **84.6** | 88.9 | 59.9 | 72.6 | 33.4 | 80.9 | 74.3 | **69.4** |
| MSAN | 58.3 | 83.4 | **89.2** | 60.3 | 70.4 | 34.4 | 81.6 | 75.0 | 69.1 |
| MoNE | 57.2 | 83.0 | 82.2 | 60.9 | **72.7** | **35.6** | **81.9** | **76.1** | 68.7 |
| SE | 58.1 | 82.7 | 84.6 | 61.3 | 71.9 | 31.8 | 67.5 | 74.5 | 66.6 |
| REAP | 58.0 | 82.3 | 82.2 | 61.3 | 71.7 | 34.4 | 80.5 | 74.7 | 68.2 |
| MAN-sum | 57.1 | 81.9 | 86.8 | 62.2 | 60.7 | 32.2 | 75.8 | 75.2 | 66.5 |
| REAP-sum | 54.4 | 80.6 | 87.0 | 62.5 | 64.9 | 32.6 | 76.2 | 73.7 | 66.5 |
| MEP | 54.0 | 81.0 | 87.7 | **62.8** | 68.0 | 32.6 | 81.6 | 74.1 | 67.7 |

| 25% kept | ARC-c | ARC-e | BoolQ | HellaSwag | MMLU | OBQA | RTE | WinoGrande | MC-8 |
|---|---|---|---|---|---|---|---|---|---|
| MAN | 42.1 | 67.9 | 82.2 | 45.9 | 52.0 | 24.4 | 75.8 | 68.1 | 57.3 |
| MSAN | **42.3** | 68.8 | 82.4 | 46.5 | **52.8** | 24.0 | 73.3 | 68.7 | 57.4 |
| MoNE | 41.0 | **69.9** | 86.0 | 48.6 | 48.9 | **27.0** | 74.7 | 68.3 | **58.0** |
| SE | 39.5 | 64.9 | 84.0 | 47.5 | 46.6 | 25.0 | **76.9** | 68.2 | 56.6 |
| REAP | 39.8 | 69.1 | **86.5** | 48.0 | 48.4 | 26.4 | 75.8 | 66.5 | 57.6 |
| MAN-sum | 30.9 | 56.2 | 84.2 | 50.9 | 32.3 | 22.2 | 76.2 | 72.0 | 53.1 |
| REAP-sum | 31.4 | 55.7 | 83.8 | 50.7 | 31.5 | 23.4 | 72.2 | **72.2** | 52.6 |
| MEP | 33.0 | 60.0 | 82.8 | **51.5** | 32.7 | 26.0 | 72.2 | 71.0 | 53.6 |

| 12.5% kept | ARC-c | ARC-e | BoolQ | HellaSwag | MMLU | OBQA | RTE | WinoGrande | MC-8 |
|---|---|---|---|---|---|---|---|---|---|
| MAN | 27.9 | 49.1 | 63.2 | 36.2 | **32.6** | 16.8 | 57.8 | 58.5 | 42.7 |
| MSAN | **28.2** | 51.9 | 62.9 | 36.0 | 29.4 | 16.6 | **61.7** | 58.6 | 43.2 |
| MoNE | **28.2** | **52.6** | 65.9 | 36.7 | 27.0 | 18.4 | 60.3 | 56.7 | 43.2 |
| SE | 26.4 | 48.4 | **71.6** | 35.4 | 27.7 | 15.2 | 54.5 | 60.2 | 42.4 |
| REAP | 27.4 | 51.6 | 68.0 | 36.4 | 28.5 | **18.6** | 61.0 | 56.7 | **43.5** |
| MAN-sum | 24.7 | 42.0 | 63.6 | 39.9 | 24.5 | 16.6 | 55.6 | 64.3 | 41.4 |
| REAP-sum | 24.0 | 42.0 | 69.3 | 39.9 | 25.2 | **18.6** | 54.2 | **66.0** | 42.4 |
| MEP | 24.1 | 43.6 | 71.1 | **40.3** | 23.7 | 18.0 | 54.9 | 64.4 | 42.5 |

Table 19: Per-task multiple-choice accuracy on Qwen3.6-35B with C4 calibration at three budgets (ARC-c/ARC-e: ARC-challenge/easy; OBQA: OpenBookQA). Bold: best in the column (ties bolded).

**Interactions among removed experts.** Table 22 gives the off-diagonal part of the Gram matrix of the removed set as a fraction of the sum of its deletion energies. It is positive in every setting, so the errors of removed experts add up rather than cancel, more so for random sets than for MEP.

| | OLMoE-1B | | | Qwen3-30B | | | Gemma-4-26B | | | Qwen3.6-35B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Experts kept (%) | 50 | 25 | 12.5 | 50 | 25 | 12.5 | 50 | 25 | 12.5 | 50 | 25 | 12.5 |
| MAN | 1.40 | 1.23 | 1.15 | 11.34 | 2.96 | 2.07 | 4.30 | 2.58 | 1.86 | 6.33 | 2.58 | 1.66 |
| REAP | 1.17 | 1.14 | 1.13 | 6.47 | 2.61 | 2.16 | 4.07 | 2.49 | 1.61 | 2.66 | 2.01 | 1.48 |
| Traffic | 3.67 | 2.28 | 1.79 | 4.89 | 2.28 | 1.61 | 7.63 | 3.97 | 2.12 | 1.84 | 1.51 | 1.28 |
| MAN-sum | 3.14 | 1.92 | 1.65 | 1.33 | 1.50 | 1.30 | 3.87 | 2.05 | 1.90 | 1.39 | 1.24 | 1.21 |
| REAP-sum | 1.15 | 1.15 | 1.61 | 1.18 | 1.36 | 1.32 | 1.84 | 1.62 | 1.30 | 1.17 | 1.13 | 1.11 |
| MEP | 1.03 | 1.02 | 1.07 | 1.02 | 1.07 | 1.13 | 1.00 | 1.00 | 1.02 | 1.00 | 1.00 | 1.01 |

Table 20: Layer reconstruction error $J(S)$ of each criterion's kept set relative to the best kept set found by greedy search with swap refinement on the same C4 Gram matrix ($1.00$: the greedy set).

| Model | $50\%$ | $25\%$ | $12.5\%$ |
|---|---|---|---|
| OLMoE-1B | 97% | 89% | 81% |
| Qwen3-30B | 100% | 97% | 92% |
| Gemma-4-26B | 98% | 92% | 83% |
| Qwen3.6-35B | 100% | 99% | 96% |
| random kept sets (expected) | 50% | 25% | 12.5% |

Table 21: Fraction of kept experts shared by MEP and by selection with the full interaction matrix (C4 calibration set, all layers). The last row is the expected overlap of two random kept sets.

**Benchmark evaluation.** The re-routed variant computes the same statistics under the routing of the pruned model: top-$k$ re-selected among a candidate pool of the $\min(2M, \frac{3}{4}E)$ experts ranked highest by REAP's saliency, gates re-normalized. Tables 23 and 24 report both variants alongside REAP and MEP under the three calibration samples. The two cells where a variant's range over the three samples separates from MEP's are Gemma-4 GSM8K at $25\%$, where the re-routed variant is worse, and Qwen3-30B HumanEval+ at $25\%$, where it is better.

## G Calibration

Table 25 gives the numbers for REAP's mixture quoted in Section 6, and Figure 7 repeats Figure 2 on all three fine-grained models. Figure 8 crosses calibration with evaluation on two axes. The domain sets are C4 (web), evol-codealpaca (code), tulu-3 personas-math (math), WritingPrompts (writing) and xlam-function-calling (agentic), each split into a calibration part and a held-out evaluation part. Each column is normalized by its best calibration, so the diagonal shows that the matched calibration set is best for every evaluation domain and language, and the bottom rows show that an equal-share mixture stays within $2.1$ times of it across domains and within $3.6$ times across languages. Language calibration sets are C4 language variants, evaluated on held-out, document-disjoint continuations of the same sources ($299{,}008$ tokens each). Figure 9 shows perplexity and MC-8 with REAP's mixture on all four models. The advantage of the summed criteria in Table 17 disappears, and at low budgets an averaged criterion or SE is best.

## H Generation and Recovery

Table 26 summarizes how far each criterion can prune under task-matched calibration before it loses $10\%$ or $25\%$ of the original model's score, and Tables 28, 29 and 30 give the values behind it. A $10^5$-token sample of REAP's mixture holds nine to eleven documents, so HumanEval+ at low budgets depends on the calibration sample for every criterion, most strongly on Qwen3-30B at $25\%$ and below (Table 24). The gaps below, taken from the first calibration sample (Tables 28–30), were tested with a paired sign-flip permutation test on the two criteria's per-problem correctness (two-sided, $20{,}000$ resamples). Gemma-4 GSM8K $+39.6$ over MoNE at $25\%$ and $+24.7$ at $12.5\%$, Qwen3-30B GSM8K $+8.5$ over MoNE and $+14.0$ over REAP at $25\%$, Gemma-4 HumanEval+

| | Experts kept (%) | | | | | |
|---|---|---|---|---|---|---|
| | 50 | | 25 | | 12.5 | |
| Model | MEP | random | MEP | random | MEP | random |
| OLMoE-1B | 25% | 27% | 37% | 44% | 50% | 55% |
| Qwen3-30B | 11% | 74% | 49% | 100% | 67% | 130% |
| Gemma-4-26B | 5% | 24% | 9% | 38% | 12% | 36% |
| Qwen3.6-35B | 3% | 10% | 5% | 16% | 9% | 18% |

Table 22: Interaction among removed experts: the off-diagonal Gram mass of the removed set, $\sum_{i \neq j \in P} G_{ij}$, as a percentage of its summed deletion energy $\sum_{j \in P} w_j$, for the kept sets of MEP and for random kept sets (C4 calibration set).

| Model | Kept | REAP | Interaction-aware | Re-routed | MEP |
|---|---|---|---|---|---|
| Gemma-4-26B | 50% | 61.7 (60.6–62.2) | 69.5 (68.0–70.5) | 71.1 (69.5–72.0) | 69.6 (68.4–71.7) |
| | 37.5% | 48.1 (47.1–49.1) | 70.2 (69.6–70.7) | 69.2 (68.6–69.5) | 69.5 (67.8–70.5) |
| | 25% | 15.6 (13.1–17.8) | 59.2 (57.5–62.3) | 46.5 (44.1–48.5) | 60.0 (58.8–61.3) |
| Qwen3-30B | 50% | 86.5 (86.2–86.7) | 86.2 (83.5–87.8) | 88.1 (84.6–89.9) | 88.2 (87.6–89.2) |
| | 37.5% | 84.1 (81.8–85.6) | 80.1 (79.4–80.7) | 78.1 (73.4–81.0) | 78.5 (76.4–80.2) |
| | 25% | 68.5 (65.6–70.4) | 78.7 (78.2–79.0) | 80.5 (79.8–81.4) | 79.3 (77.3–80.9) |
| Qwen3.6-35B | 37.5% | 94.5 (94.3–94.7) | 95.3 (95.0–95.7) | 95.2 (94.8–95.5) | 95.3 (95.2–95.4) |
| | 25% | 88.4 (85.5–89.9) | 92.7 (92.4–93.0) | 93.0 (92.6–93.3) | 92.7 (92.2–93.2) |
| | 12.5% | 24.2 (11.6–34.8) | 55.9 (45.9–64.9) | 54.2 (36.2–72.4) | 42.4 (32.9–50.8) |

Table 23: GSM8K accuracy with the math calibration set: mean over the three calibration samples, with the lowest and highest value in parentheses. Interaction-aware selection minimizes $J(S)$ with the full Gram matrix; "re-routed" is the same selection with the Gram matrix computed under the routing of the pruned model; MEP uses the diagonal of the Gram matrix. Qwen3.6-35B is shown at budgets one step lower because it has twice as many experts.

$+43.3$ over MoNE at $37.5\%$ and Qwen3.6 GSM8K $+16.0$ over REAP at $12.5\%$ returned $P < 10^{-4}$. Qwen3.6 HumanEval+ $+5.5$ and $+9.1$ over MoNE at 25 and $12.5\%$ returned $P = 0.07$ and $0.06$.

**Recovery.** Table 31 distills MoNE, REAP and MEP kept sets on Qwen3-30B with identical training. The probe is the mean accuracy on five multiple-choice tasks (HellaSwag, ARC-easy, ARC-challenge, WinoGrande, MMLU), 300 items each. Every criterion gains 2 to 7 points, and the worse selections gain more. MEP has the highest accuracy before and after distillation in each of the three samples at each budget; MoNE and REAP, within $0.6$ points of each other, change order at $50\%$.

| Model | Kept | REAP | Interaction-aware | Re-routed | MEP |
|---|---|---|---|---|---|
| Gemma-4-26B | 50% | 32.9 (29.9–34.8) | 53.9 (49.4–57.3) | 56.5 (54.3–60.4) | 49.8 (39.0–56.7) |
| | 37.5% | 3.9 (0.0–7.3) | 53.7 (48.8–58.5) | 46.1 (39.6–55.5) | 57.1 (50.6–62.2) |
| | 25% | 0.0 | 39.0 (25.0–48.2) | 17.7 (10.4–29.3) | 35.2 (18.3–45.7) |
| | 12.5% | 0.0 | 7.1 (0.0–11.0) | 0.0 | 2.0 (0.0–4.3) |
| Qwen3-30B | 37.5% | 76.8 (72.0–79.9) | 60.2 (45.7–82.3) | 65.0 (48.2–81.1) | 72.4 (51.8–84.2) |
| | 25% | 28.5 (18.3–45.7) | 27.2 (12.8–48.8) | 38.6 (27.4–55.5) | 15.0 (3.0–21.9) |
| | 12.5% | 0.0 | 4.5 (1.8–7.3) | 3.0 (0.0–4.9) | 4.3 (3.7–4.9) |

Table 24: HumanEval+ pass@1 with the code calibration set: mean over the three calibration samples, with the lowest and highest value in parentheses (a single value: zero in all three samples). Qwen3.6-35B was not evaluated with three samples on this task.

| | | REAP mixture, $10^5$ tokens | | REAP mixture, $2 \times 10^6$ tokens | |
|---|---|---|---|---|---|
| Model | Criterion | 25% kept | 12.5% kept | 25% kept | 12.5% kept |
| Qwen3-30B | REAP | 78.9 | 458.4 | 101.5 | 539.5 |
| | MEP | 179.0 | 1008 | 157.8 | 1885 |
| Gemma-4-26B | REAP | 50.4 | 106.8 | 48.3 | 104.3 |
| | MEP | 95.1 | 169.5 | 80.7 | 158.4 |

Table 25: WikiText perplexity of REAP and MEP calibrated on REAP's mixture with $10^5$ and $2 \times 10^6$ tokens.

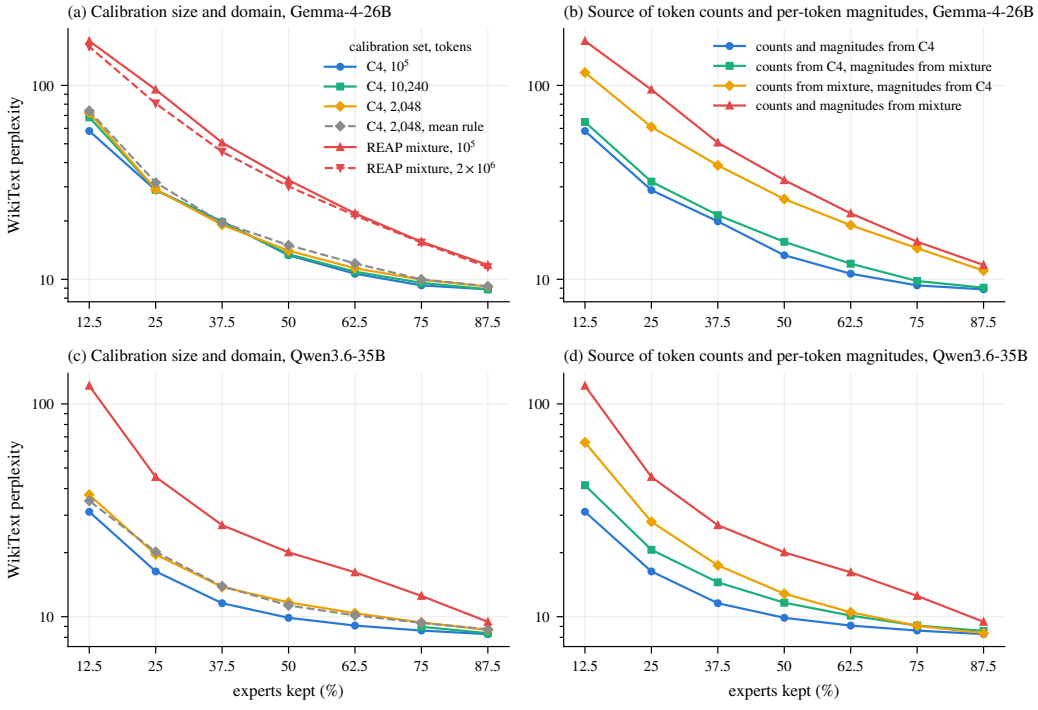


Figure 7: Figure 2 for Gemma-4-26B and Qwen3.6-35B. Left: C4 calibration sets of three sizes, the mean rule on 2,048 tokens (grey, dashed) and REAP's mixture (red; $2 \times 10^6$ tokens not evaluated on Qwen3.6-35B). Right: token counts and per-token magnitudes taken from different calibration sets.

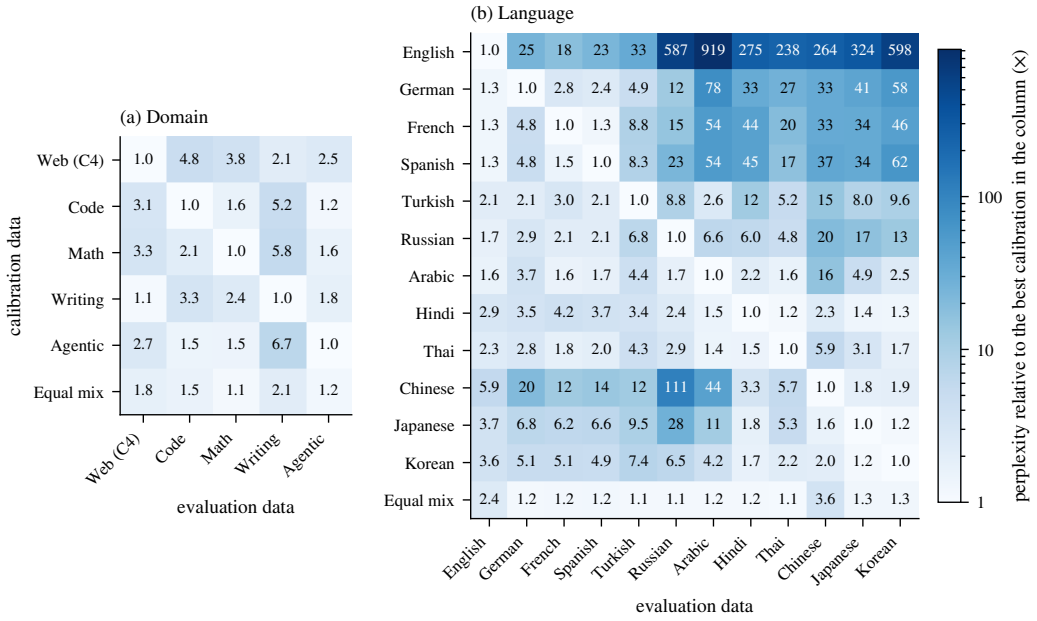


Figure 8: Perplexity of MEP on Qwen3-30B at a 25% budget for each combination of calibration and evaluation data: (a) domains, (b) languages. Each cell is divided by the best value in its column, so the color scale is comparable within columns. The bottom rows use an equal-share mixture of the calibration sets above.

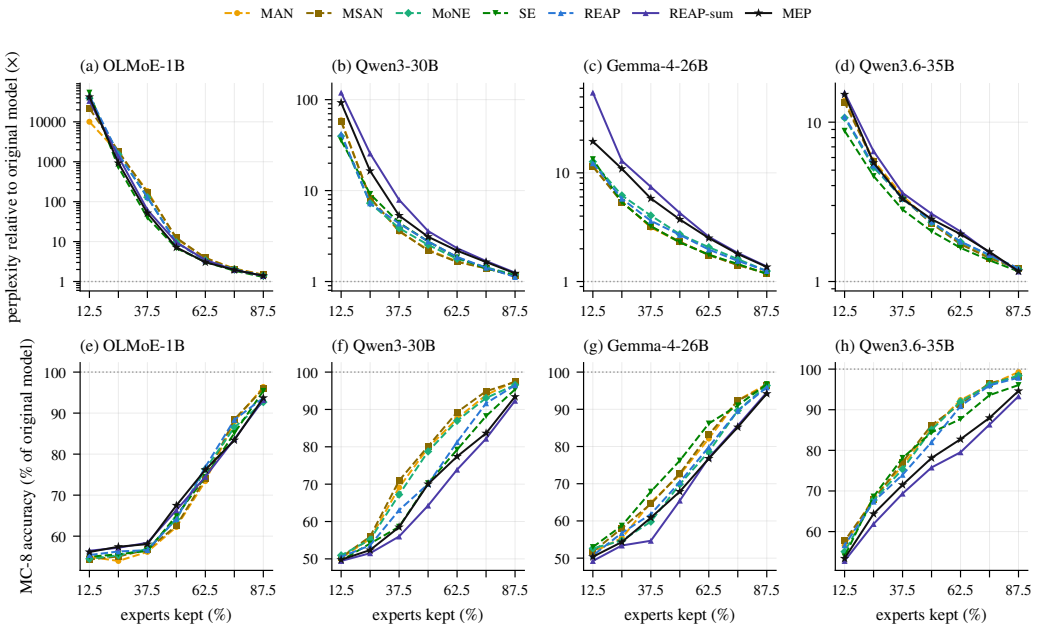


Figure 9: Criteria calibrated on REAP's mixture, evaluated on WikiText (top, perplexity relative to the original model) and MC-8 (bottom, accuracy as a percentage of the original model). Traffic and MAN-sum were not evaluated with this calibration set.

| | GSM8K | | | HumanEval+ | | |
|---|---|---|---|---|---|---|
| Criterion | Qwen3-30B | Gemma-4-26B | Qwen3.6-35B | Qwen3-30B | Gemma-4-26B | Qwen3.6-35B |
| MAN | 50 | 75 | 37.5 | 50 | none | 37.5 |
| MoNE | **37.5** | 62.5 | **25** | **37.5** | 75 | **25** |
| SE | **37.5** | 62.5 | 37.5 | 62.5 | none | 37.5 |
| REAP | 50 | 62.5 | **25** | **37.5** | 75 | **25** |
| REAP-sum | n/a | 50 | n/a | n/a | **50** | n/a |
| MEP | 50 | **37.5** | **25** | **37.5** | **50** | **25** |

Table 26: Smallest budget (% of experts kept) down to which a criterion keeps at least $90\%$ of the original model's score at every budget, on GSM8K (math calibration set) and HumanEval+ (code calibration set), first calibration sample. Bold: smallest budget in the column. "none": the criterion does not reach $90\%$ at any budget; "n/a": not evaluated. The same with a threshold of $75\%$ of the original model's score is in Table 27, and the scores are in Appendix H.

| | GSM8K | | | HumanEval+ | | |
|---|---|---|---|---|---|---|
| Criterion | Qwen3-30B | Gemma-4-26B | Qwen3.6-35B | Qwen3-30B | Gemma-4-26B | Qwen3.6-35B |
| MAN | 37.5 | 62.5 | **25** | 50 | 75 | **25** |
| MoNE | **25** | 50 | **25** | **37.5** | 62.5 | **25** |
| SE | 37.5 | 50 | 37.5 | 50 | 75 | **25** |
| REAP | 37.5 | 50 | **25** | **37.5** | 62.5 | **25** |
| REAP-sum | n/a | 37.5 | n/a | n/a | 50 | n/a |
| MEP | **25** | **25** | **25** | **37.5** | **37.5** | **25** |

Table 27: As Table 26, for a threshold of $75\%$ of the original model's score.

| HumanEval+ pass@1 | Experts kept (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 87.5 | 75 | 62.5 | 50 | 37.5 | 25 | 12.5 |
| Original: 60.4 | | | | | | | |
| 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 |
| REAP-sum | 59.8 | 58.5 | 56.1 | 54.9 | 18.3 | 1.8 | 0.0 |
| MEP | 58.5 | 58.5 | **57.3** | **56.7** | **50.6** | **18.3** | 0.0 |

| GSM8K accuracy | Experts kept (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 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 |
| REAP-sum | 72.2 | 72.7 | **73.4** | **72.0** | 64.4 | 52.8 | **32.4** |
| MEP | **72.6** | 72.3 | 71.3 | 71.7 | **70.5** | **58.8** | 26.5 |

Table 28: Gemma-4-26B: HumanEval+ pass@1 with the code calibration set (top) and GSM8K accuracy with the math calibration set (bottom), first calibration sample. Bold: best in the column (ties bolded; none when all are zero). MSAN was not evaluated on generation.

| HumanEval+ pass@1 | Experts kept (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 87.5 | 75 | 62.5 | 50 | 37.5 | 25 | 12.5 |
| 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 |
| MEP | 89.0 | 87.2 | 86.6 | 87.2 | 81.1 | 21.9 | **4.9** |

| GSM8K accuracy | Experts kept (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 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** |
| MEP | 90.5 | 90.4 | **90.5** | **89.2** | 78.8 | **79.6** | 10.5 |

Table 29: Qwen3-30B: HumanEval+ pass@1 (top) and GSM8K accuracy (bottom), as in Table 28. REAP-sum was evaluated on Gemma-4-26B only.

| HumanEval+ pass@1 | Experts kept (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 87.5 | 75 | 62.5 | 50 | 37.5 | 25 | 12.5 |
| 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 |
| MEP | **91.5** | 90.2 | **91.5** | 92.1 | 89.6 | **87.8** | **51.8** |

| GSM8K accuracy | Experts kept (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 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 |
| MEP | 95.6 | **95.9** | **96.1** | 95.5 | **95.4** | **92.2** | **50.8** |

Table 30: Qwen3.6-35B: HumanEval+ pass@1 (top) and GSM8K accuracy with reasoning disabled (bottom), as in Table 28.

| | Experts kept (%) | | | | | |
|---|---|---|---|---|---|---|
| | 50 | | 37.5 | | 25 | |
| Criterion | before | after | before | after | before | after |
| MoNE | 56.2 | 60.2 | 46.8 | 53.0 | 37.5 | 44.3 |
| REAP | 55.8 | 60.8 | 46.4 | 52.4 | 37.1 | 44.1 |
| MEP | 61.0 | **65.5** | 52.2 | **55.9** | 46.8 | **49.0** |

Table 31: Accuracy on a five-task multiple-choice probe (HellaSwag, ARC-easy, ARC-challenge, WinoGrande, MMLU; 300 items each) on Qwen3-30B before and after 500 steps of recovery distillation on FineWeb, mean over the three C4 calibration samples. Bold: best after distillation. The probe differs from MC-8, so the values before distillation differ from Table 17.