AUDITING EXPERT ARCHITECTURE REGRET IN MIXTURE-OF-EXPERTS LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Mixture-of-Experts models fix their expert architecture before training, while adaptive methods reorganize it during training when an observational statistic fires. Both rest on one unmeasured premise, that the changes such statistics suggest actually improve the model. We measure it. Expert Architecture Regret is the largest validation-loss improvement available from an audited local operation (retire, merge, or a budget-neutral composite), estimated by *paired interventional probes* against a null of no-op pairs. Each probe adapts two branches identically except for the operation. Across 407 probe results spanning two fine-grained families, seven checkpoints, random and heuristic candidates and two data distributions, six candidates cross the band. None survives re-probing on all control streams, where function-preserving placebos score the same. The instrument is not blind, since an amplitude control puts a known defect above the band and the audit finds it. But its resolution turned out to be a protocol choice. The adaptation rate moves the un-mutated model ten to fifty times further than any single expert can. Lowering it tenfold cuts the band from 0.0041 to 0.0025 nats while the control still fires, and 101 further operations give a largest effect of $+0.0003$. A sixteen-stream re-test confirms two of them, near-dead-expert retires, as gains of 1–2 $\times$ 10^{-4} nats that the pooled rule cannot flag. These gains hold on held-out WikiText text and reverse on OLMoE's own training mixture, where the two experts are not near-dead. Destroying a layer's busiest expert costs 0.0018–0.0100, so a removal gain of that size reaches 80% power only where experts are large. We state this scale argument and give a counterexample class to it. A coarse-grained model (JetMoE-8B, 8 experts, top-2) confirms the account. In it, a layer's busiest expert is worth 11–26$\times$ its band, and eight of nine audited edits lose, six of them by 2.5–13$\times$ that band. The ninth, which only a random candidate proposed, retires a busiest expert that is harmful on WikiText. It clears the band on all six streams of the validation slice, vanishes on the held-out test split (-0.0001, 95% CI -0.0004 to $+0.0002$) and reverses on RefinedWeb, JetMoE's main training source (-0.0036, every stream negative). Every positive audited LM edit we find reverses on data its model was trained on. Half a null draw's variance separates repeat runs of one configuration rather than one stream from another. More streams shrink it and $16\times$ the evaluation tokens do not, so buy control streams, not evaluation data. We release the harness, the raw per-run records, and the scripts that recompute the statistics quoted here.

1 INTRODUCTION

Sparse Mixture-of-Experts (MoE) architectures decouple total from active parameters and now underpin many frontier language models. Standard practice fixes the expert architecture before training. A complementary line of work adapts it during training or fine-tuning, adding, removing, growing or reallocating experts when an *observational* statistic (activation count, load, similarity, semantic drift, router-gate importance) crosses a threshold (Guo et al., 2025; Park and Park, 2026; YuanLab.ai, 2026; Lee et al., 2026). Whether the flagged change would improve the model is not measured, and a recent causal audit gives reason for doubt. Across three open MoEs, population-level observational statistics poorly predict the effect of removing an expert (Engmann et al., 2026). The direct measurement is missing: if we perform the operation and let the model adapt, does it improve beyond what the same adaptation achieves without the operation?

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.

A null-calibrated, paired interventional audit protocol (Section 3). The *structural utility* $\mathrm{SU}(o)$ of a candidate operation is the validation-loss difference between two branches adapted for the same τ steps on identical data, one with o, one without. EAR is the maximum SU over the audited set. Because that maximum is biased upward by noise, each EAR claim is judged against an empirical null of un-mutated branch pairs, with a conservative quantile estimator and a per-seed sign rule. Synthetic tasks with a known expert count test the protocol (Section 4).

A negative whose exceptions are marginal (Section 5). Across two fine-grained model families, seven checkpoints, heuristic and random candidates, and two data distributions (Table 6), not one operation robustly exceeds the null (Fig. 5). Retiring near-dead experts is at best *marginally* loss-improving, and only on the probe distribution. Merging the most functionally similar pairs is harmful, and most dangerous where the trigger fires hardest. Budget-neutral composites are neutral at every horizon tested ($\tau \in \{60, 180, 540\}$), even where their children demonstrably differentiate. Nominal crossings arrive at the rate the audit's own noise predicts, and none survives re-probing on all of a cell's control streams, where function-preserving placebos score indistinguishably from the real operations (Table 1).

A positive control and a power curve that explain the negative (Section 6). An amplitude control plants a known defect, and retiring the amplified expert recovers $\mathrm{SU} = +0.031$ nats at gain 16 and $+0.0067$ at gain 8. Bootstrapping the rule on its own per-seed noise puts 80% power at $\mathrm{SU} = 0.0050$ nats with two control streams and 0.0030 with six. Resolution then proved to be a protocol choice. The adaptation rate moves the un-mutated model ten to fifty times further than any single expert can, and lowering it tenfold brings the floor to 0.0028 (Section 6.1). The single-expert dynamic range is 0.0018–0.0100 nats in three layers, so a removal gain of that size reaches 80% power only where experts are large, under an assumption that Section 6 states and shows can fail.

A coarse-grained model (Section 7). JetMoE-8B, whose busiest experts are worth 11–$26\times$ its band, tests the account where the instrument has room to spare. Eight of its nine audited edits lose, six by $\geq 2.5\times$ the band. The ninth, a busiest expert harmful on WikiText that a random candidate found and no usage trigger proposes, is a positive on the validation slice that vanishes on the held-out test split and reverses on RefinedWeb, the model's main training source.

Two lessons for reuse. Usage heuristics identify safe operations only where their statistic was measured. On WikiText a random retire costs $\mathrm{SU} = -0.0015$ nats against $+0.0001$ for a heuristic-selected near-dead one (real support for efficiency-motivated pruning (YuanLab.ai, 2026)). On OLMoE's own mixture both cost -0.0011, and an 8k-token usage estimate from a single code source shares one of its sixteen picks with a 200k-token one. The audit's residual noise is adaptation-trajectory variance rather than loss-estimate sampling error, so $16\times$ the evaluation tokens does not shrink the band as we predicted (Appendix C). Buy control streams, not evaluation data.

2 Related Work

Adaptive MoE architectures. Observational triggers grow and prune the expert pool during training (Guo et al., 2025; Park and Park, 2026; YuanLab.ai, 2026) and under a fixed PEFT budget (Lee et al., 2026). Others grow it on a schedule (Jin et al., 2026), learn it when upcycling dense models (Chen et al., 2025; Liu et al., 2025), or search it before training (Jawahar et al., 2023), with mechanics from expert-merging algorithms (Nguyen et al., 2026). We do not contradict their stated goals. Retiring near-dead experts is loss-free on the probe text, which supports efficiency-motivated pruning (YuanLab.ai, 2026) and the budget-reallocation premise of adapter-level methods (Lee et al., 2026). We instead test their implicit quality premise, that flagged operations help. At convergence they do not, save two retires worth $\sim 10^{-4}$ nats on WikiText that reverse on the training mixture (Section 6.1). Our trajectory audit measures whether they help during training (Section 5). Utility-driven growth outside MoE approximates that quantity rather than measuring it (Appendix C).

Prune, merge, or grow. For removal, post-training expert pruning is standard (Lu et al., 2024), and a one-shot comparison finds pruning beats merging on generative benchmarks (Lasby et al., 2026). Our retire-and-reroute bound reaches the same ordering by direct measurement (Section 5). For merging, MC-SMoE (Li et al., 2024b) requires permutation alignment before averaging, a premise that does not hold on OLMoE (Section 5). For growth, function-preserving expansion dates to Net2Net (Chen et al., 2016), whose symmetry-breaking problem our split faces. Our three child

initialisations (perturbed copy (Komatsuzaki et al., 2023), partial re-initialization (Nakamura et al., 2025), gradient-directed (Evci et al., 2022)) come from that literature. We audit them as initialisation schemes, not as predictions a probe-scale result could refute (Appendix C).

Interventional analyses of MoE experts. A causal audit of expert removal alone, on three open MoEs, finds observational importance measures unreliable (Engmann et al., 2026). A pre-registered test finds that one frontier MoE's apparent expert modularity mostly dissolves under intervention (Salomone et al., 2026). We extend intervention from removal to merge and split, embed it in a paired, null-calibrated estimator, and connect it to the training trajectory.

3 The Audit Protocol

3.1 Structural operations

We audit three operation families at fixed tensor shapes. **Retire**(e) removes expert e from top-k competition with a router-level logit mask and zeroes its output path. **Merge**(e_i, e_j) interpolates parameters into one slot, weighted by usage, and retires the other. Since unaligned interpolation can destroy both functions, we probe an alignment-aware merge and a retire-and-reroute lower bound (Appendix C). A **budget-neutral composite** ($\Delta P{=}0$) retires an idle expert and splits a busy one into the freed slot (Fig. 3). Composites hold parameters fixed by construction, so no conclusion depends on a parameter price.

3.2 Paired structural-utility probes

Let Θ be the weights and $\mathcal{A}$ the expert architecture. For candidate o,

$$\mathrm{SU}(o) \;=\; \mathcal{L}_{\mathrm{val}}\big(\mathrm{Adapt}_\tau(\Theta, \mathcal{A})\big) \;-\; \mathcal{L}_{\mathrm{val}}\big(\mathrm{Adapt}_\tau(o(\Theta), o(\mathcal{A}))\big), \tag{1}$$

where both branches are adapted for τ steps on the identical token stream with identical optimizer hyperparameters (common random numbers; Fig. 4). The difference is control minus mutated, so positive SU means the operation helps. Adapting only the mutated branch would confound its effect with τ extra steps. A null draw is unpaired, so the band is estimated from a wider distribution than the statistic it judges. A matched null moves no verdict (Appendix B). Common random numbers cancel the data draw but not the trajectory. A control run is not reproducible (Appendix C), so a cell shares one run rather than a re-seedable stream. Controls are computed once per (layer, τ, stream), so a cell's candidates are not independent, which is the mechanism behind the control-stream offsets of Section 5. Each candidate is probed on $k{=}2$ streams and the rest go to null draws. Adaptation and evaluation data are disjoint.

We probe expansion at $3\times$ the contraction horizon ($\tau{=}180$ vs. 60) given the measured horizon bias (Fig. 7). Synthetic split utility still rises at our largest horizon, so we also sweep $\tau \in \{60, 180, 540\}$ at LM scale (Section 5) and, at a ten-fold lower rate, $\tau{=}600$ and $\tau{=}1800$ (Section 6.1).

3.3 EAR and its null

$\mathrm{EAR} = \max_{o \in \mathcal{O}} \mathrm{SU}(o)$ over the audited candidate set, where $\mathcal{O}$ excludes the no-op, so the maximum is *signed*. Here $\mathrm{EAR} < 0$ is the informative statement that every audited operation costs loss, so we keep the sign rather than the non-negative $\max(0, \mathrm{EAR})$. As a maximum over noisy estimates of a sampled candidate set, EAR is biased upward by selection and downward by coverage. We therefore calibrate it against an empirical null built from pairs of identical un-mutated branches adapted on different data streams. The band is the $\lfloor 0.9(n{-}1) \rfloor$-th order statistic (zero-indexed) of the sorted $|\mathrm{draw}|$ values. The default linear-interpolated q_{90} is wider (0.0062 vs. 0.0041 at $\tau{=}60$), so this choice can only make the negative harder to obtain. We call an architecture structurally improvable only when EAR exceeds the band and the per-seed signs agree, and we recheck borderline signals at four to six paired seeds. A maximum invites the per-candidate adjudication that manufactures findings, so we apply the rule in both directions over the whole set. The evidence for regret is any excess of upward crossings, not the fate of an individual one. Draws are pooled for the headline band, but layer-local nulls vary by a factor of 30 ($\max |\mathrm{draw}|$ from 0.0010 at layer 9 to 0.0298 at layer 7), so we report each layer against its own null (Fig. 9). The negative holds under both.

Because $\mathrm{SU}(o,s) = \mathcal{L}(\mathrm{ctrl}_s) - \mathcal{L}(\mathrm{treat}_{o,s})$ shares $\mathcal{L}(\mathrm{ctrl}_s)$, a within-cell mean confounds a control-branch offset with a real effect common to that cell, and the null cannot settle which it is, since subtracting one removes the other. An offset shared by three retires and a merge is more plausibly the control branch's, but we cannot exclude that the un-mutated model sits at a fragile point on that stream. Table 1 settles it in two ways. Re-probed on all six streams, the offset does not reappear, and the layer-4 crossers fall to between -0.0007 and $+0.0011$. The function-preserving placebos in that cell, a permutation and a compensated rescaling, score $+0.0004$ and $+0.0003$, indistinguishable from the real operations. Both readings agree that what was measured is adaptation-trajectory noise.

4 Instrument Validation on Synthetic Ground Truth

Validation uses mixture-of-latent-functions tasks with ground truth $M{=}4$ (Appendix A).

EAR tracks provisioning on one of two utilities (Fig. 6). This experiment alone prices parameters: $\mathrm{SU}_{\mathrm{adj}} = \mathrm{SU} - \lambda_p \, \Delta P / P_e$ at $\lambda_p{=}0.004$ (Appendix A). On $\mathrm{SU}_{\mathrm{adj}}$, EAR is -0.0006 ± 0.0003 under-provisioned ($N{=}2$), $+0.0028 \pm 0.0007$ matched but routing-collapsed ($N{=}4$) and $+0.0047 \pm 0.0010$ over-provisioned ($N{=}8$). It is positive exactly under misprovisioning, monotone in over-provisioning. On raw loss the same audits read $+0.0034 \pm 0.0003$, $+0.0013 \pm 0.0012$ and $+0.0009 \pm 0.0010$. The ordering reverses, so the ranking is a property of the adjusted utility, and the $N{=}2$ sign depends on λ_p, negative only above $\lambda_p{=}0.0034$. We never varied 0.004, the structural controller's default price (Appendix A). Raw loss keeps what the LM audit relies on. Contraction is free at $N{=}8$ and costly at $N{=}4$, the near-dead-expert pattern of Section 5. Every LM number in this paper is raw SU.

In the under-provisioned model split utility rises monotonically from -0.004 at $\tau{=}20$ to $+0.020$ at $\tau{=}540$, while a busy retire's cost appears instantly and is absorbed. Short horizons mis-rank expansion against contraction (horizon bias, Fig. 7). With only granularity changed, a coarse expert costs $1.14\times$ an equal-share fine group and no edit improves any of the nine models, though the synthetic does not reproduce the regime in which the floor hides the effect (Appendix C).

5 Auditing Trained MoE Language Models

Setup. We audit OLMoE-1B-7B (Muennighoff et al., 2025) (16 MoE layers, 64 experts, top-8, 5.1T tokens) and Qwen1.5-MoE-A2.7B (Qwen Team, 2024) (24 MoE layers, 60 routed experts plus a gated shared expert, top-4) in bf16 on one GPU, adapting on the training split of WikiText-103 (Merity et al., 2017) and evaluating on 197k of its validation tokens (Appendix B). Heuristic candidates per layer are the four (OLMoE) or three (Qwen) lowest-usage retires, the top two merge pairs by functional similarity (output correlation on routed tokens, weight-space cosine being degenerate here) and 2 budget-neutral composites. Each is probed on two paired seeds, and a cell's remaining streams (four of six at $\tau{=}60$, two of four at $\tau{=}180$) supply its null draws. Three trajectory passes and the layer-0 recheck use four seeds, and every rule below is applied at whatever seed count a cell carries (Table 6; Appendices B and C).

Converged OLMoE. No operation robustly exceeds the pooled band ($|q_{90}|{=}0.0041$ at $\tau{=}60$, 0.0030 at $\tau{=}180$). Retire averages $+0.0001$ ($n{=}64$), merge -0.0012 ($n{=}32$, 9% positive), composite -0.0001 ($n{=}32$). Figure 5 plots every operation against the null. Six candidates in three layers nominally cross. Five also pass the sign rule (a layer-3 composite fails it) and are the same five that exceed their layer's local null, so the 6, the 5 and the 5 are one nested set, not three findings. None survives re-probing on every control stream (Table 1).

We rest on the group-level test. We bootstrap the decision rule on the audit's own per-seed noise at $\delta{=}0$, at each candidate's own seed count (91 of the 398 carry four seeds, where requiring all four positive makes a noise firing nearly $40\times$ rarer). This predicts 4.0 upward firings, and we observe 5, inside the 95% interval 1–8. Four of the five sit in one cell, so they are not five independent events. Dependence cannot change that expectation, as $\mathbb{E}[\sum_o \mathbb{1}_o] = \sum_o \Pr(\mathbb{1}_o)$ depends on the marginals alone. Dependence changes the spread. Holding the empirical marginal fixed and inducing the measured intra-cell rank correlation of 0.17 through a Gaussian copula leaves the expectation at 4.0 and widens the interval to 1–9, which still contains the 5 we observe. The rule fires upward at the rate its own noise produces, so the crossings need no individual explanation. The four in layer 4

(three retires and a merge, $+0.0052$ to $+0.0057$) score $\approx+0.009$ on one of that cell's control streams and ≈0 on the other (Appendix C).

As a bound rather than a failure to reject, the one-sided 95% upper limit on each family's mean SU (clustered by cell, since candidates sharing a control are not independent) is $+0.0008$ for OLMoE retire, -0.0004 for merge and $+0.0006$ for composite, negative for every on-distribution family and for Qwen retire, and $+0.00002$ for Qwen composite. The exception is Qwen merge at $+0.0043$, wide because that family's mean is carried by a single pair (Section 6.1), exactly what clustered standard errors should expose. Each limit bounds a *mean*, not the maximum EAR is. The largest single SU in the first three families is $+0.0057$, $+0.0052$ and $+0.0042$, all three crossings Table 1 shows do not reproduce. The maximum has no such bound. Section 6 says what size of effect we would have detected and re-probing tests the crossings we observed, but neither is a confidence limit on $\max_o \mathrm{SU}(o)$.

Merging hurts, but not because the weights are unaligned. The alignment step that MC-SMoE (Li et al., 2024b) performs before averaging raises the merged pair's mean correlation only to 0.145 and leaves the merge indistinguishable from the naive one ($\mathrm{SU} = -0.00108$ vs -0.00113). Retire-and-reroute is 3–4$\times$ gentler (-0.00039), the ordering REAP (Lasby et al., 2026) reaches for one-shot compression. What costs loss is averaging two different functions at all (Appendix C).

Qwen1.5-MoE and random candidates. A second architecture with a very different load profile gives the same answer. Usage across the audited layers spans 0.0021–0.054, a $25\times$ range against OLMoE's $509\times$. Max EAR is $+0.0005$ at τ=60 against a band of 0.0039 and $+0.0013$ at τ=180 against 0.0022 (a layer-23 composite, $0.6\times$ its own band, the largest Qwen SU in Table 6), with retire mean -0.0002, composite -0.0004 and merge -0.0043. Its most similar pair reaches 0.84 (OLMoE's 0.76), so both models contain near-duplicates by output correlation and merging them hurts anyway. Whatever produces the null is not specific to OLMoE's imbalanced routing. Nor do candidates drawn without regard to usage behave differently. Across four OLMoE layers, 28 random retires, merges and composites give mean -0.0014, max $+0.0005$, 14% positive and none above band, a spot check of a very large space rather than a sweep. They do show the heuristics working in the other direction, picking removals that are safe where a random one is not (Appendix C).

On-distribution replication. WikiText-103 is off OLMoE's training distribution, so an improvement expressible only on the pretraining mixture would be invisible. We repeat the audit on OLMoE-mix-0924, adapting on 8M tokens at its published source shares and evaluating on 197k disjoint ones that are 94% DCLM (Li et al., 2024a), its dominant source (Appendix C). The negative holds and tightens. Across four layers and 32 operations the largest SU is $+0.00002$ against a pooled null of 0.0010. At the lower adaptation rate of Section 6.1 it tightens much further as the band falls to 0.00014 and a candidate's two seeds typically disagree by 7×10^{-5}. The result stops being a null and becomes a sign, with 15 of 16 retires and 8 of 8 merges negative (means -0.0011 and -0.0022), nothing above band and the positive control firing at $\mathrm{SU} = +0.0012$.

The replication also corrects a practical claim. On WikiText, retiring the least-used experts measured loss-free, but on the training mixture it is mildly costly (mean -0.0011, 1 of 16 positive). The reason is that an 8k-token usage estimate, drawn from a single code source, shares only one of its sixteen picks with a 200k-token one. Measured properly, there are no near-dead experts to find, and heuristic and random retires coincide (-0.00111 and -0.00110). Usage heuristics identify safe removals where their statistic was measured, and we have no evidence of it elsewhere (Appendix C).

Horizon and child initialisation. Because splitting's utility rises with adaptation horizon in the synthetic setting, we swept composites at $\tau \in \{60, 180, 540\}$ in two layers. There is no upward trend at LM scale. Layer 0 gives $-0.0012/-0.0012/-0.0014$ and layer 8 $+0.0003/-0.0005/-0.0002$, so the negative is not an artifact of too-short expansion horizons up to $9\times$ the contraction horizon. What bounds it is whether the children ever become different functions. Through 540 steps the router separates them while their outputs stay correlated at 0.994. We re-initialize a quarter and a half of the child's intermediate units from the parent's weight statistics, after Drop-Upcycling (Nakamura et al., 2025). This starts them at output correlation 0.80 and 0.58 rather than 0.999 and leaves them differentiated after adaptation. We borrow the method as an initialisation, not a prediction, since its evidence is 500B tokens of continued pretraining against 2.2M here. The verdict stands, with family means -0.0021 and -0.0003 against -0.0008 for the perturbed copy and none of the six positive on both seeds (Appendix C).

Table 1: Every candidate that nominally crossed the pooled band, re-probed on all of its cell's control streams and evaluated on the audit's validation slice, with function-preserving placebos in the same cells. The five non-placebo rows are the candidates that also pass the sign rule and exceed their layer's local null. The sixth nominal crossing, a layer-3 composite, is excluded by the sign rule and shown at the foot. Both placebos preserve the forward pass exactly. The permutation also commutes with Adam's element-wise update, so its SU is zero in expectation by construction. The rescaling preserves the function but not the optimisation trajectory, and Section 7 measures a cell where its adaptation-phase effect is nonzero, so we read it as an adaptation-speed control rather than a pure noise draw. Here both behave as noise: 20 per-stream values, mean -0.0003, s.d. 0.0026, 1 of 20 outside the τ=60 band and 3 outside each cell's own-horizon band, about what a q_{90} band should give. No crossing reproduces. The layer-0 composite has been measured three times, $+0.0042$ on two seeds, $+0.0009$ on a six-seed recheck and -0.0015 here.

layer	operation	2-stream SU	all-stream SU	streams
4	retire(e_{48})	$+0.0054$	$+0.0005$	6
4	retire(e_{55})	$+0.0057$	$+0.0008$	6
4	retire(e_{17})	$+0.0053$	-0.0007	6
4	merge(e_{21}, e_{33})	$+0.0052$	$+0.0011$	6
4	*placebo* permute(e_{21})	—	$+0.0004$	6
4	*placebo* rescale(e_{55})	—	$+0.0003$	6
0	composite(retire e_{21}, split e_{17})	$+0.0042$	-0.0015	4
0	*placebo* permute(e_{17})	—	-0.0012	4
0	*placebo* rescale(e_{21})	—	-0.0014	4
3	composite(retire e_6, split e_{24})	$+0.0032$	-0.0023	4

Across the training trajectory. Converged models show no large slack, so we ask if they ever had any. With the same protocol, we audit OLMoE's public checkpoints at seven points spanning a $256\times$ range of training tokens (0.4%, 0.8%, 1.2%, 2%, 25%, 66% and 100% of 5.1T), on four layers spanning depth (0, 5, 10 and 15), with seven candidates per (checkpoint, layer) cell. No checkpoint is elevated. The largest EAR in the grid is $+0.0022$, a τ=60 retire, against the trajectory's τ=60 null band of 0.0035, and it shows no trend with training (Fig. 10). Our first pass produced one above-band cell, layer 0 at the 2% checkpoint ($+0.0036$). A four-seed repeat returned $+0.0021$, and the high-power probe of Appendix C reproduced the point estimate but failed the sign-consistency rule, its mean carried entirely by a single adaptation-stream excursion (Appendix C). Two-seed probing at this effect size fires at about the rate its own null predicts.

6 Why the Negative Is Uniform: A Power Analysis

A negative result is only as strong as the instrument's sensitivity, so we measured it directly.

Positive control. Interpolating an expert's weights toward noise cannot create a large effect, since destroying layer 8's busiest expert costs 0.0018–0.0022 nats. Instead we amplify an expert's output projection by a gain g, freeze it so adaptation cannot undo the scaling, and audit retiring it. The audit recovers $+0.0067$ at g=8, where the injected damage is 0.091, and $+0.0003$ at g=4, where detection stops (Table 7). That bracket is too coarse to locate the threshold, and damage is zero-shot while SU is post-adaptation, so sensitivity has to be stated on the SU scale (Appendix C). Bootstrapping the decision rule on its own per-seed noise shifted by a true effect δ (Fig. 11) puts 80% power at $\delta = 0.0050$ with our two control streams and 0.0030 with six.

Effect scale of one expert. Against that floor stands one expert's dynamic range, the cost of replacing a layer's busiest expert with noise (Fig. 8). It is 0.0018–0.0022 nats in layer 8 across two runs, 0.0052 in layer 0 and 0.0100 in layer 15, so the argument is layer-conditional. Where a layer's busiest expert is worth less than the floor, the pooled protocol is underpowered for a retire gain on that scale, and the null there may reflect the instrument rather than the model. Where it is worth more, the retire null cannot be excused that way. The step from destruction damage to retire gain is an assumed scale argument, not a bound, since noise substitution and post-adaptation utility are different interventions. The argument fails for an expert whose trained routing is actively harmful, so that its removal is a gain while noise substitution costs nothing. One occurs in our own data (Appendix C).

6.1 THE FLOOR WAS A PROTOCOL CHOICE

Every probe above adapts with a freshly initialised Adam (Kingma and Ba, 2015) at 10^{-4}, a rate we fixed once and never calibrated. It moves the un-mutated model by 0.095 nats and its control streams disagree by 0.0064. At 10^{-5} the movement is 0.017 and the disagreement 0.0020. On OLMoE's own training mixture the 10^{-5} drift is 0.0003, fifty-eight times less, so the off-distribution setting we chose as the more neutral one is in fact the noisier one (Appendix C).

Re-auditing at 10^{-5} sharpens every number. A candidate's seeds disagree by a median 1.4×10^{-4} and the pooled band falls to 0.0025 over 46 draws, interval $[0.0017, 0.0028]$. The amplitude control detects all four injected gains, down to $\mathrm{SU} = +0.0012$ against its cell's own null ($+0.0059$ against the pooled band), so the probe is quieter and more sensitive, not merely rescaled. That does not close the gap. Bootstrapping the rule against the corrected band puts 80% power at δ=0.0028, against 0.0050 at the original rate (0.0019 against the matched null of Appendix B). Sensitivity improves 1.8-fold, but against layer 8's 0.0018–0.0022 it is still 1.3–$1.6\times$ too coarse, while against layers 0 and 15 it now sits comfortably below. Across 52 operations at that sensitivity the audit's rule flags nothing. Means are -0.0001 for retires (n=32), -0.0021 for merges (n=16, negative in all eight layers) and -0.0009 for composites at τ=540. No candidate exceeds the band upward with consistent signs, and the crossings that occur are all downward merges. A sixteen-stream re-test of two sub-band retires confirms both as gains, $+0.00020$ and $+0.00014$ nats, above a permutation placebo on every stream. They hold on the held-out WikiText test split ($+0.00020$ and $+0.00010$) and reverse on OLMoE's own training mixture (-0.00035 and -0.00091, every stream negative), where the two experts carry eight to ten times more traffic (Table 4). A pooled band misses gains that a quiet layer can resolve, and near-deadness is a property of the probe text. Three cells at matched $\mathrm{lr} \times \tau$ products agree (Appendix C).

The sensitivity gain is the protocol's, not the model's. The pass replicates on Qwen1.5-MoE and OLMoE's training mixture, with bands of 0.0009 and 0.00014 against 0.0039 and 0.0010 at 10^{-4}, positive controls firing at $\mathrm{SU} = +0.0019$ and $+0.0012$, and nothing above band in either (25 on Qwen, 24 on-distribution). The three settings span 101 operations without a single upward crossing. Qwen's most functionally similar merge is $5.7\times$ more damaging at the lower rate (Appendix C).

Scope of the claim. In the two fine-grained models, within our candidate families, at horizons $\tau \leq 540$, the rule, under its pooled per-seed noise, would have detected a structural improvement of $\gtrsim$0.005 nats with 80% probability at the original adaptation rate and $\gtrsim$0.003 at the lower one ($\gtrsim$0.001 in the best-resolved cell), and we found none. Apart from Qwen merge at $+0.0043$, every family *mean* is bounded above at 95% by $+0.0008$ nats. The largest single SU among all 398 candidates is $+0.0057$, a layer-4 retire that returns to $+0.0008$ when re-probed on all six streams (Table 1). The clustered bound covers the average operation, and the power statement says what size of effect we would have detected. 80% power at δ followed by a null result is evidence against an effect that large, but it is not a confidence upper limit on $\max_o \mathrm{SU}(o)$, and we have no simultaneous bound on that maximum. Re-probing every nominal crosser, as we do, tests only the crossings we observed. It cannot reach a true positive that failed to cross by noise, and it is not a substitute for a bound. Against both floors stands the 0.0018–0.0100 that destroying one expert costs. Where it is smallest it sits below both, and we do not separate instrument from model. Where it exceeds them, the audit could have resolved a retire gain of that size and did not.

7 THE GRANULARITY ACCOUNT, AND A COARSE MODEL

Section 6 leaves the retire negative resolution-bounded where a layer's busiest expert is worth less than the floor, and a measurement elsewhere. Whether that holds for these two models or for fine-grained MoEs turns on granularity. Fine-grained design (Dai et al., 2024) gives each expert a small routing share by construction, and balancing losses (Fedus et al., 2022; Zoph et al., 2022) keep it small.

Damage tracks the share of routing corrupted (Fig. 1). Destroying experts jointly drives a layer's corrupted share from 0.03 to 0.75, at a cost close to linear in that share (exponents 1.03, 1.04 and 1.53 in layers 15, 8 and 0, R^2 0.87–0.95). At the 0.125 one expert of a top-2-of-8 model carries (Appendix C), the cost is 0.006, 0.008 and 0.035 nats in layers 0, 8 and 15, two to thirteen times our best floor. At 0.25 it is 0.015, 0.019 and 0.068. The account predicts that coarse-grained MoEs are auditable at this sensitivity, a hypothesis for a coarse model to test, not a measurement of one.

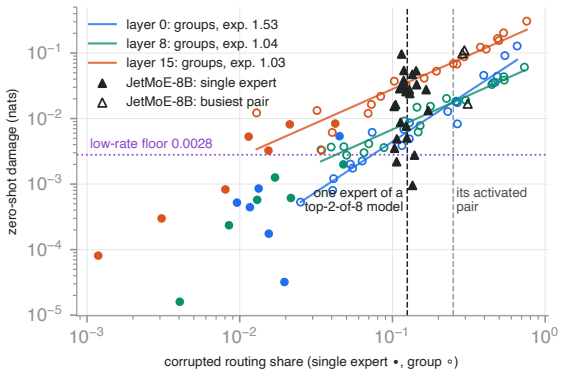


Figure 1: Zero-shot damage against the share of a layer's routing corrupted, log–log, for layers 0, 8 and 15 of converged OLMoE. Filled points destroy one expert, hollow points 2–32 experts jointly (random subsets and the busiest k). Lines are per-layer fits to the group points. Triangles show JetMoE-8B (8 experts, top-2) in layers 0, 12 and 23, single experts filled and the busiest pair hollow. The left dashed line marks the share one expert of a top-2-of-8 model carries ($1/8$, shares being normalised over a layer's experts), the right one the share carried jointly by the pair such a router activates ($1/4$). The dotted line is the low-rate detection floor.

We test the prediction on JetMoE-8B (Shen et al., 2024). It has 24 layers, 8 experts with top-2 routing, 8B parameters by its authors' count (8.5B counted from the released weights) and 2.2B active, and was trained from scratch rather than upcycled, so its experts are not copies of a dense network (Appendix C). We audit layers 12, 23 and 0, the same relative depths as OLMoE's 8, 15 and 0, at τ=60 and 10^{-5}, the rate the frozen amplitude control selects (Appendix C). The null is the model's own, 46 no-op draws with band 0.0012 (bootstrap interval $[0.0008, 0.0018]$). Its routing is balanced, with a layer's busiest expert carrying 0.14–0.17 of the slots against a uniform 0.125. Destroying it costs 0.013, 0.033 and 0.028 nats in layers 12, 23 and 0, or 11, 26 and 22 times the model's own band and 5–12 times the OLMoE floor, all three inside the 0.006–0.035 the grouped ladder predicts for one expert's share (Fig. 1). Read depth for depth, against OLMoE's per-layer fit at the same share, the busiest expert costs 1.1, 0.8 and $2.9\times$ the prediction in layers 12, 23 and 0, and layer 0's busiest pair $4.7\times$, outside that fit's scatter (Appendix C). At this granularity the floor sits far below what a layer's busiest expert is worth. The audit tests the quietest retire, a random retire and the most functionally similar merge per layer, with two paired seeds. It resolves eight of nine candidates as negative, at 1.1–$13\times$ the pooled band (six at $2.5\times$ or more), both seeds agreeing every time (retires -0.0013 to -0.0138, merges -0.003 to -0.017). All but one clear their own layer's local band as well (Appendix C). The ninth is a positive, and the only one in this paper that the audit's rule flags and that then reproduces. Retiring expert 4 of layer 0 scores SU $= +0.0031$ on the validation slice, $2.5\times$ the pooled band (seeds $+0.0030$ and $+0.0031$, all six control streams positive on re-probe). It is the layer's busiest expert, share 0.166, drawn by the random candidate and never proposable by a usage heuristic. Masking it before any adaptation lowers the loss by 0.0051 nats while destroying it with noise costs $+0.028$, so this is a competent expert whose routing on this text is harmful. On the WikiText test split (Table 5) masking still helps zero-shot (-0.0016), but after adaptation the gain is gone, at SU $= -0.0001$ (paired t 95% interval -0.0004 to $+0.0002$, three of six streams positive). On RefinedWeb (Penedo et al., 2023), the public source that dominates JetMoE's training mixture and where the expert's share falls to 0.123, masking it costs $+0.0036$ zero-shot and the paired probe gives -0.0036 on every stream, $14\times$ that cell's band (an independent second slice gives -0.0037), as the earlier OLMoE-mix proxy did (-0.0018). The function-preserving rescaling that matched the retire on the validation slice in the first re-probe ($+0.0030$) is not a placebo but a dose-dependent adaptation-speed effect. Adam's update is invariant to gradient scale, not parameter scale, so multiplying the up rows by c and the down projection by $1/c$ lets that projection move $2.8\times$ and $6.9\times$ further relative to its size at c=2 and 4. The rescaling is worth $+0.0009$, $+0.0028$ and $+0.0126$ at c=$\frac{1}{2}$, 2 and 4 on this expert, $+0.0009$ and $+0.0048$ at c=2 and 4 on a layer-0 expert whose removal costs 0.008, and nothing on RefinedWeb at c=2, the one factor tested there. It leaves the expert's routing share and output size unchanged. The match at c=2 was a coincidence of magnitude,

so the rescaling cannot say whether the gain needs removal. The held-out and training-data tests show instead that it does not transfer. Under Eq. (1) the coarse model's one positive is specific to the text it was measured on. Our controls are of three kinds: run noise, symmetry, and adaptation speed. Only the first two measure noise.

8 Discussion, Limitations, and What Would Change the Answer

Implications. (1) For adaptive-MoE triggers, at convergence and under layer-local adaptation, the operations observational statistics propose are near-neutral at best while remaining budget-positive for pruning. Quality claims for trigger-based restructuring should be measured during training or under shift, not assumed from post-hoc symptoms. (2) For architecture search, converged fine-grained basins show no audited gain above our floor, and the two we re-tested are gains of $\sim 10^{-4}$ nats on WikiText that reverse on the training mixture (Table 4). (3) For audit methodology, paired multi-stream controls with explicit nulls and sign rules are, in our data, the difference between a correct negative and six apparent findings (Appendix C).

Limitations. Our main probes use WikiText-103 and the on-distribution replication covers four layers of one model (Section 5), so distribution-specific slack remains possible. JetMoE's layer-0 retire reverses sign between distributions (Section 7). Usage comes from an 8k-token sample, too small on a heterogeneous mixture and blind to the sources it omits. Adaptation is layer-local ($\tau \leq 540$ at 10^{-4}, $\tau \leq 1800$ at 10^{-5}) and candidates are single- and pair-operations, so cross-layer regret, co-adaptation and multi-expert restructurings go untested. Sensitivity is bounded by Section 6, whose positive control covers the retire channel only. Three limits belong to the operations rather than the models (Appendix C). "Merging hurts" concerns our usage-weighted interpolation of weight-unaligned experts, not merge slack itself. Split children barely differentiate, so composite neutrality bounds perturbed-copy, partially re-initialized and gradient-directed expansion (GradMax (Evci et al., 2022)) at $\tau \leq 540$, leaving a learned or router-aware initialisation untried. Horizons past 540 are probed only at the ten-fold lower rate (Section 6.1), τ=60 bands read longer-horizon points, and re-probing resolves the control-stream offset for six candidates only (Table 1). We audit three families at ≤ 15B total parameters, two fine-grained and one coarse, so larger MoEs may differ. The coarse model's on-distribution probes cover RefinedWeb, its main source, but not its code share (Appendix C). Synthetic tasks validate the instrument, not LM-scale benefits of structural adaptation.

What would change the answer. An amplitude-based positive control at a gain the audit detects here ($g \geq 8$) that it then fails to detect (Appendix C), more upward firings than the audit's own noise predicts (Section 5), a merge rule beyond alignment scoring positive (Appendix C), any positive that holds on its model's own training distribution (Section 7), on-distribution audits of more layers or models than the four we replicate, or regret before 20B tokens, where our trajectory begins.

9 Conclusion

We turned "should this MoE restructure?" from assumption into measurement with stated power and scope. For the two fine-grained families it should not, save two retires worth $\sim 10^{-4}$ nats on WikiText that reverse on the training mixture (Table 4). Whether the retire negative is model or instrument depends on the layer. One expert's dynamic range exceeds Section 6.1's lowered floor in two of the three measured, so a retire gain of that size would likely have shown and did not. Where it is smaller, as in layer 8, we do not separate the two. Merging needs no scale argument, since low-rate merge means are negative in every layer. For composites the statement is weaker: means are zero-to-negative and no gain was detected, but with four low-rate candidates, sparse long-horizon nulls and a growing null tail we do not claim to have excluded a small one. A coarse-grained model confirms the granularity account. Its audited edits are losses save one, a gain on WikiText's validation slice from retiring a busiest expert harmful there, which only a random candidate proposed. It vanishes on the held-out test split and reverses on RefinedWeb, the model's main training source (Section 7). Every positive audited LM edit in this paper reverses on data its model was trained on (OLMoE's mixture, RefinedWeb for JetMoE). Structural adaptation may still pay during training or under shift. That case should be made there, with instruments whose sensitivity is stated.

REPRODUCIBILITY STATEMENT

Every number in this paper is computed from released per-seed measurements rather than from run summaries. The audit harness, the operations, all experiment drivers, the per-seed result records and the figure scripts are in the anonymous supplementary archive attached to this submission, and its `START_HERE.md` names the entry points. `verify_results.py` recomputes the probe count, the pooled null band, the low-rate band in all three settings, the family means of the converged, Qwen and on-distribution audits, the amplitude positive control, every row of Table 1, the coarse model's audit, re-probes and paired-difference interval, and the synthetic audit of Section 4, and prints each beside the value quoted here. `group_test_and_bounds.py` adds the group-level test and the cluster-robust upper limit on each family's mean. `verify_s21.py` and `verify_s22.py` re-check and recompute the re-test of two low-rate retires and its held-out and training-mixture runs, and `verify_s23.py` and `verify_s24.py` the coarse model's held-out, rescaling and RefinedWeb tests. All of them read the per-seed records on CPU and finish in seconds, without a GPU and without re-running any probe. Re-*measuring* does need GPUs: each adaptation branch is 60–1800 optimisation steps, and the full audit is on the order of one GPU-day. Appendices B and C give every protocol, including the integrity check whose ordering bug is described in Appendix C and corrected in the released `s17_jetmoe_audit.py`.

AI USE STATEMENT

In this work, we used generative AI tools as assistants in the following tasks for which disclosure is required: formulating the mathematical claims, proposing and refining hypotheses, designing the experiments, implementing the methods, generating the synthetic datasets, and interpreting results. They also helped clean and reformat the data. We did not use them to develop the conceptual framework. Proofs, translation and qualitative data analysis are not applicable to this work. We also used them to help with identifying related work, writing code, creating figures and polishing the text. We have reviewed all AI-assisted work and recomputed every reported number from the released per-seed records. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI. No LLM is an author.

ETHICS STATEMENT

This paper measures three publicly released model checkpoints, OLMoE-1B-7B, Qwen1.5-MoE-A2.7B and JetMoE-8B, on public corpora: WikiText-103 (Merity et al., 2017) and samples of OLMoE-mix-0924, whose largest source is DCLM (Li et al., 2024a), and of RefinedWeb (Penedo et al., 2023). Each is used under its published licence: Apache-2.0 for OLMoE and JetMoE, the Tongyi Qianwen licence for Qwen1.5-MoE, CC BY-SA 3.0 for WikiText-103, ODC-By for OLMoE-mix-0924 and RefinedWeb, and CC BY 4.0 for DCLM. The work involves no human subjects and collects no personal data. The corpora are public web and encyclopedia text, however, and can contain personal information as it was published online. We redistribute no model weights. What we release is the audit harness, the per-seed measurement records, the scripts that recompute the reported numbers and a RefinedWeb sample (tokens, 77 source documents used to check tokenization, and their URLs), which the archive ships under RefinedWeb's ODC-By licence rather than its own MIT licence. No language model was pretrained or fine-tuned to convergence here; only the synthetic tasks of Appendix A train models of their own, and Table 6 lists every audited setting with the number of probes it cost.

One of our findings could do harm if it is read out of context. We report that retiring near-dead experts is loss-free and say explicitly that this *supports* efficiency-motivated pruning (Section 2); it is not licence to prune experts freely. The same measurement repeated on the training mixture rather than on the probe distribution turns the least-used-expert retire from loss-free into mildly costly, because an 8k-token usage estimate drawn from a single code source shares only one of its sixteen picks with a 200k-token, mostly web one (Section 5). Pruning on a usage statistic estimated off distribution is therefore not something our results license, and we report no evidence that it is safe. Our use of generative AI tools is disclosed separately above.

REFERENCES

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems (MLSys)*, 3:747–769, 2021. arXiv:2103.03098.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating learning via knowledge transfer. *ICLR*, 2016. arXiv:1511.05641.

Shengzhuang Chen, Ying Wei, and Jonathan Richard Schwarz. Automatic expert discovery in LLM upcycling via sparse interpolated Mixture-of-Experts. *ACL*, pages 16703–16717, 2025.

Damai Dai, Chengqi Deng, Chenggang Zhao, et al. DeepSeekMoE: Towards ultimate expert specialization in Mixture-of-Experts language models. *ACL*, pages 1280–1297, 2024. arXiv:2401.06066.

Leonard Engmann, Christian Medeiros Adriano, and Holger Giese. From observation to intervention: A causal audit of expert importance in Mixture-of-Experts models. arXiv:2606.10703, 2026. ICML 2026 PhilML workshop, non-archival.

Utku Evci, Bart van Merriënboer, Thomas Unterthiner, Max Vladymyrov, and Fabian Pedregosa. GradMax: Growing neural networks using gradient information. *ICLR*, 2022. arXiv:2201.05125.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120):1–39, 2022. arXiv:2101.03961.

Yongxin Guo, Zhenglin Cheng, Xiaoying Tang, Zhaopeng Tu, and Tao Lin. Dynamic Mixture of Experts: An auto-tuning approach for efficient transformer models. *ICLR*, 2025. arXiv:2405.14297.

Ganesh Jawahar, Subhabrata Mukherjee, Xiaodong Liu, Young Jin Kim, Muhammad Abdul-Mageed, Laks V.S. Lakshmanan, Ahmed Hassan Awadallah, Sebastien Bubeck, and Jianfeng Gao. AutoMoE: Heterogeneous Mixture-of-Experts with adaptive computation for efficient neural machine translation. *Findings of ACL*, pages 9116–9132, 2023.

Linghao Jin, Chufan Shi, Huijuan Wang, Nuan Wen, Zhengzhong Liu, Eric Xing, and Xuezhe Ma. EMO: Frustratingly easy progressive training of extendable MoE. arXiv:2605.13247, 2026.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. arXiv:1412.6980.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training Mixture-of-Experts from dense checkpoints. *ICLR*, 2023. arXiv:2212.05055.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. REAP the experts: Why pruning prevails for one-shot MoE compression. *ICLR*, 2026. arXiv:2510.13999.

Ahin Lee, Sehyun Yun, and Taesik Gong. EPnG: Adaptive expert prune-and-grow for parameter-efficient MoE fine-tuning. *MobiSys Workshops*, pages 93–98, 2026. arXiv:2607.01789.

Jeffrey Li, Alex Fang, Georgios Smyrnis, et al. DataComp-LM: In search of the next generation of training sets for language models. *NeurIPS Datasets and Benchmarks Track*, pages 14200–14282, 2024. arXiv:2406.11794.

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. *ICLR*, 2024. arXiv:2310.01334.

Qiang Liu, Lemeng Wu, and Dilin Wang. Splitting steepest descent for growing neural architectures. *NeurIPS*, 2019. arXiv:1910.02366.

Peiyu Liu, Tianwen Wei, Bo Zhu, Wayne Xin Zhao, and Shuicheng Yan. Masks can be learned as an alternative to experts. *ACL*, pages 15800–15811, 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. *ACL*, pages 6159–6172, 2024. arXiv:2402.14800.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *ICLR*, 2017. arXiv:1609.07843.

Rupert Mitchell, Robin Menzenbach, Kristian Kersting, and Martin Mundt. Self-expanding neural networks. arXiv:2307.04526v3, 2024.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, et al. OLMoE: Open Mixture-of-Experts language models. *ICLR*, 2025 (oral). arXiv:2409.02060.

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. Drop-Upcycling: Training sparse Mixture of Experts with partial re-initialization. *ICLR*, 2025. arXiv:2502.19261.

Dung V. Nguyen, Anh T. Nguyen, Minh H. Nguyen, Luc Q. Nguyen, Shiqi Jiang, Ethan Fetaya, Linh Duy Tran, Gal Chechik, and Tan M. Nguyen. Expert merging in sparse Mixture of Experts with Nash bargaining. *ICLR*, 2026. arXiv:2510.16138.

Sumin Park and Noseong Park. How many experts are enough? Towards optimal semantic specialization for Mixture-of-Experts. *AAAI*, 40:24792–24800, 2026. arXiv:2512.19765.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. *NeurIPS Datasets and Benchmarks Track*, pages 79155–79172, 2023. arXiv:2306.01116.

Qwen Team. Qwen1.5-MoE: Matching 7B model performance with 1/3 activated parameters. Blog post, 28 March 2024. `https://qwenlm.github.io/blog/qwen-moe/`.

Tony Salomone, Deep Gandhi, and Ali Asaria. How modular is a frontier Mixture-of-Experts? A pre-registered causal test in which apparent expert modularity mostly dissolves. arXiv:2606.25092, 2026.

Yikang Shen, Zhen Guo, Tianle Cai, and Zengyi Qin. JetMoE: Reaching Llama2 performance with 0.1M dollars. arXiv:2404.07413, 2024.

Steve R. Waterhouse and Anthony J. Robinson. Pruning and growing hierarchical mixtures of experts. *Fourth International Conference on Artificial Neural Networks (IEE)*, pages 341–346, 1995.

Lemeng Wu, Bo Liu, Peter Stone, and Qiang Liu. Firefly neural architecture descent: a general approach for growing neural networks. *NeurIPS*, pages 22373–22383, 2020. arXiv:2102.08574.

YuanLab.ai. Yuan3.0 Ultra: A trillion-parameter enterprise-oriented MoE LLM. arXiv:2601.14327, 2026.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. arXiv:2202.08906, 2022.

A Synthetic Experiments

Task. There are M latent functions f_z (random two-layer MLPs). Inputs $x \sim \mathcal{N}(0, I_8)$ are shared across z and concatenated with a noisy cue embedding of z (16-d), and the targets are $y = f_z(x)$. Capacity-limited experts (width 16) cannot fit multiple functions, so correct structure genuinely matters. The model is a top-1 MoE with per-expert router rows and Switch-style auxiliary balancing (α=0.01 unless stated).

Audit validation (Fig. 6). Fixed models were trained 6k steps to convergence, with three seeds per $N \in \{2, 4, 8\}$. Every retire, merge and split candidate was probed at τ=80 with two probe seeds, and each model has four no-op null draws. Figure 6 reports two utilities side by side, the raw SU of Eq. (1) and a parameter-adjusted one, $\mathrm{SU}_{\mathrm{adj}} = \mathrm{SU}_{\mathrm{raw}} - \lambda_p \, \Delta P / P_e$, where ΔP is the operation's change in parameter count, P_e the parameter count of one expert, and λ_p=0.004 the price of one expert's worth of parameters. A retire or merge is thus credited $+0.004$ and a split charged 0.004, while the no-op nulls (ΔP=0) are unchanged. The positive means at N=4 and N=8 are this credit net of a raw removal cost near zero. The raw utility of the best retire or merge is -0.0005 to -0.0019 at N=4 and $+0.0001$ to $+0.0019$ at N=8, against raw $\max|\mathrm{null}|$ of 0.0008–0.0014 and 0.0006–0.0020. At these two sizes the instrument therefore certifies an expert that can be removed for free, which is what over-provisioning means at this price. Because $\Delta P / P_e = \pm 1$ for every audited operation, the price moves each split down and each retire or merge up by exactly λ_p, so the adjusted pattern is conditional on it. The N=2 mean is negative only for $\lambda_p > 0.0034$ (all three seeds only above 0.0037). At 0.004 its $|\mathrm{EAR}| \leq 0.0009$ lies inside the local null (0.0013–0.0022) in every seed, so it is a sign and not a resolved negative. The best N=4 and N=8 removals clear their local nulls in every seed only above λ_p=0.0029 and 0.0019. The value λ_p=0.004 is the structural controller's default price and is used unchanged in every synthetic run that prices parameters. Raw readings are maxima over every candidate. The audit's record kept the six best by adjusted utility, and a re-run

that keeps every row (`exp0b_multiseed_full.py`) reproduces those rows, the nulls and the adjusted EAR exactly. Results are in the main text. The best operation in each seed is a split at $N{=}2$ (3/3 seeds) and a retire or merge of a redundant expert at $N{=}4$ and $N{=}8$ (6/6).

Granularity (Section 4). On the $K{=}8$ task the fine model has $E{=}64$, top-8 routing and width 16. The coarse models have $E{=}8$ with top-1 and width 128, and $E{=}8$ with top-2 and width 64. Routing weights are the softmax over all experts taken at the top-k indices without renormalisation, for every k, so the granularities differ only in E and k, and setting $k{=}1$ reproduces the original model bit for bit. Each model is trained with three seeds for 6k steps on the same data stream. For each model we measure the zero-shot damage from replacing each expert's MLP with scale-matched noise, with routing intact, and in the fine model also from six random 8-subsets and the eight busiest. Each model also gets ten no-op null draws at $\tau{=}80$, and retire (idlest and random), split (busiest) and merge (most similar pair) are probed with two seeds. An edit counts as resolved when $|\mathrm{SU}|$ is above the model's own band and both seeds have one sign. Top-1 routing at $E{=}K$ is fragile under the balancing loss, and two of the three top-1 coarse seeds converged with one or two dead experts. This is why that configuration resolves 6 of 12 edits against the top-2 model's 11 of 12.

Structure recovery. A probe-driven structural controller recovers ground truth from both directions. It draws candidates from usage and similarity heuristics and decides by paired probes with hysteresis, cooldown and a parameter price. From $N_0{=}1$ it grows to exactly $N{=}4$ (NMI 1.0 against latent labels) in five of five seeds, with zero variance in the final count. The consolidating direction was run once, with seed 0. From $N_0{=}16$ the controller executes twelve retires, all by step 3,250 of 9,000, settles at exactly $N{=}4$ (NMI 1.0) and accepts no further edit.

Decision-signal ablation. Operation machinery and candidate generation are identical across conditions, and only the decision rule differs. Probe-based decisions are the only condition that recovers true structure with zero variance, both on a static task (final $N{=}4$ in 5/5 seeds, validation loss 0.169 ± 0.010) and on a distribution-shift task whose true complexity moves from 2 to 5 to 3 (final $N{=}3$ in 5/5 seeds, Fig. 2). Random decisions give unstable, bloated structures ($N = 5.6 \pm 1.8$, loss 0.221 ± 0.022), and threshold heuristics stall ($N{=}2$ in 5/5 seeds, loss 0.290 ± 0.013). Grow-only ablations cannot contract, going from 2 to 5 and staying there.

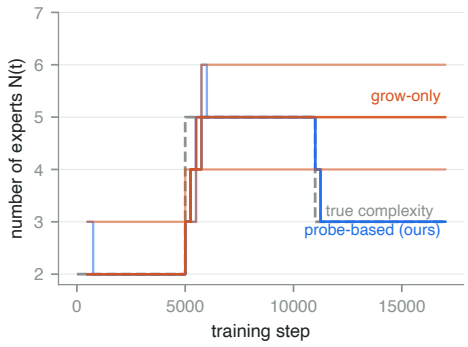


Figure 2: Synthetic distribution-shift task. Probe-based structural decisions track the true complexity as it moves from 2 to 5 to 3 across 5 seeds, and grow-only controls cannot contract. In this setting the instrument does find and act on structural slack.

Load-balancing controls. On uniform mixtures, strong balancing ($\alpha{=}0.1$) prevents the routing collapse that produces slack in fixed models. There, forced-uniform load coincides with correct specialization, and an "evolved beats initialized" contrast disappears under this control. On skewed mixtures ($P(z) = (0.55, 0.25, 0.12, 0.08)$), no balancing coefficient yields structure that is both aligned with ground truth and reliable across seeds. Strong balancing reaches NMI 0.769 ± 0.004, and weak balancing is a seed lottery, with NMI 0.965 ± 0.050 and a validation-loss range of 0.095–0.178. Probe-driven structural adaptation does yield such structure (NMI 0.9994 ± 0.0013, validation loss 0.139 ± 0.002).

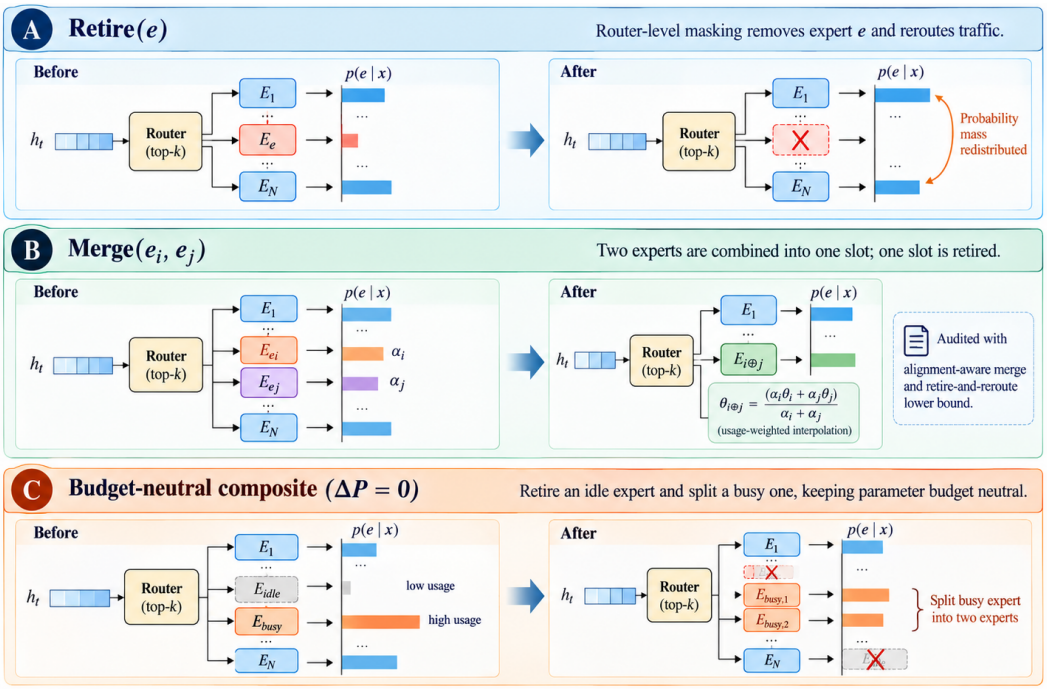


Figure 3: The three audited operations (Section 3), drawn schematically on one MoE layer. The figure shows no measured data. (A) Retire removes the expert from top-k competition, so its routing probability is redistributed over the others. (B) Merge interpolates two experts into one slot, weighted by usage, and retires the other. An alignment-aware merge and a retire-and-reroute bound are audited as well (Appendix C). (C) The budget-neutral composite retires an idle expert and splits a busy one into the freed slot, so the parameter count is unchanged. Every operation edits a single layer at fixed tensor shapes.

Noise null. Each audited model carries its own no-op nulls at the same horizon as its candidates, τ=80. Their $|\text{draw}|$ reaches 0.0008–0.0014 at N=4 and 0.0006–0.0020 at N=8, below every positive EAR measured there (0.0022–0.0035 and 0.0041–0.0059 respectively).

B Protocol Details

Terminology. Table 2 collects the terms the audit uses in a fixed sense, and Figs. 3 and 4 draw the operations and the paired probe.

Operations at fixed shapes. Retire: additive -10^4 mask on the router logit (installed by replacing the router forward with a masked equivalent), gate row and expert output slice zeroed. Merge: usage-weighted interpolation of fused expert tensors and gate rows into slot i; slot j retired. Composite split: victim slot overwritten with a perturbed copy of the parent (σ equal to $0.02\times$ the parent's weight standard deviation), gate rows pushed apart symmetrically by $0.15\,\|g\|$; victim's mask cleared.

Adaptation. We adapt with Adam at lr 10^{-4}. The trainable set is the audited layer's MoE block (the fused expert tensors, the router and, for Qwen, the shared expert). Both branches see identical seeded streams, with $\tau \in \{60, 180, 540\}$ depending on the operation family. Evaluation uses 48 batches of 4×1024 tokens (197k) from the validation split, disjoint from the training-split adaptation data. The usage and similarity statistics that choose heuristic candidates are computed on the first one or two of those batches. The coarse model uses twelve, and the on-distribution 200k re-estimate uses fifty for usage and eight for similarity. There is no separate held-out reporting split. Outside the $16\times$ probes and the held-out re-test of Table 4, every heuristic candidate's SU in Sections 5 and 6.1 is measured on the text it was chosen on (random candidates come from a seeded generator). That selection keys on usage and similarity rather than on SU, but it is the same text. Text a candidate was not chosen on enters in three places. The $16\times$ probes evaluate on 3.15M tokens, over 90% of

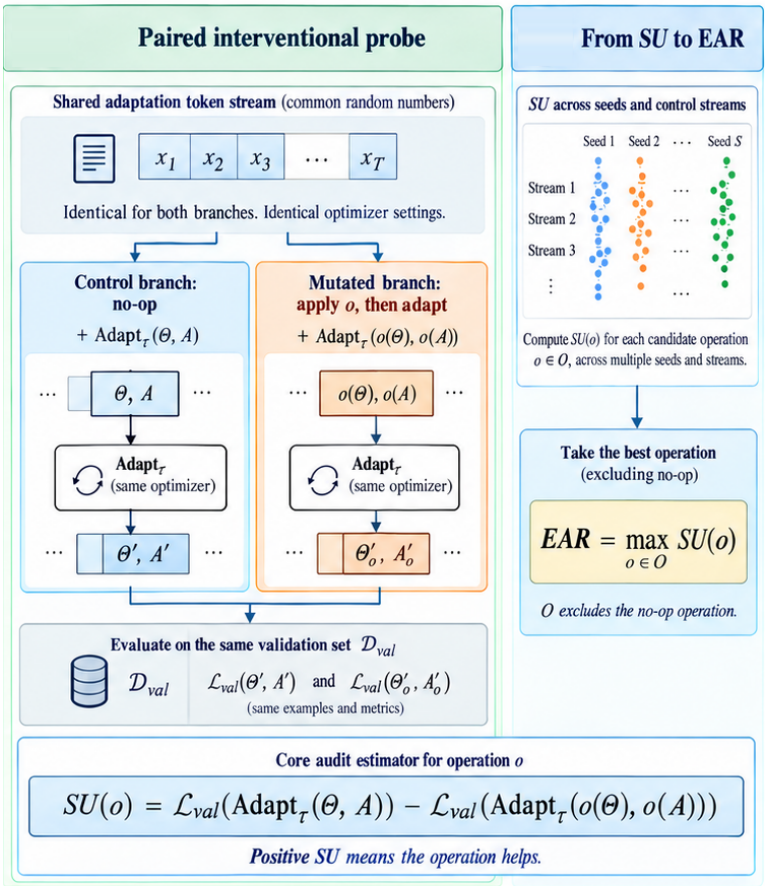


Figure 4: The paired structural-utility probe (Section 3.2) and the audit statistic (Section 3.3), drawn schematically. The figure shows no measured data. Control and mutated branch adapt for τ steps on the identical token stream with identical optimizer settings and differ only by the operation, and Eq. (1) is the difference of their validation losses. EAR is the maximum over the audited candidates, excluding the no-op. It is read only against the empirical null of Section 3.3, which pairs un-mutated branches on different streams, together with the per-seed sign rule. In the scatter a seed indexes a control stream (Table 2).

them from training-split documents disjoint from the adaptation slice (Appendix C). The held-out and training-mixture runs of Table 4 use such text, and so do the coarse model's test-split, OLMoE-mix and RefinedWeb probes (Section 7, Table 5).

Null construction. A null draw is $\mathcal{L}(\text{Adapt}_\tau^{(s_a)}(\Theta)) - \mathcal{L}(\text{Adapt}_\tau^{(s_b)}(\Theta))$ for distinct streams $s_a \neq s_b$ on un-mutated copies. This is the same estimator as SU with the operation removed and the pairing broken, and so an upper bound on paired-probe noise. Draws are constructed from disjoint stream pairs within a cell (streams $2{-}3$ and $4{-}5$ at $\tau{=}60$, $2{-}3$ at $\tau{=}180$), so no stream appears in more than one draw and the draws within a cell are independent. The pooled sign balance is 63 positive to 66 negative over 129 draws. Draws are pooled across layers for the headline band and reported per layer in Fig. 9. The band is the conservative lower q_{90} order statistic (Section 3.3). A null draw differences two unpaired branches while SU differences a paired one, and the rule thresholds a two-seed mean of such differences, so the null is not the estimator's own sampling distribution. It is wider, which makes the band conservative, but it is mis-specified, so we measure the gap directly. At 10^{-4}, the Var of a within-candidate seed difference is $0.20\times$ that of a null draw. That ratio pools sixteen layers against a null the early layers dominate, so we also build the *matched* null directly, at no extra compute, from the 36 identical re-runs above. A matched draw differences $\mathcal{L}(\text{Adapt}_\tau^{(s)}(\Theta))$ between two independent executions of the same (layer, s, τ, rate). It is the estimator with the operation removed and the pairing kept. Over layers $\{0, 2, 4, 6, 8, 12\}$ at 10^{-5} these draws have s.d.

Table 2: Terms used in a fixed sense throughout the audit.

term	meaning
probe; control and mutated branch	A paired probe adapts two copies of the model for τ steps on the identical token stream: the *control branch* (un-mutated) and the *mutated branch* (operation o applied); $\mathrm{SU}(o) = \mathcal{L}(\mathrm{ctrl}) - \mathcal{L}(\mathrm{mutated})$, positive when the operation helps (Section 3.2).
control stream	The seeded token stream a probe's two branches share, indexed by seed s, and by extension the one control run adapted on it; each candidate is probed on two such streams (four or six in the rechecks and the three earliest trajectory passes), and a cell's remaining streams are spent on null draws (Section 3.2, Section 5).
null draw	$\mathcal{L}(\mathrm{Adapt}_\tau^{(s_a)}(\Theta)) - \mathcal{L}(\mathrm{Adapt}_\tau^{(s_b)}(\Theta))$ for two distinct streams on un-mutated copies: the SU estimator with the operation removed and the pairing broken, taken over disjoint stream pairs within a cell (below).
band	The noise threshold: the $\lfloor 0.9(n-1) \rfloor$-th order statistic (zero-indexed) of the sorted $\lvert\mathrm{draw}\rvert$ values over the pooled draws, written $\lvert q_{90} \rvert$ (Section 3.3); a layer's *local* band uses that layer's draws alone (Fig. 9).
floor	The minimum detectable effect: the SU at which the decision rule reaches 80% power when bootstrapped on the audit's own per-seed noise. It is 0.0050 nats at 10^{-4} with two control streams and 0.0028 at 10^{-5} (Section 6, Section 6.1).
cell; cell-group	A *cell* is one audited (model, checkpoint, layer): its candidates adapt the same MoE block, and its control branches are computed once per (τ, stream) and shared. A *cell-group* is a cell's candidates at one horizon τ, i.e. the set read against one shared control pair; its control-stream offset is the spread across seeds of the per-seed mean SU over those candidates (100 groups in the main audit, Section 5).
sign rule	The decision rule fires only when the mean SU exceeds the band *and* the per-seed signs agree (all positive for an upward firing); the crossing count of Section 5 applies it in both directions (Section 3.3).
composite; share	A *composite* is the budget-neutral operation ($\Delta P{=}0$): retire an idle expert and split a busy one into the freed slot. An expert's or group's *share* is the fraction of a layer's routing it carries, that is, its usage summed over the group on the estimation sample (Section 7).

0.00105 and $\max \lvert\mathrm{draw}\rvert$ 0.0034, giving a band of 0.0015. Averaged into disjoint pairs, which is the two-seed mean the rule actually thresholds, they give s.d. 0.00084 and a band of 0.0014. Against the unpaired 0.0025 that is a factor of 1.6–1.7, less than the 2.2 the variance ratio alone would predict. These draws span two passes and so carry any between-pass drift as well as run-to-run noise, and removing each layer's mean drift gives 0.0013. Re-running the power bootstrap of Section 6 against the matched band moves the low-rate floor from 0.0028 to 0.0019, or to 0.0018 at the two-seed-mean band and 0.0016 with drift removed. The rule's false-positive rate at $\delta{=}0$ rises only from 0.0000 to 0.0014, an order of magnitude inside the band's nominal 10%. The negative sharpens and no verdict moves. Of the 101 low-rate operations, 42 cross downward and none crosses upward, against 35 and none at the unpaired bands. Only the OLMoE band is measured matched. For the Qwen and on-distribution bands we scale the published ones by the same 0.63 and do not present the result as measured. The best operation anywhere in that pass sits at 0.20 of the matched band, against 0.13 of the unpaired one, and the composites are untouched because they fail the sign rule at any band. Per layer the matched band is not usable. The reason is the order statistic, not anything about these particular draws. Those bands run 0.00005 to 0.0028, each estimated from six draws, where the estimator of Section 3.3 lands on the fifth of the six. A rank-k-of-n band of this kind is exceeded by a fresh draw from the same distribution with probability $(n{+}1{-}k)/(n{+}1)$, whatever that distribution is, provided the draws are exchangeable. That is $2/7 = 29\%$ here, against 13% at the $n{=}46$ pooled band and 11% at $n{=}129$. A six-draw per-layer band therefore admits pure noise nearly a third of the time by construction, and the in-sample figure of 17% understates it. A leave-one-out check, which returns 12 of 36 here, cannot make the same point. With six draws and this order statistic it gives exactly 2 of 6 for every distribution, forced by the ranks alone, so it measures the rule and not the tail. Crossing a band this permissive settles little either way, and its single-draw exceedance is not

the rate of the two-seed-mean-and-sign rule. Of the retires, 9 of 24 cross it, seven of them with both seeds agreeing and two of them upward. The two upward cases retire e_3 of layer 8 at $+0.000096$ ($1.1\times$ its layer's band, seeds $+0.000098$ and $+0.000094$) and e_8 of layer 12 at $+0.000168$ ($3.5\times$, seeds $+0.000154$ and $+0.000182$), and both are near-dead experts. A band this permissive cannot settle them either way, so we re-tested both directly (Appendix C). On sixteen control streams each, fourteen of them new, they score $+0.000200$ and $+0.000138$, are positive on every stream, and beat a permutation placebo on every stream. These are gains of order 10^{-4} nats, about 0.020% and 0.014% in perplexity, and they lie 9.5 and 14 times below the matched floor of 0.0019. They hold on the held-out WikiText test split and reverse on OLMoE's own training mixture (Table 4). The audit's decision rule does not flag them, and the targeted test confirms them on WikiText, where they make these two retires marginally better than free. On the training mixture both cost loss. Per-layer results elsewhere are read against layer-local unpaired nulls, and the audit as a whole against the pooled band. The layer-local nulls are three draws per layer, two at τ=60 and one at τ=180, summarised by $\max|\mathrm{draw}|$ rather than by a quantile (Fig. 9). At 10^{-4} we keep the unpaired band. The only matched draws available at that rate are the ten per-stream permutation placebos of Appendix C plus four re-run controls (n=14, s.d. 0.0029, band 0.0025). These come from cells selected for their excursions, and one of them reads 0.0082, above the unpaired band itself. Table 3 decomposes the noise. The per-run component is measured directly, by re-running 36 control branches at configurations already probed in an earlier pass and differencing them. It is the reason a fixed stream does not reproduce, and it makes the offset of Section 5 a property of one run rather than of a stream.

Table 3: Noise components. The first two split a null draw and the last two split the reported SU. These are different denominators, so the shares do not sum. The 36% is the fraction of variance in reported SU means that removing the per-cell, per-seed mean absorbs, while 48% is the fraction of a null draw's variance attributable to re-running rather than re-streaming. The run component is strongly depth-dependent (s.d. 0.0014 at layer 0, 0.00004 at layer 12), which is why early layers dominate the band and why two draws per layer under-sampled it (Section 6.1). The 46% quoted in Appendix C ('Shared controls within a cell') uses the unrounded s.d. 0.00144 and 0.00156.

component	how estimated	s.d.	share
run-to-run	36 identical re-runs, 10^{-5}	0.0010	48% of null variance
stream-to-stream	remainder of the null	—	52% of null variance
cell offset	spread of per-seed candidate means	0.0014	36% of SU-mean variance
candidate probe	within-candidate seed differences	0.0016	per seed

The last two rows predict the band scaling with control-stream count used in Fig. 11.

Hardware and compute. Every OLMoE and Qwen probe ran in bf16 on one GPU with the base model held in host memory, while the coarse model stayed resident on the GPU (Appendix C). The earlier passes ran on one 84 GB card. Our logs record it only as an 84 GB card, without the SKU. These passes are the converged OLMoE and Qwen audits, the 0.4%, 2%, 25% and 66% trajectory passes and their rechecks, the random candidates and the dynamic-range sweeps of Fig. 8. Every pass from the 0.8% and 1.2% checkpoints onward ran on one 48 GB RTX 4090-class card, a 48 GB variant of that part (nvidia-smi reports 49,140 MiB). These include the amplitude positive control, the high-power, on-distribution and multi-stream re-probes, the low-rate re-audit and the follow-ups of Appendix C. The exceptions are the runs that followed them, which used a 32 GB virtual GPU: the two forward-only analyses, group destruction (Section 7) and routing by source (peak 16.6 GiB), the coarse-model audit with its re-probe and clean re-run, the GradMax composite, and the re-test of two low-rate retires in Appendix C. That re-test's held-out and training-data runs used two 32 GB cards of another type that reproduce the earlier base losses (Appendix C). A full 16-layer converged audit took 3.1 h, the six-seed L0 recheck 23 min, and a trajectory audit ∼80–100 min per checkpoint plus checkpoint download. Peak memory is the base model plus one probe copy (≈28 GB for OLMoE in bf16).

C Protocols for the Follow-up Experiments

Appendix B describes the main audit. Everything added afterwards is specified here, and all of it uses the same probe, band estimator and decision rule unless noted.

Multi-stream re-probe and placebos (Table 1). Every candidate that nominally crossed the pooled band was re-probed on all six control streams at τ=60 (four at τ=180). Each cell has two placebos, both exact symmetries of the network. The permute placebo applies one permutation to an expert's gate rows, up rows and down columns together. The rescale placebo multiplies the up-projection rows by c=2 and divides the down projection by c. A forward pass verifies that the two preserve the function, to 5×10^{-8} and 0 respectively. This gives 20 per-stream placebo values in all (6+6+4+4).

Alignment-aware merge and retire-and-reroute (Section 5). Both run in layers 0/5/10/15 of converged OLMoE. The alignment permutation maximises the correlation between the two experts' intermediate activations on 4096 routed tokens and is found by solving a linear assignment problem with the Hungarian method. The mean matched correlation is 0.145, against -0.001 for the identity. Zero-shot damage is $\mathcal{L}$ after surgery, before either branch adapts.

Split differentiation (Section 5). One composite per layer in layers 0 and 8 is adapted for 540 steps, with diagnostics at $0/60/180/540$. The diagnostics are the output correlation of the two children on 4096 routed tokens and their routing overlap, the fraction of tokens selecting both among those selecting either.

Low-rate re-audit (Section 6.1). The re-audit runs at 10^{-5} and τ=60 with two paired seeds and six control streams. On OLMoE/WikiText it covers layers 0, 2, 4, 6, 8, 10, 12 and 15, with four retires and two merges each, plus composites at τ=540 in layers 0 and 8. On Qwen it covers layers 0, 6, 12, 18 and 23, with three retires and two merges each. Qwen's 151,936-token vocabulary makes fp32 logits 2.5 GB at batch 4, so that pass uses batch 2 with τ=120 and 96 evaluation batches. This keeps both token budgets identical to every other run (245,760 adaptation, 196,608 evaluation). The on-distribution pass covers layers 0, 5, 10 and 15.

Null draws at 10^{-5}. The WikiText band uses 46 draws. Of these, 16 come from the audit itself (two per layer) and 30 are control-only draws added afterwards, from layers 0, 2, 4, 6, 8 and 12 with ten streams each, paired disjointly into five. The Qwen and on-distribution bands rest on 10 and 8 draws and are correspondingly uncertain. They are 0.0009 with bootstrap interval $[0.0002, 0.0012]$ and 0.00014 with $[0.00005, 0.00021]$. The WikiText correction (0.0017 at 16 draws, 0.0025 at 46) should therefore be read as a warning about both.

Positive controls at 10^{-5}. Each control amplifies the busiest expert's output projection by g, freezes it, and audits retiring it. On WikiText, layer 8, $g \in \{8, 4, 3, 2\}$ gives SU $= +0.0681$, $+0.0128$, $+0.0059$ and $+0.0012$, all detected. On Qwen layer 12, $g \in \{8, 3\}$ gives $+0.0219$ and $+0.0019$, and on-distribution layer 8, $g \in \{8, 3\}$ gives $+0.0243$ and $+0.0012$.

Re-initialized composites (Section 5). These run in layers 0 and 8 at τ=540, on the same cells and control streams as the perturbed-copy baseline. A fraction $r \in \{0.25, 0.5\}$ of the child's intermediate units is re-initialized by sampling $\mathcal{N}(0, \sigma^2)$, with σ the s.d. of the corresponding parent tensor. The parent is left intact, since re-initializing an expert that already carries traffic would impose a zero-shot cost unrelated to whether expansion pays. There are six candidates in all.

On-distribution random and quietest passes (Section 5). Both passes run at 10^{-4} and τ=60 in layers 0/5/10/15. The random pass draws 28 candidates without regard to usage, 12 of them retires. The quietest pass selects the four lowest-usage experts per layer from a 204,800-token on-distribution estimate rather than the 8,192 the main audit used, and adds two merges from the same larger sample, for 24 candidates.

Low rate with restored step budget (Section 6.1). Layer 0 is audited at 10^{-5} with τ=1800, which matches the lr $\times\ \tau$ product of 10^{-4} at τ=180. The run uses two paired seeds, four streams and one null draw.

Dynamic range versus routing share (Section 7, Section 6). This run also supplies the per-layer dynamic-range figures quoted in Section 6. Destroying a layer's busiest expert costs 0.0052 in layer 0 (e_{41}), 0.0018 in layer 8 (e_{18}) and 0.0100 in layer 15 (e_{26}). The layer-8 value is consistent with the λ=1 point of the positive-control sweep (0.0022) on the same expert with the identical perturbation. Over three noise draws the same expert costs 0.0017–0.0021 against 0.0022, a gap at the edge of the draw-to-draw spread. The three layers are therefore directly comparable, and the layer-8 figure is not model-wide. Three layers are too few to give the range across all sixteen. The run covers layers 0, 8 and 15 of converged OLMoE, with eight experts per layer, log-spaced by usage rank, that span 0.0009 to 0.048 of the routing. Damage is zero-shot. The expert's two tensors are replaced by

Gaussian noise matched to each tensor's s.d. and the loss is re-evaluated with no adaptation, so the measurement carries no trajectory noise. The model is restored bitwise after each expert and the base loss re-checked. Three experts per layer are repeated with two further noise draws to establish the floor, with median spread 0.0001. Three of 24 experts show non-positive damage, all at the quiet end. They are dropped by the log-log fit, which therefore has an upward-biased low-usage arm, and the fit uses the remaining 21 over 1.44 decades.

Low rate with headline step budget (Section 6.1). Layers 4 and 8 are audited at 10^{-5} with τ=600, which matches the $\mathrm{lr} \times \tau$ product of 10^{-4} at τ=60. Each layer gets one retire, one merge and one composite, with two paired seeds, four streams and one null draw each.

Re-initialized composite trajectory (Section 5). This uses layers 0 and 8 with r=0.5 at 10^{-4}, and an independent control pair at each of $\tau \in \{60, 180, 540\}$. The child-pair output correlation and routing overlap are measured on the same runs that give SU.

Band versus horizon. Only control branches are run, with no operations. They cover layers 0 and 8, with six streams at each of $\tau \in \{60, 180, 540\}$ at 10^{-4}. We report the median of the three disjoint-pair draws per cell and each branch's deviation from its cell median. We avoid the all-pairs band because it lets one stray branch enter five of fifteen differences, which inflates an upper quantile.

Coarse-grained model (Section 7). JetMoE-8B (Shen et al., 2024) has 24 layers and 8 experts with top-2 routing and was trained from scratch. The released checkpoint holds $8{,}522{,}237{,}952$ parameters when the tied input and output embedding is counted once, and our loader logs the same number. Its authors round this to 8B and count 2.2B as active per token. We run in bf16 on one 32 GB card with the model resident on the GPU. The audited block is snapshotted to host memory and restored per branch, and each 4096-token step is two accumulated micro-batches (exact). Only the MLP experts are audited, and the attention mixture is untouched. The router takes top-k on raw logits and then a softmax over the k, so retire masks the logit before top-k and zeroes the expert's output and router rows. The audited layers are 12, 23 and 0. The rate is chosen at layer 12 by the frozen amplitude control ($g \in \{8, 4, 2\}$) at 10^{-5} and 10^{-4}. The audit runs at τ=60 and 10^{-5}. In each layer it probes the quietest retire, a random retire and the most functionally similar merge, with two paired seeds and six control streams. There are 46 pooled no-op draws, 38 control-only plus the sweep's and audit's spare pairs. No composite is audited on the coarse model, whose τ=60 audit covers retire and merge only, so the composite negative rests on OLMoE and Qwen. The dynamic range uses zero-shot scale-matched noise on every expert, with three draws on the busiest, and on the busiest pair. Losses are given in nats per token and in bits per byte, where the base loss of 1.884 nats is 0.6905 bpb. The run's final integrity sweep aborted. It re-checked layer 12 while the other two layers still held their last branch's adapted weights. This was an ordering error in the check and not in the measurements, since every branch resets its own layer and both arms of a pair share the same state of the others. The reported file is therefore assembled from the per-phase checkpoint (42 draws) and the run log (4 draws). The check is corrected in the released `experiments/s17_jetmoe_audit.py` (Reproducibility Statement). The measurements are anchored independently of that check. Every branch begins with `block_reset` from a snapshot taken before any adaptation. That snapshot covers exactly the four tensors a probe or an operation can touch, and a CPU test asserts that its restoration is bitwise. Eight whole-model checks passed before the ninth aborted. The recorded control losses also carry no trace of the residual. Layer 23's six controls average 1.87959 against a base of 1.88433, whereas the preceding layer-12 branch would have left them 0.0053 lower, four times the band. Section 7's layer-0 cell was then reproduced by `s19`, a separate process that loads the checkpoint afresh and touches no other layer. It matches to six decimals on the zero-shot quantity (-0.005083 against -0.005083) and to 0.00002 on SU. None of this amounts to a re-run of all nine candidates with the anchor checked at every branch. The released `experiments/s20_jetmoe_clean_rerun.py` does exactly that, asserting bitwise restoration of every audited layer and reproduction of the whole-model base loss before each branch, and we report its result. A fresh process that loads the checkpoint again and rebuilds the probe data by the same path gives a base loss of 1.884327 against the 1.884330 recorded, and all 46 branch anchors passed. At every one of them the four-batch anchor loss reproduced the original exactly. All nine candidates come back with the same sign, and six of the nine zero-shot damages agree to six decimals. The largest SU difference anywhere is 0.0006, half the pooled band, at layer 0's quietest retire, where -0.0038 becomes -0.0044. The layer-0 positive returns at $+0.0028$ against $+0.0031$. Differences are largest in layer 0 and vanish in layer 23, the same depth ordering of run-to-run noise that Table 3 reports. The failed sweep was a bug in the check, not in the measurements.

Re-probe of the coarse-model positive (Section 7). The re-probe targets expert 4 in layer 0 of JetMoE-8B, at τ=60 and 10^{-5}. On WikiText we measure zero-shot damage on the audit's 197k validation tokens and on the test split, with the routing redistribution masking causes. Retire, permute and rescale each run on the cell's six control streams, with the audit's 46-draw pooled band as reference. Permute applies one permutation over the expert's gate rows, up rows and down columns. It is exact in real arithmetic, but bf16 reduction order leaves a zero-shot residual of $+0.00045$ nats on validation and $+0.00017$ on test. Rescale (up rows $\times 2$, down projection $\times \frac{1}{2}$) is bitwise exact in bf16. The on-distribution proxy is OLMoE-mix, with 2M adaptation tokens at its published shares and 197k evaluation tokens that are 98% DCLM web text and 2% algebraic-stack. It uses six control streams plus ten control-only streams, which give eight disjoint draws and a band with bootstrap interval. The retire runs on six streams and each placebo on three. As for the OLMoE crossings, the verdict rule requires a mean above the band with every stream positive. Clearing the permutation placebo and reproducing off the validation slice are reported separately from that verdict. The rescaling, an adaptation-speed control (Table 5), cannot say whether the gain needs removal. For the OLMoE crossings, which fell to zero when re-probed, the all-stream and sign criteria alone were decisive, so the placebos there served as calibration. This candidate passes those criteria, so under that rule it is a positive. The held-out and distribution tests show that it does not transfer, without overturning it under Eq. (1).

GradMax composite. The setup is that of the re-initialized composite: layers 0 and 8, $\tau \in \{60, 180, 540\}$, 10^{-4}, two seeds and an independent control pair per horizon. Here the child's re-initialized half is built after the inverse, outgoing-zero variant of GradMax (Evci et al., 2022). Those units' down-projection columns are zeroed so they contribute nothing at initialization. Their up-projection rows are the mutually orthogonal directions that maximize the gradient norm on the parent's routed tokens, rescaled to the parent's row norms. The parent is untouched.

Grouped destruction (Section 7). In layers 0, 8 and 15, $k \in \{1, 2, 4, 8, 16, 32\}$ experts are replaced jointly by scale-matched noise, zero-shot, with routing left intact. For each $k > 1$ we destroy three random subsets and the k busiest experts. The eight k=1 points are log-spaced by usage rank as in the dynamic-range run. Usage is estimated on 49k tokens, and the corrupted share is the summed usage of the destroyed experts. Three configurations per layer are repeated with two further noise draws, with spread ≤ 0.001. The reported per-layer exponents and R^2 come from log-log fits per layer over the group points ($k > 1$) with positive damage.

Routing by source (Section 5). Each of five OLMoE-mix sources is split into two document-disjoint halves of 500k tokens. The sources are DCLM, StarCoder (the shard fetched is C source), pes2o, open-web-math and algebraic-stack. Routing counts for all 16 layers are recorded in one forward pass per half. Between-source divergence is the mean Jensen–Shannon divergence over the four half-pairs of a source pair. Within-source divergence is the divergence between a source's own halves and serves as the sampling control. Domain preference uses the pooled halves per source, and the quiet set is the four least-used experts.

Run-to-run reproducibility (Table 3). The same 36 control branches, at 10^{-5} and τ=60 in layers 0/2/4/6/8/12 on streams 0–5, were run once in the low-rate audit and again in the null-draw pass, and the component is the difference between the two runs. The passes ran days apart, so the component includes any environment drift between them.

Re-test of two low-rate retires (Appendix B, Table 4). The re-test retires e_3 of layer 8 and e_8 of layer 12 of converged OLMoE at 10^{-5} and τ=60, with the low-rate audit's evaluation slice and adaptation sampler unchanged. The new streams are new adaptation runs, not new evaluation text. The re-test therefore confirms the gains across adaptation streams on the text the audit flagged them on, and the second run below tests them on held-out text and on the training mixture. Each cell has sixteen control streams, the audit's own two as an anchor and fourteen new ones. Every stream carries the retire, a permutation placebo and a c=2 rescale control of the same expert. Six of the streams also carry a sibling near-dead retire that the audit had measured negative, and six repeat-run controls give a matched null measured in the cell itself. Both symmetries are checked before adaptation. Rescaling changes the zero-shot loss by exactly 0, and permutation changes it by $+0.000083$ and -0.000012, a bf16 summation-order residual. In layer 8 the retire scores $+0.000200$ (paired Student-t 95% interval $+0.000159$ to $+0.000241$) and is positive on all 16 streams. On the 14 new streams alone it scores $+0.000194$ ($+0.000150$ to $+0.000239$), all positive. Minus the permutation it is $+0.000189$ ($+0.000157$ to $+0.000222$) and minus the rescale $+0.000152$ ($+0.000119$ to $+0.000186$), each

positive on all 16. The placebos score $+0.000011$ and $+0.000048$, both intervals containing zero, and the sibling stays negative (-0.000124, on 5 of 6 streams). In layer 12 the retire scores $+0.000138$ ($+0.000111$ to $+0.000165$), positive on all 16, and $+0.000135$ ($+0.000105$ to $+0.000165$) on the 14 new ones alone. Minus the permutation it is $+0.000133$ ($+0.000107$ to $+0.000160$) and minus the rescale $+0.000133$ ($+0.000109$ to $+0.000157$), each positive on all 16. The placebos score $+0.000005$ each, and the sibling scores -0.000637, negative on all 6. Most of the gain arises during adaptation, about nine-tenths of it in layer 8 and three-fifths in layer 12. Before adaptation the two retires move the loss by only -0.000023 and -0.000057. The re-test ran on a 32 GB virtual GPU and the audit on a 48 GB card. On the anchor streams the adapted control losses reproduce to within 1.3×10^{-4}, but the per-stream SU does not reproduce exactly. Layer 8's anchor streams give $+0.000185$ and $+0.000301$ against the audit's $+0.000098$ and $+0.000094$, and repeat runs of one configuration on the virtual GPU differ with a standard deviation of 4.5 to 7.9×10^{-5}. The sign therefore reproduces across hardware, and the magnitude only to within that noise. The script `verify_s21.py` checks the run's completeness, its placebos, the sibling's sign and the data path before it recomputes every figure quoted here.

Held-out and training-data tests (Section 6.1, Section 7, Tables 4 and 5). For OLMoE, the re-test's design is run again with every adapted branch evaluated twice from the same weights, once on the audit's validation slice and once on the first 197k tokens of the WikiText-103 test split, built by the same path. Separately, it is run with adaptation on the OLMoE-mix adaptation pool of Section 5 and evaluation on a document-shuffled rebuild of its evaluation pool. The rebuild holds the same documents in the same per-source quantities and is token-identical to the original pool before shuffling. Its first 197k tokens are 99.3% DCLM and 0.7% pes2o (166 documents), where the unshuffled slice was 93.7% DCLM and 6.3% algebraic-stack. For JetMoE layer 0, the six control streams carry the retire of e_4, its permutation, its rescaling at $c \in \{\frac{1}{2}, 2, 4\}$, the rescaling of e_1 at $c \in \{2, 4\}$, the retire of e_1 and six repeat controls, and each branch is evaluated on the validation slice and the test split. A rule written into the driver before the run fixes e_1. It excludes e_4's most functionally similar expert, ranks the rest by their distance to e_4 in destruction damage and routing share, and takes the first whose zero-shot removal costs more than 10^{-3} on the validation slice. The RefinedWeb test draws 2M adaptation tokens and two disjoint 205k-token evaluation pools (366 and 335 documents from 17 and 14 crawl dumps) from 48 row groups of as many shards, each pool evaluated on its first 197k tokens. The retire, the permutation and the c=2 rescaling run on six streams, with ten control-only streams for a band. These runs used two 32 GB cards that reproduce the earlier base losses to all six recorded decimals (2.224024 for OLMoE, 1.884327 for JetMoE). The scripts `verify_s22.py`, `verify_s23.py` and `verify_s24.py` gate completeness, bookkeeping, placebo exactness and those anchors before recomputing every figure quoted.

Reconciling the probe counts (Section 1, Table 6). Table 6 lists every audited setting. The 175 trajectory probes are six passes of four layers at seven candidates per (layer, checkpoint) cell (three retires, two merges and two composites), plus a seven-candidate four-seed recheck. The settings of the main audit sum to 398 records. Six of them repeat an operation already counted, because the recheck re-probes six of its seven candidates and substitutes a seventh. A further 9 candidates in three cells were probed at $16\times$ evaluation power. Six of them are re-probes and three are composites that this count had not otherwise reached, giving $398{+}9 = 407$ results over $392{+}3 = 395$ distinct operations. The controls that follow the main audit are reported separately and not folded into that number. They are the multi-stream re-probes and placebos of Table 1, the merge-rule and split-differentiation experiments of Section 5, the 101-operation low-rate re-audit of Section 6.1, and the placebo-controlled re-test of two of its retires. Five candidates, the layer-4 cluster and a layer-0 composite, exceed their own layer's local null (Fig. 9), and they are precisely the ones that vanish when re-probed on every control stream (Table 1).

Symptoms the triggers key on (Section 1, Section 5). OLMoE's expert usage spans 0.0001–0.062 across layers against a uniform 0.0156, exactly the imbalance an activation-count or load trigger treats as slack. Yet retiring near-dead experts is at best *marginally* loss-improving. It gives mean SU $= +0.0001$ on WikiText, -0.0001 in the lower-rate re-audit of Section 6.1, whose band is $1.6\times$ tighter, and -0.0011 on the training mixture. The two we re-tested on sixteen streams gain $+0.000200$ and $+0.000138$ on WikiText and lose 0.00035 and 0.00091 on the training mixture (Table 4). Merging the most functionally similar pairs is harmful in a bimodal way. OLMoE scores -0.0012. On Qwen the family mean of -0.0043 comes from one near-duplicate pair at -0.039,

Table 4: Re-test of the two low-rate retires (Appendix C), in units of 10^{-4} nats. The upper block uses the audit's own 197k-token validation slice, the text on which the two were first flagged. Its new streams are new adaptation runs, not new text. The lower block is a second run of the same design. It is evaluated again on the validation slice and on the WikiText-103 test split, which the candidates were not chosen on, and adapted and evaluated on OLMoE's own training mixture (a document-shuffled evaluation pool). Counts are streams out of 16 on which the quantity is positive. Converged OLMoE at 10^{-5} and τ=60, with sixteen control streams per cell, the audit's own two and fourteen new. Intervals are paired Student-t 95% over streams, and every difference is taken stream by stream. In the upper block both differences are positive on all 16 streams in both cells. The sibling is a near-dead retire in the same cell that the audit had measured negative. The audit's decision rule does not flag these gains. The re-test confirms them on WikiText, held-out text included, and the training mixture reverses them.

	layer 8, retire e_3	layer 12, retire e_8
first-pass SU, the audit's two streams	$+0.96$	$+1.68$
SU, all 16 streams	$+2.00$ [$+1.59$, $+2.41$]	$+1.38$ [$+1.11$, $+1.65$]
SU, the 14 new streams alone	$+1.94$ [$+1.50$, $+2.39$]	$+1.35$ [$+1.05$, $+1.65$]
streams with SU > 0	16/16	16/16
retire $-$ permutation placebo	$+1.89$ [$+1.57$, $+2.22$]	$+1.33$ [$+1.07$, $+1.60$]
retire $-$ rescale control (c=2)	$+1.52$ [$+1.19$, $+1.86$]	$+1.33$ [$+1.09$, $+1.57$]
permutation placebo	$+0.11$ [-0.38, $+0.59$]	$+0.05$ [-0.28, $+0.38$]
rescale control (c=2)	$+0.48$ [-0.04, $+1.00$]	$+0.05$ [-0.19, $+0.29$]
sibling retire (audit: negative)	e_{46}: -1.24, 5/6 negative	e_{63}: -6.37, 6/6 negative
zero-shot $\mathcal{L}_{\text{edited}} - \mathcal{L}_{\text{original}}$	-0.23	-0.57
repeat-run s.d., six pairs	0.79	0.45
perplexity improvement	0.020%	0.014%
SU, validation slice, second run	$+1.80$ [$+1.34$, $+2.26$], 16	$+1.50$ [$+1.24$, $+1.76$], 16
SU, WikiText test split	$+2.00$ [$+1.72$, $+2.29$], 16	$+1.01$ [$+0.54$, $+1.49$], 15
retire $-$ permutation, test split	$+2.12$ [$+1.59$, $+2.65$], 16	$+1.07$ [$+0.72$, $+1.42$], 15
sibling retire, test split	-1.32, 0 of 6	-4.91, 0 of 6
SU, OLMoE training mixture	-3.53 [-3.91, -3.16], 0	-9.10 [-9.41, -8.80], 0
retire $-$ permutation, mixture	-3.65 [-3.98, -3.31], 0	-8.98 [-9.26, -8.70], 0
zero-shot, test split / mixture	-0.75 / $+3.13$	$+0.78$ / $+8.97$
usage, validation (8k) / mixture (197k)	0.0008 / 0.0082	0.0011 / 0.0089

Table 5: The retire of e_4 and the rescaling control in JetMoE-8B layer 0 (Section 7, Appendix C), on the validation slice the audit measured, on the held-out WikiText test split and on RefinedWeb, in units of 10^{-3} nats. Every value is SU against the un-mutated control of the same stream, with a paired Student-t 95% interval over six control streams and the number of positive streams. Rescaling multiplies an expert's up rows by c and its down projection by $1/c$, and its zero-shot change is exactly 0 everywhere. e_1 is a layer-0 expert whose removal costs loss, chosen by a rule fixed before the run.

	WikiText validation	WikiText test	RefinedWeb
retire e_4	$+3.01$ [$+2.66$, $+3.36$] 6	-0.12 [-0.38, $+0.15$] 3	-3.59 [-3.77, -3.41] 0
permute e_4	-0.02 [-0.36, $+0.31$] 3	$+0.09$ [-0.25, $+0.42$] 3	-0.01 [-0.25, $+0.24$] 3
rescale e_4, c=$\frac{1}{2}$	$+0.92$ [-0.45, $+2.30$] 5	$+1.08$ [-0.16, $+2.32$] 5	—
rescale e_4, c=2	$+2.80$ [$+2.59$, $+3.01$] 6	$+2.70$ [$+2.46$, $+2.95$] 6	$+0.01$ [-0.15, $+0.18$] 3
rescale e_4, c=4	$+12.60$ [$+11.96$, $+13.24$] 6	$+11.78$ [$+11.14$, $+12.42$] 6	—
rescale e_1, c=2	$+0.95$ [$+0.79$, $+1.11$] 6	$+1.07$ [$+0.84$, $+1.31$] 6	—
rescale e_1, c=4	$+4.79$ [$+4.03$, $+5.55$] 6	$+4.85$ [$+3.90$, $+5.81$] 6	—
retire e_1	-8.33 [-8.99, -7.67] 0	-7.17 [-7.93, -6.40] 0	—
zero-shot, retire e_4	-5.08	-1.64	$+3.63$
routing share of e_4, 49k tokens	0.166	—	0.123

against a median of -0.0004 for the other nine, so similarity-triggered merging is most dangerous exactly where the trigger fires hardest.

Evaluation size and the band (Section 1, Section 6.1). We predicted that $16\times$ the evaluation tokens would shrink the noise band $\approx 4\times$ throughout. It does so in well-behaved cells, where nulls fall to 0.0008 and, at that five-fold finer resolution, every candidate remains at or below zero, our strongest-resolution negative. Rare adaptation-stream excursions ($+0.033$ on a single seed), however, are untouched by evaluation size and dominate the tail. The experiment overturned our own stated prediction, and the guidance we would now give is the opposite, namely to buy control streams and seeds rather than evaluation data.

Placebo scores and dynamic-range precision (Section 1). In the re-probed cells of Table 1 the placebos score $+0.0003$ to $+0.0004$ in layer 4 and -0.0012 to -0.0014 in layer 0, indistinguishable from the real operations beside them. Both placebos preserve the function exactly, so their structural utility is zero by construction at zero-shot, up to a bf16 reduction-order residual of $+0.00045$ nats for the permutation. The per-layer dynamic ranges of Section 6 rest on three noise draws each, with spread ≤ 0.0004.

Utility-driven structural adaptation beyond MoE (Section 2). Deciding structural growth by estimated loss reduction goes back at least to the growing and pruning of hierarchical mixtures of experts (Waterhouse and Robinson, 1995). Modern instances include Splitting Steepest Descent (Liu et al., 2019), Firefly descent (Wu et al., 2020), and Self-Expanding Networks' natural expansion score (Mitchell et al., 2024). These rely on local (Taylor/Fisher) approximations of a structural change's utility. Our probes instead measure utility empirically over a finite adaptation horizon, which prices in optimizer and router dynamics. Our horizon study also shows that the measured utility of splitting changes sign with horizon, which cautions against zero-horizon estimates.

Adaptive MoE architectures (Section 2). DynMoE (Guo et al., 2025) adjusts top-k and both adds and removes experts during training, triggered by activation counts. MASS (Park and Park, 2026) and progressive expandable MoE (Jin et al., 2026) grow the pool. Layer-adaptive expert pruning (LAEP), introduced in the Yuan3.0 Ultra pretraining report (YuanLab.ai, 2026), prunes under-utilized experts during trillion-scale pretraining, and EPnG (Lee et al., 2026) prunes and grows under a fixed budget in PEFT settings. SIMoE (Chen et al., 2025) and learned-mask methods (Liu et al., 2025) discover expert structure when upcycling dense models. AutoMoE (Jawahar et al., 2023) searches heterogeneous architectures for neural machine translation before training, and expert-merging algorithms (Nguyen et al., 2026) supply merge mechanics. These methods operate during pretraining, during fine-tuning, or for inference efficiency.

Prune, merge, or grow (Section 2). The systematic one-shot comparison that finds pruning beats merging measures that ordering on generative benchmarks (Lasby et al., 2026). MC-SMoE (Li et al., 2024b) makes permutation alignment a prerequisite before averaging grouped experts. GradMax (Evci et al., 2022) chooses new-unit directions by gradient rather than by perturbation. Drop-Upcycling's evidence for partial re-initialization is a 500B-token continued-pretraining comparison, and its claim is explicitly "in the long term" (Nakamura et al., 2025). The three child initialisations we audit are perturbed copy, partial re-initialization, and gradient-directed (Section 5 and below). Our measurement that a perturbed-copy child stays 0.994-correlated with its parent after 540 steps is the same phenomenon observed at probe scale.

Compression literature (Section 2). Each family we audit has a literature that measures compression cost rather than quality gain, and those findings bracket ours. The one-shot comparison finds that pruning beats merging because merging forfeits fine-grained routing control (Lasby et al., 2026).

Run-to-run non-reproducibility (Section 3.2). This is the evidence behind the run-to-run row of Table 3. It properly belongs in Appendix B and is collected here because it also settles the layer-4 offset of Section 5. Common random numbers remove something narrower than "stream-level movement". A probe's two branches share the token stream but are separate optimisation runs, and those runs are not reproducible. Re-running 36 identical (layer, stream, rate, horizon) control branches moves the loss by 0.0005 on average (21% of the low-rate band) and by up to 0.0034, because bf16 kernels are not bitwise deterministic and 60 Adam steps amplify the difference. Those 36 re-runs come from two passes days apart, so strictly they measure a *repeat-run* difference and bound run-level noise from above. Three checks indicate that the environment is not what they caught. Each pass re-evaluates the un-mutated model on the identical validation tokens before adapting anything, and the two agree to the 5×10^{-6} of our logged precision, while that same fingerprint moves by 4×10^{-5} in later sessions. The 36 differences also carry no per-pass offset (17 of 36 positive,

mean $+0.0001$), and their s.d. falls $39\times$ from layer 0 to layer 12, which a library or driver change would not do. Finally, the same quantity measured inside one process is larger, not smaller (mean $|\mathrm{SU}|{=}0.0025$ over ten per-stream values at 10^{-4}). That quantity is the permute placebo, whose SU vanishes in exact arithmetic because an elementwise optimiser commutes with the permutation, so only non-associativity can move it. None of this excludes a change confined to the backward path, which a deterministic-kernel comparison would settle. Half the variance of a null draw (48%) is this repeat-run component rather than any property of the stream. The component is sharply depth-dependent, with s.d. 0.0014 at layer 0 against 0.00004 at layer 12.

Shared controls within a cell (Section 3.2). Sharing one control branch per (layer, τ, stream) across a cell's candidates has a consequence, which we quantify rather than assume. Removing the per-cell, per-seed mean absorbs 36% of the variance of the reported SU values, and the stream component accounts for 46% of per-seed variance. The two differ because seed averaging reduces one component and not the other. This sharing is the mechanism behind the control-stream offsets of Section 5.

The layer-4 offset on re-probing (Section 3.3). Re-probed on all six streams, the offset is gone. The re-runs above explain why it could not have been a property of the (cell, stream) pair. A control branch is not reproducible even at fixed stream. Identical re-runs move by 0.0005 on average, and layer 4 is among the least reproducible layers (measured at 10^{-5}, and we assume the depth ordering carries to 10^{-4}). What the first pass caught was therefore one run's excursion, and a second run of the same configuration lands elsewhere.

Horizon calibration (Section 3.2). Composite neutrality at $\tau{=}180$ would bound slack only up to that horizon, since the synthetic split utility is still rising at our largest audited horizon. Horizons beyond 540 are untested at the headline rate.

Synthetic provisioning and horizon audits (Section 4). The parameter adjustment there is flat because $\Delta P/P_e$ is exactly integral in these models. At $N{=}2$ the best operation is correctly a split whose short-horizon utility sits below the parameter price. At $N{=}8$ the best operations are retires and merges of redundant experts, and over the three training seeds every positive mean exceeds its local null, so op-type attribution is correct as well as sign. The raw readings exceed the local null in $3/3$, $1/3$ and $1/3$ seeds, and the raw ordering across N reverses the adjusted one. In raw terms $N{=}2$ is highest because the under-provisioned model has the most loss to buy with parameters. We report both utilities rather than let Fig. 6 imply that the ranking is a property of loss, and λ_p appears nowhere outside that figure. On the horizon side, retiring busy experts costs -0.5 to -0.8 (MSE) instantly in the over-provisioned model and up to -1.4 in the under-provisioned one, where every expert is essential. The cost shrinks as survivors absorb the traffic. By $\tau{=}540$ two of the three retires still sit at -0.03, while the third has crossed to $+0.004$. That third retire removes the busiest expert, which carries 53% of tokens. Zero- or short-horizon estimates systematically mis-rank expansion against contraction.

Granularity with everything else fixed (Section 4). The LM audit explains its negative partly by granularity, and Section 7 measures the share-to-damage relation inside OLMoE, but only a synthetic model can change granularity alone. On a task with $K{=}8$ latent functions we train a fine model ($E{=}64$, top-8, expert width 16) and two coarse ones ($E{=}8$, top-1 at width 128 and top-2 at width 64) on the same stream for the same 6k steps, three seeds each. The models are matched in total expert parameters (29,952 against 29,728) and in active parameters per token (3,744, 3,716 and 3,720). The top-2 model has half the total. Share is the mechanism. Destroying eight of the fine model's experts at once, a corrupted share near $1/8$, costs a median 0.44 (MSE), and destroying one of the coarse top-1 model's experts, the same share, costs 0.50, a ratio of 1.14. A coarse expert is worth what a fine group summing to its share is worth. The within-router specialisation that steepens single-expert curves in OLMoE ('Grouped destruction' below) does not carry across granularity here. No edit improves any of the nine models. Of 36 retire, merge and split candidates, none clears its model's band with a positive sign. At $E{=}K$ every resolved edit is large and negative (the top-2 model resolves 11 of 12). At $E{=}8K$ the fine model is over-provisioned, with 25–32 of its 64 experts below the band, and retiring the idlest measures exactly zero in all three seeds. That retire is free and not improving, much as the LM's near-dead experts are free and at best marginally improving. The test has a limit. Even at $E{=}64$ the synthetic band is two to three orders of magnitude below a single expert's worth, with the busiest expert at 120–760 bands and the median expert at 3, whereas in OLMoE the busiest is 0.6–3.6 floors. The synthetic isolates the mechanism and shows that edits do not pay at coarse

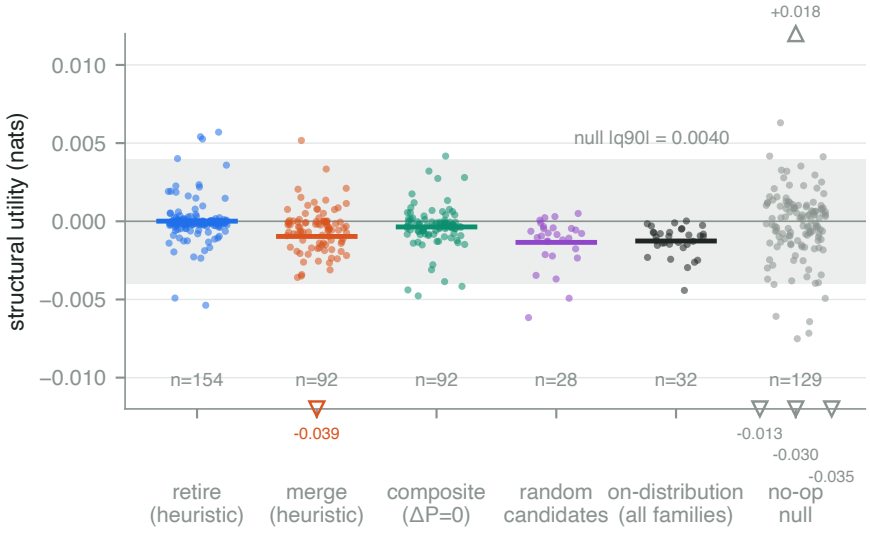


Figure 5: All 398 probed operations by candidate family, including the on-distribution replication on OLMoE-mix, against 129 no-op null draws (pooled band $|q_{90}| = 0.0040$). Bars are family means. No measured distribution rises above pure estimation noise, and for merge, random, and on-distribution candidates it is shifted below the null. Five of the 527 points lie outside the plotted range and are drawn as hollow arrowheads at the axis edge, labelled with their values. They are Qwen's 0.84-similarity merge at -0.039 and four null draws at $+0.018$, -0.013, -0.030 (layer 7, Section 3.3) and -0.035. Because four of the five are null draws, plotting them widens the null's visible spread more than the operations'. The fifth, the Qwen merge, is the largest point in the figure in absolute value.

granularity. It does not reproduce the regime in which the floor hides the effect, and we do not read it as doing so (Appendix A).

Probe cost of the LM audit (Section 5). One branch run, adaptation plus evaluation, costs 12–21 s. Excluding model load, the full 16-layer OLMoE audit took 3.1 h of probe time and the five-layer Qwen audit (5 of its 24 MoE layers) took 0.9 h.

Sign symmetry of the crossings (Section 5). Crossings are also sign-symmetric, with 5 up and 9 down. We do not lean on that, because several of the downward firings are real damage rather than noise, namely Qwen's -0.039 near-duplicate merge and the uniformly negative on-distribution families.

Cell-group offsets (Section 5). A cell-group is the set of candidates of one (model or checkpoint, layer) cell probed at one horizon τ in one pass, which share the same control runs. Its offset is the spread across streams of their mean SU. The main audit contains 100 such groups. Layer 4's offset, 0.0079, is the second largest. On one stream every layer-4 candidate scores $\approx+0.009$, and on the other ≈ 0. The largest, 0.0099, belongs to the two τ=180 composites at layer 0 of the 1.2% checkpoint, whose four-stream means run from -0.0039 on stream 2 to $+0.0061$ on stream 3. Neither candidate crosses the band, which is why the offset is a diagnostic and not a rejection rule. The layer-3 composite (seeds ≈ 0 and $+0.0064$) and the layer-0 composite ($+0.0042$) sit in cell-groups ranked 5th and 12th. A six-seed recheck of the layer-0 composite and its sibling returned $+0.0009$ and -0.0021, with 2/6 seeds positive each.

Merge similarity and alignment (Section 5). The most similar merged pairs correlate 0.76 (OLMoE) and 0.84 (Qwen) in output while their weights are nearly orthogonal. In audited OLMoE layers the most similar pair by weight-space cosine is $\approx$0.003, a degenerate criterion, which is why merge candidates are selected by output correlation on routed tokens. Most audited pairs are far less alike, with median output correlation 0.09 and 0.06. The near-duplicates invite the reading that naive interpolation, rather than merging itself, is what costs loss. That reading does not hold. MC-SMoE (Li et al., 2024b) aligns experts before averaging by solving the assignment problem that best matches

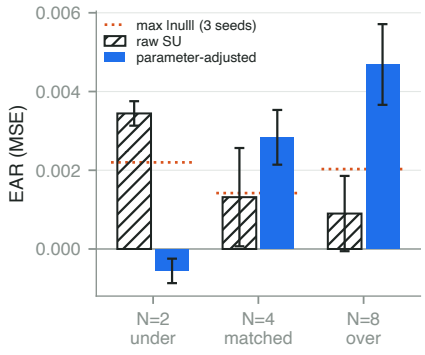


Figure 6: Synthetic validation of EAR against provisioning (truth $M{=}4$), mean $\pm$ s.d. over 3 training seeds. Values are reported both as raw SU (hatched) and parameter-adjusted ($\lambda_p{=}0.004$ per expert, Appendix A), and dotted lines are each architecture's $\max|\text{null}|$. On the adjusted utility EAR is positive exactly where structure is redundant or collapsed, negative under provisioning shortfall, and monotone in over-provisioning. On raw loss the ranking reverses. A split buys loss with parameters, so the under-provisioned model scores highest. The $N{=}2$ adjusted value is a sign inside its null and depends on λ_p (Appendix A). The contraction result, free at $N{=}8$ and costly at $N{=}4$, holds on both.

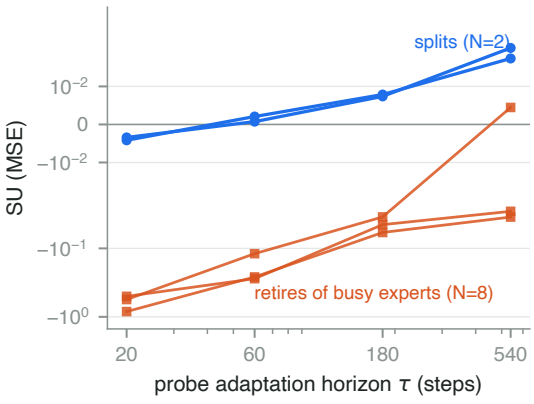


Figure 7: Probe-horizon bias (symlog axes). Split utility (blue, under-provisioned $N{=}2$) increases monotonically in τ and changes sign. The cost of retiring busy experts (orange, over-provisioned $N{=}8$) appears instantly and is absorbed with adaptation, and by $\tau{=}540$ one of the three curves has crossed zero.

e_j's intermediate units to e_i's. Here that alignment raises the mean correlation only to 0.145, so these experts are not permutations of one another. The aligned merge is indistinguishable from the naive one both before adaptation (zero-shot damage $+0.00167$ vs $+0.00169$ nats) and after (SU $= -0.00108$ vs -0.00113, four layers). The retire-and-reroute lower bound is three to four times gentler ($+0.00042$ zero-shot against $+0.00169$, SU $= -0.00039$ against -0.00113). If one of two similar experts must go, deleting it therefore beats averaging it in. REAP (Lasby et al., 2026) reaches the same ordering for one-shot compression by benchmark score, and here it is obtained by intervention on the loss.

Qwen's load profile within a layer (Section 5). The $25\times$ and $509\times$ usage ranges quoted for Qwen and OLMoE are across the audited layers. Within a single layer, where candidates are actually selected, the ratios are 4–$13\times$ and 9–$329\times$ respectively.

The WikiText heuristic contrast (Section 5). On WikiText, retiring a randomly chosen expert scores SU $= -0.0015$ on average (worst case -0.0049), while retiring a heuristic-selected near-dead one costs $+0.0001$. On OLMoE's own mixture, heuristic retires cost -0.00111 ($n{=}16$, four layers) and randomly chosen ones -0.00110 ($n{=}12$, same four layers and rate, from a separate random-candidate pass). These are two independent measurements that agree to 10^{-5}, not one number quoted twice, and the advantage is gone. The random pass is a spot check of a very large candidate space rather than an exhaustive sweep. It rules out a gross heuristic-selection artifact but not a rare improving operation.

The on-distribution sample (Section 5). The 12M-token OLMoE-mix-0924 sample follows the published source shares: DCLM 95.5%, StarCoder 2.5%, pes2o 1.4%, and open-web-math and algebraic-stack 0.3% each. The 8M adaptation pool is sampled at random offsets and so follows those shares. The 4M evaluation pool, however, is concatenated source by source without shuffling. As a result, the slice the audit evaluates on, its first 197k tokens, is 93.7% DCLM and 6.3% algebraic-stack,

with no StarCoder, pes2o or open-web-math. The 8k usage sample that chooses candidates, its first 8k, is algebraic-stack alone. The on-distribution numbers are therefore measured on DCLM-dominated web text, which is what the mixture mostly is, with candidates chosen on mathematical code, and the 8k-versus-200k usage comparison changes source as well as size. A shuffled pool is the fix. We have built it from the same documents, only reordered, and used it for the two retires of Table 4, whose experts it shows carrying 0.008–0.009 of the traffic. We have not re-run the on-distribution audit on it.

Limits of the 8k usage estimate (Section 5). On WikiText the least-used experts measured loss-free at mean $+0.0001$, with 20 of 64 positive, and in the lower-rate re-audit of Section 6.1 the same family measures -0.0001 over $n{=}32$. On the training mixture the cost does not vanish for experts our routing sample scores at zero usage (-0.0023 for one such expert). Re-estimating usage on 200k on-distribution tokens instead of 8k changes almost completely which experts look near-dead. The two rankings share one of their sixteen picks over four layers, and in three of four layers they share none. The experts the 8k sample scored at 0.0–0.0002 usage carry 0.013–0.022 at 200k, a hundredfold underestimate. It arises because the 8k sample is the evaluation pool's first 8k tokens, all algebraic-stack, while the 200k one is 94% DCLM, so the change is one of source as much as of size (the next paragraph shows the quiet set turning over across sources). The least-used expert in these layers still carries 0.0042–0.0127 of the routing against a uniform 0.0156, which puts the quietest one at 27% of even load rather than at zero. The eight merges of quietest pairs in the same pass average -0.0024, and none is positive. A better estimator would not recover free removal on the training distribution, because it is not there. Routing is also more concentrated on that algebraic-stack sample (mean max usage 0.102 vs. 0.046).

Usage estimates across sources (Section 5). Given 500k tokens, a usage estimate becomes stable within a source, where two document-disjoint halves of the same source agree on 3.1 of their four quietest experts. Across sources the quiet set turns over almost completely, agreeing on 1.0 of four. Routing itself moves with source far more than sampling can explain. The Jensen–Shannon divergence between the routing distributions of two sources is 0.125 bits against 0.0022 between two halves of one source, a factor of 58 that holds in every layer (46–74), and 15.5 of 64 experts per layer send more than twice their mean share to one source, against none in the within-source control. Code against prose is the widest split, and the two mathematics sources are nearly indistinguishable. A usage estimate taken on encyclopaedic prose therefore misses the experts that serve code, however many tokens it uses.

Differentiation of split children (Section 5). Over 540 adaptation steps of a split, the router does separate the children. In layer 8 the fraction of tokens selecting both falls from 0.88 to 0.65, and their combined share of the layer's traffic falls by 40% (0.100 to 0.060). Their outputs, however, stay correlated at 0.994, down from 0.999 at initialisation, and in layer 0 neither moves. The composite result is therefore not that expansion does not pay, but that expansion into a perturbed copy does not pay within 540 steps, because the copy is still the same function. The partially re-initialized children start at output correlation 0.80 and 0.58 rather than 0.999, and in layer 8 they stay differentiated after adaptation (correlation 0.51, routing overlap 0.55 against 0.88 for a perturbed copy), so the composite negative covers partial re-initialization as well as the perturbed copy. What we borrow from Drop-Upcycling (Nakamura et al., 2025) is the initialisation, not a prediction. Its evidence is a 500B-token continued-pretraining comparison whose stated claim is long-run, whereas $\tau{=}540$ here is 2.2M tokens (540 steps of 4096), smaller by a factor of $2{\times}10^5$. Nothing in our result contradicts it. The only question is whether the same initialisation buys anything a local edit can collect. A third initialisation, gradient-directed growth in the sense of GradMax (Evci et al., 2022), is probed over the same cells and horizons below ('Re-initialized children at longer horizons') and lands in the same place.

The trajectory grid (Section 5). OLMoE's public checkpoints at 0.4%, 0.8%, 1.2%, 2%, 25%, 66% and 100% of 5.1T tokens are 20B/41B/62B/104B/1.28T/3.38T/5.12T tokens, with base losses 2.965, 2.769, 2.685, 2.598, 2.402, 2.314 and 2.224. Each (checkpoint, layer) cell carries seven candidates, $7 = 3 + 2 + 2$: the three lowest-usage retires, the two most functionally similar merge pairs, and two budget-neutral composites that split the two busiest experts into the two idlest slots. That is one candidate fewer than the converged audit's eight per layer, the omitted one being the fourth-lowest-usage retire. The 100% point is read from the converged audit's own cells for these four layers, which carry that fourth retire. The trajectory's $\tau{=}60$ null band of 0.0035 rests on 38 draws, and the $\tau{=}180$ band is 0.0021 over 25. The 0.4%, 0.8% and 1.2% passes use four paired seeds and

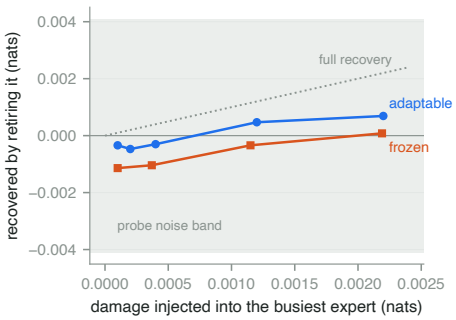


Figure 8: Positive control. Damage injected into the busiest expert (x-axis) vs. loss recovered by retiring it (y-axis), with and without freezing the damage. Both curves stay far below full recovery and inside the noise band.

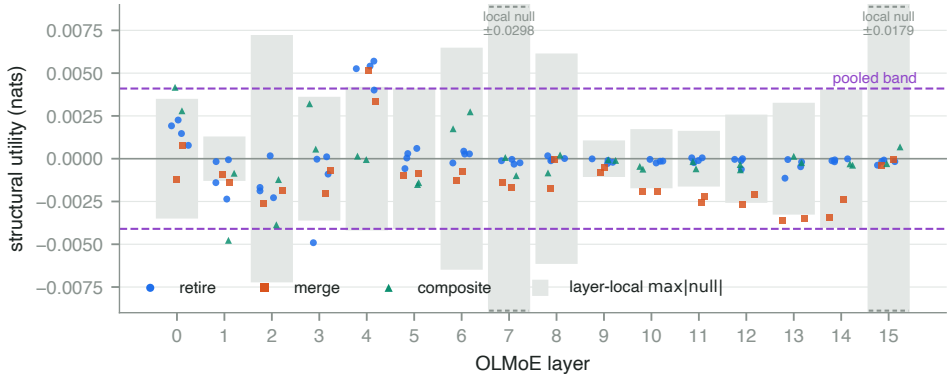


Figure 9: Converged OLMoE, per layer. Each candidate's SU is shown against that layer's own null (grey) and the pooled band (dashed). The local null is $\max|\text{draw}|$ over the three local draws every layer carries, two at τ=60 (streams 2−3 and 4−5) and one at τ=180 (streams 2−3), so each local null is itself a three-draw statistic. Local nulls span a factor of 30 across layers. Layer 7's ±0.0298 and layer 15's ±0.0179 run past the axis and are capped with a dashed line and labelled. The pooled band is therefore too wide for quiet layers and too narrow for noisy ones. Layers 9–13 resolve the negative three to four times more tightly than the headline figure suggests. Five candidates sit above both their local null and the pooled band: layer 4's cluster of four, and one layer-0 composite. Table 1 shows that none of them reproduces on the full set of control streams.

the rest use two. Figure 10 plots the τ=60 EAR, so its 100% point for layer 0 reads $+0.0022$ while Section 5 quotes $+0.0042$ for that layer's τ=180 composite. These are two horizons of the same cell, not two measurements of one quantity. We do not report per-checkpoint means, because the 2% pass re-probed only layer 0 and its mean is not comparable with the four-layer means elsewhere.

The one above-band trajectory cell (Section 5). The two-seed trajectory pass produced exactly one above-band cell, layer 0 at the 2% checkpoint ($+0.0036$), which was a candidate for early-training slack. Two independent re-measurements disposed of it. A four-seed repeat returned $+0.0021$, with local null draws of $+0.0013$ and $+0.0035$ bracketing the estimate. The high-power probe described in Appendix C reproduced the point estimate ($+0.0036$) but failed the sign-consistency rule, with two of four seeds negative and the mean carried entirely by a single adaptation-stream excursion of $+0.014$. The same stream inflated an unrelated merge candidate by $+0.015$. The sign-consistency rule and multi-seed rechecks exist for exactly this reason.

Usage heuristics on the training mixture (Section 5). On WikiText, retiring the least-used experts measured loss-free, whereas on the training mixture it is mildly costly (mean -0.0011, 1 of 16 positive). The cause is our own estimator. Re-estimating usage on 200k on-distribution tokens instead

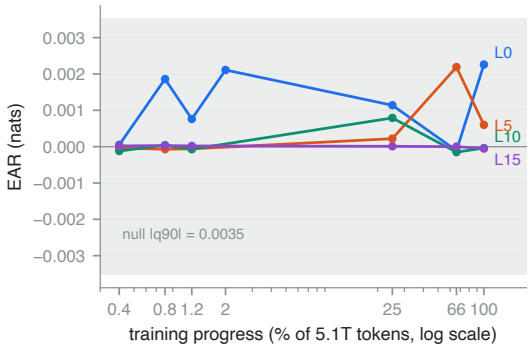


Figure 10: EAR (τ=60) across OLMoE's training trajectory, four layers, after the four-seed re-measurement of the one cell above band in the two-seed pass (2%, L0). No checkpoint is elevated above the pooled null band.

of 8k changes almost completely which experts look near-dead, and the two rankings share one of their sixteen picks over four layers. Measured properly, there are no near-dead experts to find. The quietest expert in these layers still carries 0.0042–0.0127 of the routing against a uniform 0.0156, and auditing those genuinely quietest experts costs -0.00074 (n=16, 15 of 16 negative), better than the -0.00111 the 8k picks cost but not free. "Near-dead" is a property of the sample used to define it, of its size and of its distribution. Nor does the contrast that motivated our heuristic claim survive the move. On OLMoE's own mixture, heuristic and randomly chosen retires coincide at -0.00111 and -0.00110 (Appendix C). Usage heuristics identify safe removals when usage is estimated on the probe distribution, and we have no evidence they do so otherwise.

Experiment inventory (Section 5). Every audited setting is listed in one place, so that the counts appear only once.

Sensitivity and power (Section 6, Section 6.1). The main text states the power analysis at the resolution the claims need. This part gives the rest of it: the positive-control sweep and its units, the assumption that turns destruction damage into a bound on retire gain, the band estimate at 10^{-5}, the cells that restore the step budget the lower rate removes, and the evaluation-token prediction that our own experiment overturned.

The amplitude sweep (Section 6). At g=16 the injected damage is 0.436 nats and retiring recovers $+0.0314$, which is 7.6× the band with both seeds positive. At g=8 the damage is 0.091 and the recovery $+0.0067$. Below that, detection stops. The sweep establishes that a real improvement of SU $\approx$ 0.007 nats is found at our standard settings (τ=60, two paired seeds, 197k probe-validation tokens) and that one of $+0.0003$ is not. It does not by itself locate the threshold, since the two gains bracketing it differ by a factor of five. Its units are also coarse. "Damage" is measured zero-shot on the amplified model, whereas SU is measured after both branches adapt, so the two columns are related by a recovery fraction (7.4% at large g, 1.7% at g=4). That fraction is a property of *this* defect, because amplification is something the router can partly route around. Structural slack, by definition, is what adaptation cannot remove, and would convert at a higher rate.

The power curve (Section 6). The bootstrap uses 922 centred per-seed values (s.d. 0.0025), heavy-tailed by construction. The positive control's detection at SU $= +0.0067$ and non-detection at $+0.0003$ bracket the same curve. Adding seeds at a fixed band does not move the curve, because the sign rule tightens as fast as the mean stabilises. Adding control streams does move it, to 0.0040 at k=4 and 0.0030 at k=6, because the band narrows with them. This puts numbers on the guidance about where to spend probe compute.

From destruction damage to retire gain (Section 6). Treating destruction damage as a bound on retire gain is an assumption rather than a result, and we state it as one. Noise substitution, masked-and-rerouted removal and post-adaptation utility are three different interventions, and we have not proved $\mathrm{SU}_{\mathrm{retire}} \leq$ destruction damage. Three considerations support the assumption. Noise substitution destroys the expert's function and also injects an active harm, so it should cost more than

Table 6: Every audited setting, so that the counts appear once. "ops" are paired probes, each on two control streams unless noted, and "nulls" are no-op draws contributing to that setting's band. Bands are the order statistic of Section 3.3 at τ=60/τ=180. With n draws the band is the rank-k order statistic for $k = \lfloor 0.9(n-1) \rfloor + 1$, which a fresh null draw exchangeable with them exceeds with probability $(n+1-k)/(n+1)$. That probability is 10% only for large n, 13% at n=46 and 29% at n=6, and these are reference values, since the pooled draws are heterogeneous across layers (Appendix B). "pos. ctrl" is the smallest injected defect the audit recovers in that setting, judged against that row's band. The WikiText low-rate control also clears its own cell's null down to $+0.0012$ (Section 6.1). Counting the nine candidates re-probed at $16\times$ evaluation power as separate results gives the 407 probe results over 395 distinct operations quoted in Section 1. Random candidates share the converged model's nulls. The OLMoE low-rate row's τ=540 applies to its four composites only, and Qwen's low-rate horizon is 120 steps at half the batch, token-matched to the others (Appendix C). The last column is the largest first-pass SU. Six of these nominally crossed their band, and none survives the all-stream re-probe of Table 1. † The coarse model's one positive, which clears the band on all six control streams and which Section 7 adjudicates.

setting (layers audited)	data	τ	ops	nulls	band	pos. ctrl	largest SU
Main audit, 10^{-4}*: 398 probe results, 392 distinct operations, 129 null draws, pooled band* 0.0040							
OLMoE-1B-7B (16)	WT	60/180	128	48	.0041/.0030	+.0067	+.0057
random candidates (4)	WT	60/180	28	*shared*	.0041/.0030	—	+.0005
OLMoE, 6 checkpoints × 4	WT	60/180	175	63	.0035/.0021	—	+.0036
Qwen1.5-MoE-A2.7B (5)	WT	60/180	35	10	.0039/.0022	—	+.0013
OLMoE-1B-7B (4)	mix	60/180	32	8	.0010	—	+.00002
Low-rate re-audit, 10^{-5}*: 101 operations, each setting under its own fired positive control*							
OLMoE-1B-7B (8)	WT	60/540	52	46	.0025	+.0059	+.0003
Qwen1.5-MoE-A2.7B (5)	WT	120	25	10	.0009	+.0019	+.0003
OLMoE-1B-7B (4)	mix	60	24	8	.00014	+.0012	+.00004
Coarse-grained model, 10^{-5} *(Section 7)*							
JetMoE-8B, 8 experts top-2 (3)	WT	60	9	46	.0012	+.048	+.0031†

Table 7: Amplitude positive control on the busiest expert of a converged layer (layer 8, expert 18, two paired seeds, band 0.0041). Damage is measured zero-shot, before either branch adapts, while SU is the post-adaptation paired difference. The ratio between them is the recovery fraction. It is not constant (7.4% at large damage, 1.7% at g=4), and Section 6.1 shows that it is not even a property of the defect, since the same g=8 damage recovers 7.4% at this adaptation rate and 75% at a ten-fold lower one. Sensitivity is therefore stated on the SU scale in Fig. 11, not in units of injected damage.

gain g	damage (nats)	SU(retire)	recovered	detected
16	0.436	+0.0314	7.2%	yes
8	0.091	+0.0067	7.4%	yes
4	0.017	+0.0003	1.7%	no
2	0.0030	−0.0009	—	no
1.5	0.0011	−0.0010	—	no

removing the function alone can buy. Retiring an expert already replaced by noise recovers 3.7% of the 0.0022 nats of damage when the noise is frozen and 31% when it can adapt (Fig. 8). And no amplitude-control recovery exceeds 7.4% of the injected damage (Table 7). One case breaks the assumption, and our own data show that it is not hypothetical. An expert whose trained routing is actively harmful gains from removal, while noise substitution need cost nothing at all. Three of the 24 experts we noise-destroyed show non-positive damage (Appendix C), so the quantity is not even bounded below by zero. Of the six experts for which we measured both, four have a larger measured retire SU than damage (layer 0's e_{21} scores $+0.0023$ against damage -0.00002), and all of them lie inside the band, which is exactly the regime the argument would have to exclude. Our clearest expert harmful on WikiText, JetMoE's layer-0 e_4 (Section 7), does satisfy the ordering with room to spare (0.0051 nats of zero-shot removal gain and SU $= +0.0031$ against 0.028 of destruction damage). Its

layer's dynamic range, however, is $22\times$ its own band, so it does not test the case that matters. In layer 8 we therefore do not separate instrument from model, and we do not claim that a retire gain there is impossible. Either way, the positive control validates only the retire channel, so for merge and composite we can state the measured floor but no analogous scale argument.

Calibrating the adaptation rate (Section 6.1). The rate sweep uses six control streams per cell. Adaptation at 10^{-4} moves the un-mutated model by 0.095 nats, ten to fifty times the dynamic range of a single expert depending on the layer, and the streams disagree with each other by 0.0064. At 10^{-5} the movement is 0.017 and the disagreement 0.0020. Most of that movement turns out not to be learning at all. When the same 10^{-5} probe runs on OLMoE's own training mixture instead of WikiText, the control branch drifts by 0.0003, fifty-eight times less than the 0.017 it drifts on WikiText, and the stream disagreement falls in step to 0.0002. The model has nothing left to gain on the distribution it was trained on, so the WikiText probe spends most of its adaptation budget moving toward WikiText. This is why the off-distribution setting, which we chose as the more neutral one, is in fact the noisier one. It also sharpens the earlier guidance, since the reason to probe on the deployment distribution is not only that usage statistics are estimated there (Section 5) but also that the estimator itself is quieter there.

The band at 10^{-5} (Section 6.1). The median disagreement between a candidate's two seeds falls from $\sim 10^{-4}$–10^{-3} to 1.4×10^{-4}, and in the quietest cells to 10^{-6}. The band falls too, but by less than a first 16-draw estimate suggested, and the difference matters enough that we describe how we learned it. Two draws per layer gave 0.0017 over 16 draws. Noise at this rate is strongly depth-dependent, with a layer-0 draw reaching 0.0043 where layer 12 stays at 0.0005, so 16 draws under-sample the tail badly, and their bootstrap interval of $[0.0009, 0.0043]$ is wide enough to be uninformative. Thirty further control-only draws put the pooled figure at 0.0025 over 46 draws, with interval $[0.0017, 0.0028]$. Re-run at the same rate, the amplitude positive control detects all four injected gains, down to $\mathrm{SU} = +0.0012$ against that cell's own null and $+0.0059$ against the pooled band. The quieter instrument is thus more sensitive and not merely rescaled, which is the check this claim requires.

Limits of the lower floor (Section 6.1). The floor has come down by a factor of 1.8, and whether that suffices depends on the layer. Against layer 8's 0.0018–0.0022 it is still 1.3–$1.6\times$ too coarse (it used to be 2.3–$2.8\times$), while against layer 0's 0.0052 and layer 15's 0.0100 it now sits comfortably below. The honest reading is therefore narrower than the one the 16-draw band would have licensed, and it depends on the layer. Lowering the rate does not give the pooled rule 80% power for an improvement the size of a small layer's expert, though where experts are large it now has that power for a retire gain of that size. It does make the audit roughly twice as sensitive, and across 52 operations at that sensitivity its rule flags nothing. Retire's best case anywhere is $+0.0003$, and 32 of 32 fall below the band. Merge is negative in every one of the eight layers, which at this resolution makes the harm unambiguous rather than marginal. Budget-neutral composites are probed at $\tau{=}540$ so that expansion is not starved of adaptation, and they average -0.0009 over four candidates with a best case of $+0.0002$.

Qwen and the on-distribution replication (Section 6.1). The Qwen pass uses token-matched batching, since its 152k-token vocabulary would otherwise not fit the probe on one card. Retire means are -0.0002 on Qwen and -0.0011 on-distribution. Qwen also shows the rate-dependence of repair on a real candidate rather than an injected defect. Its most functionally similar pair is layer 6's e_{37} and e_{40}, the 0.84-similarity pair of Section 5. Merging it costs -0.039 at 10^{-4} and -0.224 at 10^{-5}, so the same operation on the same pair comes out $5.7\times$ more damaging, with the two seeds 0.0012 apart. Consistent with the recovery-fraction result above, the higher rate had been repairing most of the merge's damage out of the adaptation budget, and the low-rate number is the closer estimate of what the merge itself destroys. We report the pair rather than the family mean because at both rates the mean (-0.023 and -0.0043) is that one candidate. Qwen's other nine merges have a median cost of -0.0003 at this rate (-0.0004 at 10^{-4}, Section 5), so the pair carries the family mean at both. The pair's measured similarity is itself sample-dependent (0.84 on the original probe sample, 0.93 on this one), the same estimator fragility we document for usage in Section 5.

Low rate with the step budget restored (Section 6.1). Three cells probe the regime this leaves open. A lower rate with a proportionally longer horizon restores the step budget in small steps. At 10^{-5} with $\tau{=}1800$, the same $\mathrm{lr} \times \tau$ product as 10^{-4} at $\tau{=}180$, layer 0 gives -0.0022 for the retire and -0.0024 for the composite, where the same operations at 10^{-4} gave $+0.0023$ and $+0.0042$.

The signs disagree between regimes, but the 10^{-4} values here are exactly the layer-0 crossings that Table 1 showed do not reproduce, so the disagreement is with an artifact rather than with a result. Two further cells instead match the $\mathrm{lr} \times \tau$ product of the headline τ=60 protocol (10^{-5} at τ=600) and cover layers 4 and 8 with three operations each. We chose layer 4 because it produced the audit's largest excursion, the $+0.009$ that did not reproduce. All six results are negative or sign-split, and none is positive on both seeds. The retires give -0.0003 and -0.0014, the merges $+0.0007$ (seeds $+0.0019$ and -0.0004) and -0.00004, and the composites -0.0017 and -0.0023, against null draws of 0.0052 and 0.0009. Across three cells at two matched horizons the restored-budget regime agrees with the τ=60 audit. One cell could not have shown that, which is why a single cell counts only as a probe of the regime.

Caveats of the low-rate audit (Section 6.1). A lower rate means less adaptation. The composites above were given nine times the contraction horizon to compensate, but that is still $3.3\times$ less total step budget than τ=180 at 10^{-4}. Their τ=540 cells also carry only one null draw each, one of which is -0.0094, so the composite bound here is looser than the retire and merge bounds beside it. The recovery fraction of Table 7 is also rate-dependent. The same g=8 defect that recovers 7.4% at 10^{-4} recovers 75% at 10^{-5}, because the surviving experts no longer have the step budget to route around it. Structural utility measured at a low rate is therefore closer to the un-repaired effect of an operation, and further from what a model would enjoy after real continued training. Which of the two a practitioner wants is a choice the protocol should expose rather than fix, and we report both.

Evaluation-set size and band width (Section 6). Before running these high-power probes, we predicted that $16\times$ the probe-validation tokens (3.15M, drawn from the validation split plus a document-disjoint slice of the training corpus) would shrink the band $\approx 4\times$ by $1/\sqrt{n}$. The measurements bear this out for the typical case and not for the tail. In a well-behaved cell (converged layer 8) the τ=60 null draws fall to 0.0008, a $5\times$ improvement and slightly better than the predicted $4\times$. At that five-fold finer resolution all three candidates remain at or below zero (-0.0001, -0.0006, -0.0005), which is the strongest form of our negative. In two other cells, however, a single adaptation-stream excursion produced per-seed values of $+0.033$ and $+0.014$. Each excursion inflated every candidate sharing that control identically (retire $+0.033$ and merge $+0.028$ in one cell, $+0.014$ and $+0.015$ in the other) and pushed null draws to 0.0080. The layer-7 0.0298 of Section 3.3 is a third instance of the same phenomenon, not the same event. It is a τ=60 null draw between two null-only control streams (4 and 5) of layer 7 in the main audit at 197k evaluation tokens, and the layer-7 candidates, on streams 0 and 1, carry no offset (0.00005). The $+0.033$, by contrast, is stream 1 of the $16\times$ re-probe of converged layer 0, and the $+0.014$ is stream 0 of the 2% layer-0 cell. Three unrelated control runs in three cells, at 197k and 3.15M evaluation tokens alike, thus each land 0.01–0.03 nats from their siblings. Evaluation-set size cannot damp this, because the residual noise is variance of the short adaptation trajectory itself, not sampling error of the loss estimate. Our practical guidance for future audits is therefore the opposite of our own advance prediction. Compute is better spent on more control streams and seeds than on a larger evaluation set, the same conclusion Bouthillier et al. (2021) reach for benchmark comparisons, where randomizing more sources of variation beats estimating any one of them more precisely. A sign-consistency rule should also be kept, since it is what disqualifies the four inflated candidates in those two cells (2/4, 1/4, 2/4 and 1/4 positive seeds).

Granularity measurements (Section 7). This part gives the grouped-destruction fits and the cautions that travel with them, the normalisation of routing shares, the two composite-trajectory follow-ups, the control-only band-versus-horizon run, and the coarse model's rate selection, remaining damages and control taxonomy.

Grouped destruction fits and cautions. Shares are summed usage on 49k tokens. These are sums over busy or over many experts, whose relative error is small where the 8k estimate of one quiet expert's share was not. The costs Section 7 quotes at shares 0.125 and 0.25 are medians of the measured group points within $\pm 28\%$ of the target share, and the per-layer fits agree (0.006, 0.009, 0.035 at 0.125 and 0.018, 0.018, 0.072 at 0.25). Two cautions travel with the prediction. One is that a single busy expert costs more than a random group summing to the same share (single-expert exponents run 1.25–1.88 in the same layers), so within-router share carries a specialisation component that an architecturally fixed share need not. In a synthetic model where granularity alone is varied (Section 4), though, that component vanishes, with a coarse expert costing $1.14\times$ a fine group of equal share, and no edit improves the coarse models either. The other is that the figure is the cost of corrupting an eighth of OLMoE's routing, not a measurement of a coarse model, whose experts may

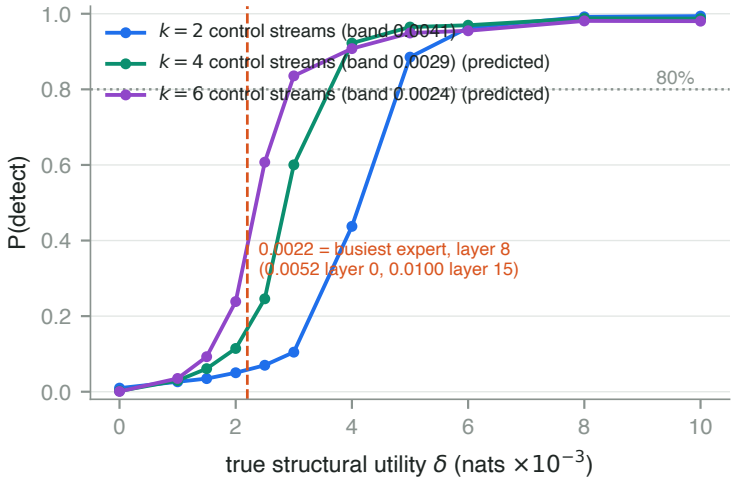


Figure 11: Power of the decision rule (mean above band and all seeds positive), bootstrapped from the audit's own centred per-seed noise and shifted by a true effect δ. The noise sample is 922 values with s.d. 0.0025, the 398 candidates' per-seed values less the 56 belonging to the random-candidate family, which is a spot check rather than part of the audited set. Curves are for k control streams per candidate, and the band narrows with k as the null does. At our k=2, 80% power arrives at $\delta = 0.0050$ nats, above the 0.0018–0.0022 nats that destroying layer 8's busiest expert produces (dashed), which is why the pooled protocol is underpowered for a retire gain on that scale in layers whose experts are that small.

be more or less redundant than that. It is a hypothesis for a coarse model to test. Extrapolating the single-expert fit instead gives a bootstrap interval of 0.003 to 0.08 at the single-expert share (a factor of 28, against 0.005 to 0.36 at 0.25), which the grouped measurement narrows to 0.006–0.035.

Normalisation of routing shares (Section 7). Shares are normalised so that a layer's experts sum to one. OLMoE's uniform share is $1/64$=0.0156 and a top-2-of-8 model's is $1/8$=0.125, so one expert of such a model carries a share of $1/8$, and $1/4$, the second share Section 7 quotes, is what the two experts its router activates carry between them. Both costs are medians of the measured group points, which makes them interpolation inside the audited range of grouped shares rather than extrapolation beyond it. The ladder itself is measured inside a single model, across layers 0, 8 and 15 of OLMoE.

Re-initialized children at longer horizons. Section 5 showed that partially re-initialized split children genuinely differentiate and still score negative at τ=540. The natural rejoinder is that a differentiated child is worse at first and better later, so for the re-initialized composite we have the whole trajectory rather than its endpoint. Its SU runs -0.0006, $+0.0036$, -0.0009 in layer 0 and $+0.0002$, -0.0010, -0.00002 in layer 8 across $\tau \in \{60, 180, 540\}$, with no trend toward zero. One point, layer 0 at τ=180, nominally exceeds the τ=180 band of 0.0030. Its seeds are $+0.0074$ and -0.0002, so under the decision rule it is sign-split and does not fire (the τ=540 point in the same layer is likewise split, -0.0029 and $+0.0011$). The only cell positive on both seeds is worth $+0.0002$, five per cent of the band. The τ=540 points are read against a band whose tail is heavier than the τ=60 estimate (next paragraph), which only widens the margin by which they fail. A third initialisation, gradient-directed growth in the sense of GradMax (Evci et al., 2022), gives the same picture. Across $\tau \in \{60, 180, 540\}$ the composite scores -0.0001, $+0.0018$ and $+0.0008$ in layer 0 with the seeds split at every horizon, and -0.0005, $+0.0012$ and -0.0007 in layer 8 with the seeds agreeing and the largest at 0.4 of the band. No cell fires, and the children stay differentiated (correlation 0.59–0.72, layer-8 routing overlap falling from 0.85 to 0.65). Three ways of starting a child (a perturbed copy, a partial re-initialization, and a start chosen to maximise its own gradient) end in the same place. The children stay differentiated the whole way, with output correlation 0.67 and 0.52, flat in τ, and in layer 8 the router keeps separating them (overlap falls from 0.79 to 0.46). A ninefold horizon range therefore does not turn differentiation into utility.

Null tail and probe horizon. Bands are estimated from τ=60 draws and then used to read trajectory points at τ=180 and 540. A control-only run with six streams at each horizon in layers 0 and 8, and no operations, shows that this is not innocuous. The centre of the null distribution barely moves (median |draw| 0.0012, 0.0017, 0.0019), but its tail grows. Of twelve control branches per horizon, none lands more than 0.01 from its cell median at τ=60 or 180, while two do at τ=540 (largest deviation 0.0078, 0.0041, 0.0224). Long-horizon SU carries more tail risk than a τ=60 band represents. The direction favours our conclusion, since a wider true band at τ=540 leaves every candidate that failed to exceed the drawn band still failing. We therefore report the effect rather than revise verdicts, but an audit reporting a positive at long horizon would need its own band there. The same run does not reproduce the main audit's ordering of the τ=60 and τ=180 bands (0.0041 against 0.0030). Restricted to these two layers it gives the opposite direction, and neither sample (16 draws there, 6 here) can resolve a 90th-percentile statistic, so that ordering should not be read as a property of the horizon.

Coarse-model rate selection and damages. JetMoE-8B is audited at the rate its own frozen amplitude control selects. At 10^{-5} the control drift is -0.011, and a g=4 defect is detected at SU $= +0.048$ with 86% recovered, while g=2 is not. At 10^{-4} the drift is -0.102. The median expert of layers 12, 23 and 0 costs 0.004, 0.031 and 0.029 nats to destroy, and the busiest pair, at a share near 0.3, costs 0.017, 0.099 and 0.110. Read against the per-layer group fit of the depth-matched OLMoE layer (8, 15 and 0) at each point's own share, the busiest expert costs 1.1, 0.8 and 2.9 times the prediction and the busiest pair 0.8, 1.2 and 4.7 times. The residual scatter of those fits is 0.6–1.5, 0.35–3.6 and 0.4–2.2× (Fig. 1). The prediction holds in direction everywhere and in size in layers 12 and 23. In layer 0 the busiest pair is 4.7× the fit, outside its group scatter, while the busiest single expert (2.9×) sits below OLMoE layer 0's own busiest expert, which is 4.1× the same fit. At nearly the same share, 0.10 to 0.17, the 24 single experts span 0.0010 to 0.096 nats, and three of them fall below the 0.0028 floor. Share thus sets the order of magnitude across granularity, while within a model the layer and the expert matter more. Against each layer's own band the eight resolved candidates run 0.9–61×, and only layer 12's quietest retire, -0.0013 against a local band of 0.0016, fails to clear its own layer's band. The OLMoE-mix proxy cell has a band of 0.00022 over eight draws, so the layer-0 retire's -0.0018 there is also 4.5× the largest single draw, with both placebos within a band of zero.

Coarse-audit verdict and control taxonomy. At coarse granularity the audit resolves a layer's busiest expert with room to spare, and the held-out and training-data tests decide whether a positive transfers. The layer-0 candidate improves on the un-mutated model on the validation slice but not on the held-out test split, and on RefinedWeb and the OLMoE-mix proxy it is a loss (Table 5). Because a placebo inert at zero-shot is not inert under adaptation (Section 7), our controls fall into three kinds rather than two. A no-op pair is a run-noise control, and its utility is zero in expectation. A permutation is a symmetry control. It commutes with Adam's element-wise update, so it is a symmetry of the whole adaptation trajectory and not only of the forward pass, and its cell mean stays inside its cell's band in every cell we measure it. Rescaling an expert's up rows by c and its down projection by $1/c$ is an *adaptation-speed* control. It preserves the function but not the trajectory, because it changes how far Adam's steps move each projection relative to its size. On WikiText its effect grows with the factor on the harmful expert and, about a third as large, on one whose removal costs loss ($+0.0009$, $+0.0028$ and $+0.0126$ at $c=\frac{1}{2}$, 2 and 4 for e_4, and $+0.0009$ and $+0.0048$ at c=2 and 4 for e_1). On RefinedWeb, at c=2, it is nil. In the OLMoE cells at c=2 it stays within noise, at $+0.0003$ and -0.0014 in Table 1, $+0.000005$ and $+0.000048$ in Table 4, and -0.000008 to $+0.000037$ in its second run. Only the first two kinds measure noise. The third measures how much faster adaptation toward the probe text buys. A retire-minus-rescale difference at one factor therefore depends on the factor chosen and cannot say whether a gain needs removal. For the layer-0 retire at c=2 that difference is $+0.0001$ in the first re-probe and $+0.0002$ in the run of Table 5 (95% t intervals -0.0005 to $+0.0006$ and -0.0001 to $+0.0005$). Whether the gain transfers is decided instead by the held-out and training-data tests.

Scope limits of Section 8. Section 8 mentions each of these only as a clause. Three of them follow from how the operations are implemented rather than from the models.

Distribution. JetMoE's layer-0 retire is a reproducible $+0.0031$ on WikiText's validation slice, -0.0001 on its test split and -0.0036 on RefinedWeb. The two OLMoE retires of Table 4 hold on the test split and cost -0.00035 and -0.00091 on OLMoE's training mixture. An audit run on one probe distribution alone would report gains that do not transfer.

The merge rule. Our merge is deliberately the simplest rule, usage-weighted parameter interpolation, applied to the most functionally similar pair available in each layer. Those pairs reach output correlation up to 0.76 on OLMoE and 0.84 on Qwen (median 0.09 and 0.06), yet their weights are essentially orthogonal (cosine ≈ 0.003). "Merging hurts" is therefore a statement about interpolating weight-unaligned experts, not about the existence of merge slack. The two obvious repairs do not change the sign. An alignment-aware merge is indistinguishable from the naive one, and retire-and-reroute is gentler but still negative (Section 5). What remains open is a stronger merge rule than these two, such as routing-similarity grouping (Li et al., 2024b) or Nash-bargaining merges (Nguyen et al., 2026).

Split children. Split children are perturbed copies of their parent (σ $0.02\times$ the parent's weight s.d.), and we measure that they barely differentiate within $\tau \leq 540$ (Section 5). The router separates them, but their outputs stay 0.994-correlated. Composite neutrality therefore bounds perturbed-copy, partially re-initialized and gradient-directed (GradMax (Evci et al., 2022)) expansion at $\tau \leq 540$, and 'Re-initialized children at longer horizons' reports the full $\tau \in \{60, 180, 540\}$ trajectory of the latter two. We have not tried a child chosen by a criterion other than noise, the parent's weight statistics or the parent's own gradient, such as a learned or router-aware initialisation.

Horizons and re-probes. Horizons past 540 are probed only at the ten-fold lower rate (Section 6.1), where three cells at $\tau{=}600$ and $\tau{=}1800$ carry retires and merges as well as composites. The control-stream offset of Section 5 is resolved by re-probing rather than by argument (Table 1), but only for the six candidates we re-probed, not for every cell on every stream.

Cost and reach of the controls. Paired multi-stream controls with explicit nulls and sign-consistency rules cost about twice what single-branch probing costs, and we document every one of the six apparent findings they overturn. Pruning remains budget-positive, and our random-candidate contrast shows the statistics genuinely support it.

Candidate generation. Usage comes from an 8k-token sample, which is too small on a heterogeneous mixture and, at any size, blind to the sources it omits. This limits the candidate generator, not the probes.

Band construction. Bands estimated at $\tau{=}60$ are used to read longer-horizon points, whose growing null tail ('Null tail and probe horizon') displaces no verdict.

The coarse model's on-distribution probe. JetMoE's mixture is assembled from public datasets that we did not reassemble (in the first phase, 64% RefinedWeb, 25% StarCoder GitHub and 8% Pile). We probe RefinedWeb directly and, earlier, the web text of OLMoE's mixture. Both agree in sign, and neither covers the code share.

Two falsifiers. The first is an amplitude-based positive control at a gain the audit detects here ($g{\geq}8$) that it then fails to detect, which would indict the instrument rather than the models. The second is more upward firings than the audit's own noise predicts, which the bootstrap of Section 5 would show directly.

Audited scope and closing figures (Section 9). The audit covers 407 probe results, two model families, seven checkpoints, heuristic and random candidates, and adaptation horizons of 60 and 180 steps, extended to 540 where expansion needed it. Across all of it, every family *mean* is bounded above at 95% by $+0.0008$ nats save one, and the largest single SU of all 398 is $+0.0057$, back to $+0.0008$ on re-probe. Nominal crossings arrive at the rate the audit's own noise predicts. A power analysis states what size of effect the audit would have detected, which is evidence against a large maximum but not a bound on one. One expert's whole dynamic range runs from 0.0018 in layer 8 to 0.0100 in layer 15. At our original rate the floor was 0.0050, above the smallest of those, so in layers of that kind the negative said more about the instrument than the model. Lowering the rate brings the floor to 0.0028 and the band to 0.0025, and the rule flags none of 101 operations in three settings. A re-test does, however, reproduce two $\sim 10^{-4}$-nat retire gains on WikiText, held-out text included, and these reverse on the training mixture. Low-rate merge means sit at -0.0021 on OLMoE and -0.0022 on-distribution, and composite seeds agree to 10^{-4}. That single experts have so little range is consistent with a granularity account of these architectures, which Section 7 quantifies. Corrupting an eighth of a layer's routing, which is what one expert of a top-2-of-8 model carries when shares are normalised over a layer's experts, costs 0.006–0.035 nats, two to thirteen times our best floor, so such a model is predicted to be auditable at this sensitivity. In JetMoE-8B, with eight experts and

top-2 routing, a layer's busiest expert is worth 11–26 times the model's own band. Retiring layer 0's busiest expert, a gain on WikiText's validation slice, gives nothing on its test split and costs 0.0036 on RefinedWeb, 14 times that cell's band (0.0018 on the web text of OLMoE's mixture), with its zero-shot sign reversed.

The coarse-model verdict (Section 9). Eight of the nine audited JetMoE-8B edits are negative beyond the band, six of them by $2.5\times$ or more. The ninth is a positive under Eq. (1) on WikiText's validation slice that reproduces on all six control streams, is -0.0001 on the held-out test split (95% interval -0.0004 to $+0.0002$), and reverses on RefinedWeb (-0.0036, all six streams negative) as on the OLMoE-mix proxy (-0.0018). The rescaling that matched it on the validation slice is an adaptation-speed effect that grows with its factor and appears for an expert whose removal costs loss too (Table 5). The dynamic range the closing argument turns on is what replacing a layer's busiest expert with noise costs, and the retire gain it is compared against is one the size of a single expert.