MoE Mirage: Selection Noise and Routing-Weight Geometry in Token-Level Counterfactuals

Anonymous authors
Paper under double-blind review

Abstract

Token-level "best expert" headroom in sparse mixture-of-experts (MoE) models is a same-sample best-of-K maximum: each candidate expert is installed, the downstream loss is scored, and the best is kept, on the same tokens. We show that, at the resolution we can test, this quantity behaves as selection noise whose scale is set by the intervention operator. Selected on one half of the continuation and scored on the other, no operator retains headroom above the experiment's resolution (identity-preserving swap: -0.0002 nats, 95% CI $[-0.0006, +0.0002]$; planted advantages detected with 80% power from 0.0021), and an exact matched-noise null closely reproduces the primary same-sample values and the checkpoint ordering of their ratio. Published operators change support, weight shape, total mass or normalization together with the expert; we introduce an **identity-preserving swap** that changes only the expert, and full-commitment forcing reports more same-sample headroom than it in 199 of 202 cells ($\hat{\kappa} = 5.48$), as the null predicts. The operator also changes a decision that takes no maximum: one-shot pruning by mean expert effect yields operator-dependent masks, forcing's ranking partly reproduces across document halves while the swap's does not, and on two OLMoE releases forcing-ranked masks fall below random masks in benchmark accuracy. We recommend that token-level routing maps state their operator and scoring horizon, cross-fit candidate selection and scoring, and report a matched-noise null to separate persistent opportunity from selection noise.

1 Introduction

Counterfactual routing analysis asks which expert should have served a token, and its answers are used as a decision map: which layer leaves routing opportunity unused, which expert to prune, which router to retrain. The reported quantity is a maximum: install each candidate expert, score the downstream loss, keep the best. **Three mechanisms shape that maximum, and none of them requires a better expert to exist.** The intervention sets how strongly a candidate perturbs the output; selecting and scoring on the same tokens turns that perturbation into positive apparent headroom; and the winner for the next token does not persist over the continuation. A token-level routing oracle should therefore not be read as detectable persistent opportunity unless the candidate is selected on independent outcomes and its gain exceeds a matched selection-noise null.

At a top-m MoE layer a token's routing state is $R = (S, \tilde{w}, M, N)$: the *support* S (which experts are selected), the *weight shape* $\tilde{w}$ (the relative routed weights on S), the *total mass* $M = \sum_{j \in S} w_j$, and the *normalization semantics* N (whether selected weights are renormalized). Five interventions are in current use. Recent MoE work zeroes an expert's output (Engmann et al., 2026), ablates it with proportional renormalization (Hu et al., 2026), resamples fixed-k route sets (Yoon et al., 2026), or forces a gate logit into the top-m set (Bandarkar et al., 2026; Fayyaz et al., 2026). The fifth is full commitment of the routed mass to one expert, the literal reading of "force token t through expert e." Each changes expert identity together with at least one coordinate of R (Table 1). None of these works compares conventions, and, to our knowledge, none includes a control that changes identity alone; the closest, a weight-nearest control expert in expert-level patching (Lu et al., 2026), matches router weight approximately by selection rather than holding it fixed by construction.

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

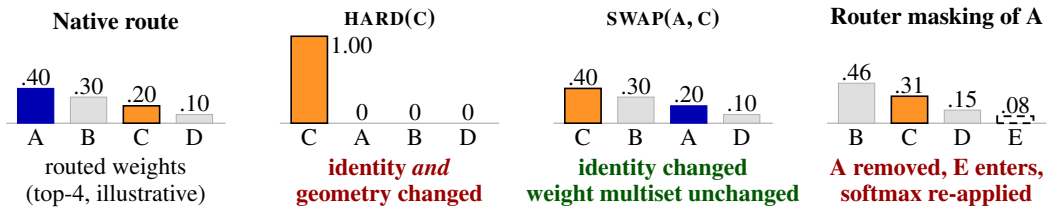

Figure 1: **What each operator changes.** Weights are illustrative; bars are in rank order. Forcing C (HARD) collapses the routed weights onto one expert and, on non-renormalizing routers, also changes the total routed mass. The identity-preserving swap trades the roles of A and C: identity changes, the routed-weight multiset does not. Router masking of A removes A and re-applies the softmax, so the next-ranked expert E enters: two identities and the weight shape change.

We therefore introduce the **identity-preserving swap** (Figure 1), a token-local expert transposition that exchanges the gate logits of the candidate and the native top-1 expert before the model's native routing runs. For an in-committee candidate, support size, the routed-weight multiset, total mass and normalization are unchanged; only the assignment of expert identities to routed roles changes. This differs from function-preserving MoE permutations, which relabel experts and gates jointly (Tran et al., 2025), and from fixed-routing causal tracing, which restores expert computations without exchanging routed roles (Lu et al., 2026). On 7 architectures, full-commitment forcing reports more headroom than the swap in **199 of 202 matched cells**, with $\hat{\kappa} = 5.48$ over the full causal suffix. The published conventions also disagree with each other: on one matched grid, ablate-and-renormalize reports $0.64\times$ the headroom of fixed-k resampling.

Chosen on one half of the suffix and scored on the other, no operator retains headroom detectable at the experiment's resolution, and an exact matched-noise null closely reproduces $\hat{\kappa}$ checkpoint by checkpoint (Section 5). Cross-fitting removes selection optimism but cannot say which routing-state coordinate set the scale; the swap isolates that coordinate but, scored on the same sample, still reports a maximum. Both are needed.

Why this is not only the winner's curse. That a maximum over noisy estimates is optimistic is textbook. What is specific to MoE analysis is that token-level counterfactual routing reports exactly this maximum, the intervention operator sets how large the artifact is, and the operator also changes mean-based pruning rankings that take no maximum. The issue can arise in any modular system scored by best-of-K counterfactuals; sparse MoE separates it cleanly because the routing state is explicit.

Contributions.

1. **Control.** We write each convention as a map on the routing state (support, weight shape, total mass, normalization) and introduce an identity-preserving swap that, to our knowledge for the first time in MoE counterfactual analysis, changes expert identity while holding support size, the routed-weight multiset, total mass and normalization exactly fixed (Section 3).
2. **Mirage.** Same-sample headroom is positive under every operator, while cross-fitted headroom is below detectable resolution and a matched-noise null closely reproduces the same-sample values and their checkpoint ordering (Section 5).
3. **Mechanism.** Weight shape and, on non-renormalizing routers, total mass widen the per-candidate effect distribution without moving its centre, and the maximum turns that width into reported headroom (Section 6).
4. **Consequence.** Pruning by mean expert effect, which uses no maximum, still depends on the operator: forcing and the swap induce different, differently reproducible rankings, and on both OLMoE releases forcing-ranked masks fall below the random-mask median in benchmark accuracy. The effect is checkpoint-dependent and no criterion we test reliably beats random masks on accuracy, so this shows that operator choice matters, not that the swap finds better experts (Section 7).
5. **Standard.** A reporting checklist that separates operator, selection, scoring and numerical resolution (Section 9).

2 Related Work

MoE counterfactual metrics. Engmann et al. (2026) audit whether observational routing statistics predict interventional expert importance and find they largely do not; their axis is observation to intervention, ours intervention to intervention. Their estimand and the Counterfactual Expert Impact of Hu et al. (2026) are means over tokens for a fixed expert. Ours is a maximum over alternative experts at a fixed token, which is dispersion-sensitive in a way a mean is not. The published ranking conclusion we re-test, that router probability barely predicts interventional importance, holds under all 5 operators (Appendix D.3). Yoon et al. (2026) resample equal-compute route sets, recomputing the gate weights, and never commit a token fully to one expert, and Bandarkar et al. (2026) raise a gate logit only until the expert is selected. Li (2026) decouples dispatch from aggregation and reports that the router's own weights are rarely the loss-optimal mixture over the experts it selected; that intervention holds expert identity, expert computation and total selected mass fixed and varies only the weights, a geometry-only intervention in our terms and the dual of the swap, which holds the weight multiset fixed and varies only identity. Pruning methods allocate expert budgets from per-layer importance scores (Liu et al., 2026b;a; Li et al., 2026); Section 7 tests what our operators imply for that decision.

Ablation sensitivity and permutation controls. For dense components, zero ablation overstates optimal-ablation effects by a median factor of about nine (Li & Janson, 2024), switching the ablation method can reverse circuit-method rankings (as noted by Seth et al., 2026), and apparent expert modularity moves with metric and corpus (Salomone et al., 2026). The spread we measure is comparable, not larger; what differs is the object: a dense component has no routed mass, so an identity-isolating control has no counterpart there. Permuting units is an established symmetry primitive (Entezari et al., 2022; Ainsworth et al., 2023); in an MoE, permuting router rows and expert tensors jointly leaves the layer's function exactly unchanged (Tran et al., 2025; Wood & Stringer, 2026). The swap permutes the router output alone, at a single token, which is the deliberate asymmetry. Fixed-routing causal tracing is the closer analogue. Activation patching (Meng et al., 2022) restores a component's output after routing has run, so it does not feed back into routing: it asks what an expert's computation contributes under the observed route, not what changes when another expert fills that routed role. Lu et al. (2026) come closest: to rule out router weight as a confound in expert-level patching under fixed routing, they choose as control the co-active expert whose clean router weight is nearest the target's, and separately equalize patch norms. That matching is approximate and achieved by selection; a transposition exchanges the two experts' roles, so the routed-weight multiset is invariant by construction. Two further operators touch identity without isolating it: in a hardware-scheduling study, Huang et al. (2026) replace a selected expert with a near-boundary unselected one that keeps its own gate weight, changing support membership and the weight multiset together, and Gu (2026) pins a clean route set with its clean weights, an interchange intervention that transplants the whole routing state rather than one coordinate of it. To our knowledge, no prior work varies expert identity with the native routed-weight multiset exactly preserved.

Coalitional attribution and oracle inflation. Shapley-style values (Shapley, 1953; Lundberg & Lee, 2017), applied to MoE experts by Zhang (2026b), average over which *other* experts are present, so they are insensitive to the incumbent's identity at one token; they still need a rule for what removing a player does to the routing state, so they inherit the same conventions. Oracle headroom that does not convert into realizable gain recurs elsewhere (Chen, 2026; Garg & Sagtani, 2026; Zhang, 2026a; Wei et al., 2026; Sugumar & Gopinath, 2026); closest is Chen (2026), who shows that router-benchmark oracles built from single correctness draws sit above a reproducible oracle, and to whom the separation of action-class limits from information limits is due; we show the same artifact inside the model, where the intervention sets its scale.

3 What a Counterfactual Holds Fixed

Fix an MoE layer ℓ with E experts and top-m routing. For token t the gate produces logits $r_t \in \mathbb{R}^E$, a selected set S_t with $|S_t| = m$, and routed weights $w_t \in \mathbb{R}^m_{\geq 0}$ over S_t. Let $u_\ell(\cdot)$ be the downstream utility, the negative teacher-forced loss of the suffix after layer ℓ. An intervention is a map $\mathcal{I} : (S_t, w_t, e) \mapsto (S'_t, w'_t)$ expressing "involve expert e." Table 1 lists what each convention does to the routing state.

Table 1: Intervention operators by the routing-state coordinates they hold fixed (✓) or change (✗). Every operator changes expert identity; only the swap changes nothing else. Rows are stated for an in-committee candidate $e \in S_t$, the primary population throughout. ‡For $e \in S_t$, support cardinality and membership are invariant (two members exchange roles); for $e \notin S_t$, cardinality is preserved and one index is exchanged (Appendix C.3). †Mass is preserved on routers that renormalize the selected weights (Granite, Qwen3, Mixtral) and not on OLMoE or Qwen1.5-MoE. §As a convention, departs from the native rule on routers whose rule differs (zeroing on renormalizing routers; explicit rescaling on non-renormalizing ones); in every measured row the router's own rule computes the weights (Appendix C.1). Router masking is also the pruning mechanism of Section 7; unlike ablate-and-renormalize it admits the next-ranked expert, so two identities change (Appendix A.1). Sources: zeroing (Engmann et al., 2026), ablate and renormalize (Hu et al., 2026), router masking (Salomone et al., 2026), route resampling (Yoon et al., 2026), logit forcing (Bandarkar et al., 2026).

Convention	Operation	Support	Weight shape	Total mass	Normal-ization	Identity only
zero, no renormalization	$w_{t,e} \leftarrow 0$, rest fixed	✗	✗	✗	✗§	✗
ablate and renormalize	drop e, rescale survivors	✗	✗	✓†	✗§	✗
router masking	$r_{t,e} \leftarrow -\infty$ before top-m	✗	✗	✓†	✓	✗
route resample	resample S'_t with $\lvert S'_t \rvert = m$	✗	✗	✓†	✓	✗
logit force	raise $r_{t,e}$ into the top-m	✗	✗	✓†	✓	✗
hard (full commitment)	all routed weight to e	✗	✗	✓†	✓	✗
hard, native mass	full commitment at native mass	✗	✗	✓	✓	✗
identity-preserving swap	exchange logits of e and top-1	✓‡	✓	✓	✓	✓

Definition 1 (Same-sample identity-preserving and full-commitment headroom)**.** *For operator $\mathcal{I}_A$ let $r_A = \mathcal{I}_A(S_t, w_t, \text{native}_t)$ be that operator applied naming the* native *expert; then $H^A_{\ell,t} = \max_{e \in \mathcal{K}_t} u_\ell(\mathcal{I}_A(S_t, w_t, e)) - u_\ell(r_A)$. For the swap, $\mathcal{I}_{\text{swap}}(S_t, w_t, \text{native}_t) = (S_t, w_t)$ exactly, so the reference is the unintervened route and its delta is identically zero: H^{swap} is the reported consequence of reassigning which expert serves the token with the routed-weight multiset fixed, selected and scored on the same tokens. For $\mathcal{I}_{\text{hard}}$ the reference is the native expert at full routed mass, a real intervention, and subtracting it removes that operator's candidate-independent offset. Cross-fitted counterparts (Section 5) test whether any survives.*

H^{hard} concerns what a single expert at full weight could have done, a question about capacity; H^{swap} concerns a different routing decision with the weights held fixed, the question the router-suboptimality literature intends. "Best expert" always means the candidate maximizing a stated utility under a stated intervention. The swap reassigns two roles, so its estimand is "what if e and the incumbent traded places." No operator can move exactly one index while holding the weight multiset fixed (Appendix A.1).

The swap is a control, not a ground truth. The swap is a diagnostic control and baseline, not an inference-time method. A planted-importance benchmark cannot adjudicate between operators: whichever intervention matches the generative form of the planted target wins by definition, as two planted constructions show (Appendix A.4). Because there is no intervention-independent scalar importance, the swap is justified methodologically by what it uniquely preserves: it isolates candidate identity from routing-geometry perturbation.

Remark on intervention semantics. The swap is not activation patching. Patching restores or replaces an expert's *output* after routing has run, so it leaves the routing decision and its weights untouched and does not feed back into routing. The swap acts one step earlier: transposing two gate logits before top-m and softmax makes the router itself compute the same weight distribution, reassigning only which index receives which weight. On fused routers that select inside the kernel (DeepSeek-MoE) the same statement holds for the dispatched `(index, weight)` tuple, whose weights are left untouched (Appendix A.1).

The identity-only action class. Write $\mathcal{A}_{\text{id}} = \{\mathcal{I}_{\text{swap}}(\cdot, \cdot, e) : e \in \mathcal{K}_t\}$: the interventions reachable by changing which expert receives a given routed weight, with the multiset $\{w_{t,j}\}$ fixed. For a rule g of any hypothesis class and observables, its recovery fraction against a reported headroom H is

$\mathcal{F}(g \mid H) = \mathbb{E}_t[u_\ell(g(t)) - u_\ell(\text{native}_t)] \,/\, \mathbb{E}_t[H_{\ell,t}]$. Throughout, R is the routing state, $\mathcal{F}$ a recovery fraction and ρ a Spearman correlation.

Remark 1 (Action-class ceiling)**.** *For any rule g with action class $\mathcal{A}_{\text{id}}$,*

$$\mathcal{F}\big(g \mid H^{\text{hard}}\big) \;\le\; \mathbb{E}_t\Big[H^{\text{swap}}_{\ell,t}\Big] \,/ \mathbb{E}_t\big[H^{\text{hard}}_{\ell,t}\big] \;=\; 1/\kappa, \qquad \kappa \;:=\; \mathbb{E}_t[H^{\text{hard}}_{\ell,t}] / \mathbb{E}_t[H^{\text{swap}}_{\ell,t}], \tag{1}$$

independently of the observables available to g, its hypothesis class and its supervision.

The bound is elementary (proof in Appendix A.2); its content lies in the measured $\hat{\kappa}$ and in what cross-fitting does to it (Section 5). It is a ceiling on *reported* same-sample forced headroom for identity-only selectors, not a claim that router optimization is futile: at $\hat{\kappa} = 5.48$ an identity-only rule can recover at most 18.2% of that reported quantity, and after cross-fitting both numerator and denominator fall below detection. It binds this action class only: router retraining, pruning, and changes to gate temperature or to m move the weights and fall outside it.

4 Experimental Setup

Models. We use 7 open-weight sparse-MoE architectures with different gate implementations. They are **OLMoE-1B-7B** (Muennighoff et al., 2024) (64 experts, top-8; the 0924 and 0125 releases), **Qwen1.5-MoE-A2.7B** (Qwen Team, 2024) (60 routed plus shared experts, top-4) and **DeepSeek-MoE-16B-base** (Dai et al., 2024) (64 routed plus 2 shared, top-6). The others are **Granite-3.0-3B-A800M-base** (IBM Granite Team, 2024) (40 experts, top-8), **Qwen3-30B-A3B-Instruct-2507** (Qwen Team, 2025) (128 experts, top-8; the only instruction-tuned checkpoint), and **Mixtral-8x7B** and **Mixtral-8x22B** (Jiang et al., 2024) (8 experts, top-2; the most concentrated routers here). Shared experts are unchanged between matched intervention arms, so they are not directly manipulated; nonlinear downstream interactions mean they need not cancel algebraically from the utility difference.

Regimes and protocol. Seven regimes are grounded in named public benchmarks with recorded revision and content hashes: WikiText-103 prose, HumanEval, GSM8K, XNLI in six languages, LAMBADA, and two perturbations of WikiText-103. Models are frozen. At layer ℓ and token t we intervene with candidate e, rerun only the suffix after ℓ, and score teacher-forced downstream NLL. The candidate set is the model's own committee: $K'{=}4 \le m$ on every top-4, top-6 and top-8 checkpoint, so $\mathcal{K}_t \subseteq S_t$. Mixtral routes top-2, so its two committee members are joined by two uniformly drawn outside experts. Both arms are scored on the matched site set of each cell, and intervals are clustered as stated with each number.

Instrument validation. Ten failure checks run before any number is admitted: the intervention changes routing, downstream loss responds, top-m logic is intact, targets are aligned, cached-suffix reruns equal full reruns, and revisions are pinned. A planted positive control is recovered ($\rho = +0.641$ swap, $+0.731$ hard) while a row-permuted negative arm loses its routing value, and an admission gate excludes any cell whose intervention does not bind (Appendices A.3 and A.4).

5 What Reported Headroom Measures

Reported headroom selects the best candidate and scores it on the same tokens. We first test whether any of it survives independent scoring, then read the same-sample magnitudes in that light. *Notation.* $\hat{\kappa}$ is the forcing-to-swap ratio of summed cell-mean headroom over all 202 admitted cells; $\hat{\kappa}_{115}$ is that ratio on the 115-cell cross-fit population, $\hat{\kappa}_{41}$ on the strict convention grid of Table 4, and $\hat{\kappa}_{\text{site}}$ pools sites rather than cells. Medians of cellwise ratios are named in words.

Cross-fitted headroom. We choose the candidate on one interleaved half of each site's suffix and score it on the other, $H^{\text{CF}} = \frac{1}{2}\big[u_B(e^*_A) - u_B(\text{ref}) + u_A(e^*_B) - u_A(\text{ref})\big]$, over the 115 cells scored under both forcing and the swap (53,863 sites, 7 of the 8 checkpoints, as Mixtral-8x22B has no split-half rows; Figure 2), the common matched population of every forcing-versus-swap comparison; each operator alone has split-half rows in more cells (119 forcing, 116 swap), reported separately in Table 3. The halves have equal length and the same distribution over causal distance, and cost

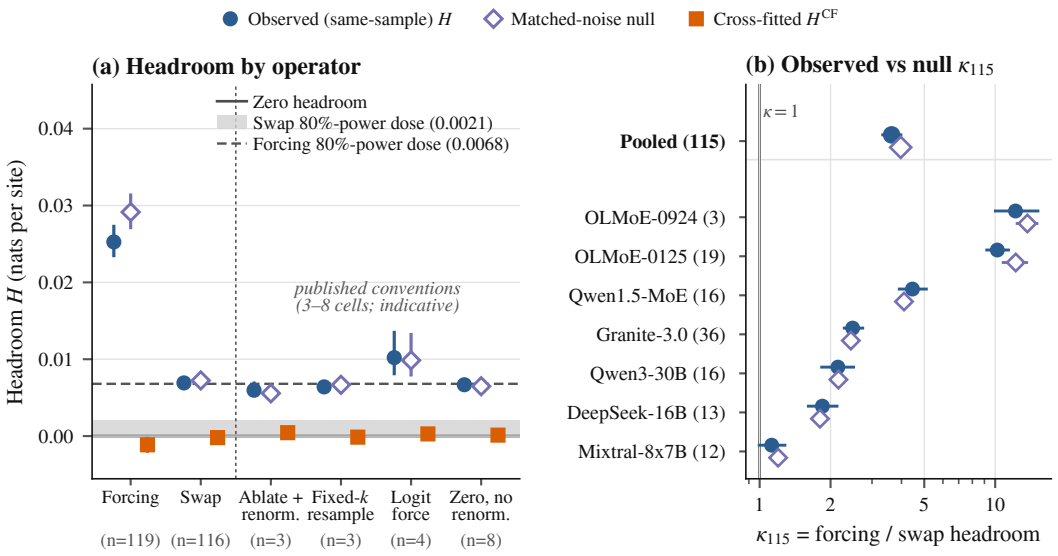

Figure 2: **Reported token-level headroom is a scale-dependent maximum.** (a) Per intervention operator (cell counts n under each label; 7 checkpoints): same-sample headroom, the matched-noise null (no persistent candidate advantage by construction), and cross-fitted headroom (selected on one half of the continuation, scored on the other), with 95% bootstrap intervals clustered by document (all of its sites, in every layer, resampled together). The grey band and dashed line mark the swap (0.0021) and forcing (0.0068) 80%-power detection floors; published conventions were collected on few cells and are indicative only. The null closely reproduces the same-sample values and the checkpoint ordering; no persistent advantage is detected above the floors. (b) Per checkpoint, $\hat{\kappa}_{115}$ observed and under the same null (log scale; pooled 3.63 observed, 3.97 null).

no extra forward pass. Same-sample headroom on the half-length data is 0.0300 nats (forcing) and 0.0092 (swap); cross-fitted it is -0.0011 $[-0.0022, -0.0001]$ and -0.0002 $[-0.0006, +0.0002]$, and the four published conventions behave the same way on the few cells collected for each (3, 3, 4 and 8). A preregistered power analysis detects a persistent advantage planted in one candidate with 80% power from 0.0021 nats (swap) and 0.0068 (forcing; Appendix B.5), so the statement is that no operator retains cross-fitted headroom detectable at that resolution, not that the headroom is exactly zero. Because the swap's floor is close to its same-sample value, the cross-fit alone cannot exclude a small persistent component; the stronger evidence is that the matched-noise null below, which has no persistent component by construction, closely reproduces the observed value (0.0072 against 0.0069 for the swap, 0.0291 against 0.0253 for forcing) and the checkpoint ordering of the ratio ($\hat{\kappa}_{115} = 3.97$ under the null against 3.63 observed); cross-fitted, forcing exceeds the swap in only 54 of 115 cells. Precision moves the scale but not this reading: in a matched fp32 re-audit of 13 of 21 WikiText cells $\hat{\kappa}_{\text{site}}$ rises from 5.33 (bf16) to 10.06 (fp32) and neither operator's fp32 cross-fit is detectably positive; a four-cell deterministic check agrees (Appendix A.3).

Positive residuals under forcing. With the intervened token excluded from selection, forcing retains $+0.0007$ $[+0.0003, +0.0013]$ nats and the swap none detectable; Granite-3.0 supplies 79.6% of that residual (without it, $+0.0002$ $[-0.0001, +0.0006]$). Outside the router's committee, forcing an unselected expert retains $+0.0037$ $[+0.0007, +0.0065]$ nats over 19 cells, about 12.1% of its same-sample value, while the swap retains -0.0005 $[-0.0010, +0.0001]$. Forcing's reference is the native expert at full routed mass, so these residuals concern the forcing action class, not the unintervened model (Appendix C.3). Coarsening changes the picture: one static expert per cell, chosen on the registry's documents and scored on fresh ones, gains $+0.0015$ $[+0.0003, +0.0031]$ nats under forcing and $+0.0000$ $[-0.0002, +0.0002]$ under the swap, a policy-level estimand (Appendix B.5).

The per-token winner. Selection that transferred would also make winners reproducible; they are not: forcing's split-half winner agrees with itself 35.2% of the time and the swap's 34.1%, against 33.3% chance among the three non-reference alternatives, and a winner chosen on one half has

regret 0.0094 nats on the other (swap) against 0.0097 for a random candidate. Neither wide margins nor short windows rescue it (Appendix B.3), so decisions should not consume the per-site argmax.

The matched noise null and why the maximum inflates. Write each candidate's same-sample contrast as $\delta_e = \mu_e + \sigma\,\varepsilon_e$, with μ_e its persistent advantage, σ the scale at which the operator perturbs the output, and ε_e zero-mean finite-window variation. If $\mu_e = 0$, then $\mathbb{E}[\max_e \delta_e] = \sigma\,\mathbb{E}[\max_e \varepsilon_e] > 0$ while $\mathbb{E}[\delta_{e^*_A}$ on $B] = 0$: same-sample headroom grows with σ and cross-fitted headroom vanishes, so $H^{\mathrm{hard}}/H^{\mathrm{swap}} \approx \sigma_{\mathrm{hard}}/\sigma_{\mathrm{swap}}$. The null estimates the right-hand side without a distributional model: under exchangeable halves and $\mu_e = 0$ the full-suffix contrast $(\delta_e^A + \delta_e^B)/2$ and the half-difference $(\delta_e^A - \delta_e^B)/2$ share a joint law across candidates, so the maximum of the half-differences, averaged over both signs, is a draw of null same-sample headroom keeping each candidate's own noise law and the cross-candidate correlation. Persistent headroom would show as observed headroom above this null *and* as positive cross-fitted headroom.

The same-sample magnitudes. On the full registry, forcing reports more headroom than the swap in **199 of 202 matched cells**, with median paired difference $+0.0220$ nats and median ratio $4.06\times$; per-checkpoint $\hat{\kappa}$ ranges from $1.12\times$ (Mixtral-8x7B) to $11.90\times$ (OLMoE), and Qwen3-30B-A3B-Instruct, the only instruction-tuned checkpoint, gives 20 of 20 at $\hat{\kappa} = 2.28$. By the null above, $\hat{\kappa} = 5.48$ is a ratio of perturbation scales of the reported estimand, not a ratio of detectable persistent opportunity. Candidate restriction does not drive it: exhaustive $K'{=}E{=}8$ evaluation on Mixtral gives $\hat{\kappa} = 1.12$ $[1.01, 1.27]$, matching its committee value. There, raising K' from 2 to 8 lifts forcing's same-sample headroom from 0.0039 to 0.0091 nats, as does its null (0.0040 to 0.0097), while its cross-fit stays at -0.0007 $[-0.0014, -0.0000]$ (Table 6). The checkpoint is the generalization unit; weighting checkpoints equally gives a median of $2.65\times$ (Appendix B.2); under every preregistered reweighting $\hat{\kappa} \in [2.65, 6.09]$ and no observed value detectably exceeds its null (Appendix B.2). The ratio also falls with the scoring horizon, from $7.85\times$ at the intervened token to $2.28\times$ at $h{=}64$ over 6 checkpoints (Appendix B.4), and exceeds one at every horizon.

Published conventions disagree with the control and with each other. Forcing is our operator, so we also score the published conventions against the swap on matched sites: over 41 cells, ablate-and-renormalize gives $\hat{\kappa} = 1.40$ $[1.19, 1.71]$, fixed-k resampling 2.19 $[1.51, 3.31]$ and forcing 4.37. Scored directly against each other on one population (Table 4), the two conventions in print disagree: ablate-and-renormalize reports $0.64\times$ $[0.47, 0.85]$ the headroom of fixed-k resampling. Logit forcing (Bandarkar et al., 2026) is the identity map on a committee pool, since every candidate is already selected, so it is quarantined from the matched convention grid.

6 Routing-Weight Geometry Sets the Scale of Reported Headroom

Routing-weight geometry changes the dispersion of candidate effects, and the same-sample maximum converts that dispersion into apparent headroom (the σ of Section 5). Forcing differs from the swap in two coordinates of R: it collapses the weight shape onto one expert and, on routers that do not renormalize (both OLMoE releases and Qwen1.5-MoE), raises the total mass from the native top-m mass to about one. We separate the two coordinates, then show that they act on the scale of the candidate-effect distribution.

Total mass: forcing at native mass. The HARD_NM operator keeps full commitment but rescales the expert branch to the native total mass (Table 1), so $H^{\mathrm{hard_nm}} - H^{\mathrm{swap}}$ is shape-associated and $H^{\mathrm{hard}} - H^{\mathrm{hard_nm}}$ is mass-associated. On OLMoE (6 cells, 3,072 matched sites) forcing at native mass gives $\hat{\kappa} = 4.10$ $[3.46, 4.84]$ against 10.58 for plain forcing; the shape increment is $+0.01567$ $[+0.01291, +0.01889]$ nats and the mass increment $+0.03277$ $[+0.02614, +0.03947]$. Across OLMoE, Qwen1.5-MoE and Granite-3.0, weight shape contributes on every router, and the absolute mass increment follows each router's normalization semantics (OLMoE $>$ Qwen1.5-MoE $>$ Granite $= 0$) while its *share* of forcing's surplus does not (Table 5). The increments are descriptive, not an additive causal decomposition.

Weight shape: a dose-response. A logit boost b from the swap sharpens weight shape with identity fixed; rescaling the expert branch sets total mass with shape fixed. Within single batches (4,480

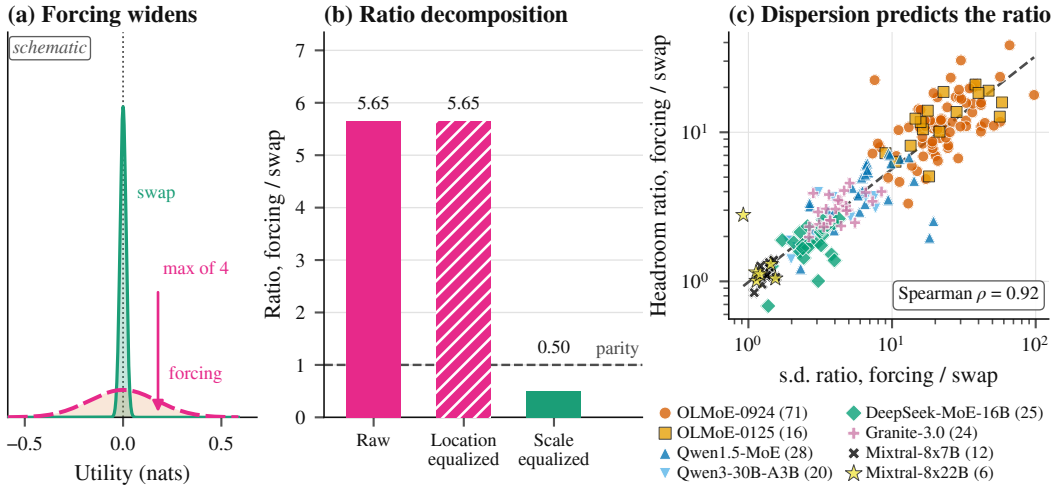

Figure 3: **Routing-weight geometry sets dispersion, and the maximum turns dispersion into headroom.** (a) Schematic: forcing widens the per-candidate utility distribution without moving its centre, and a maximum over candidates tracks the width. (b) Equalizing location leaves the site-pooled ratio $\hat{\kappa}_{\text{site}}$ at $5.65\times$; equalizing scale moves it to $0.50\times$, over-correcting by about a factor of two (Appendix C.2). (c) Across 202 cells from 8 checkpoints, the dispersion ratio tracks the same-sample ratio (Spearman $\rho = 0.92$).

positions) headroom rises monotonically with concentration (Spearman $+1.00$), mass adding less; directional only (Appendix C.1).

The estimator effect: scale, not location. Forcing leaves the centre of the per-candidate utility distribution in place and widens it. A maximum over candidates is far more sensitive to width than to centre (Figure 3). Over 202 cells, equalizing scale collapses the site-pooled ratio $\hat{\kappa}_{\text{site}}$ from $5.65\times$ to $0.50\times$. Equalizing location leaves it unchanged, as it must: each arm subtracts its own native-candidate reference, so an offset common to every candidate cancels exactly. The dispersion ratio tracks the inflation (Spearman 0.92) and the location difference does not (-0.38). Fitted on all but one checkpoint, dispersion predicts the held-out one with a worst-case relative error of 29% (OLMoE; Appendix C.2).

Concentration alone is a coarse correlate. More concentrated routers inflate less across checkpoints ($\rho = -0.79$) but not within them (sign reverses in 6 of 8); a two-expert model with $1/\kappa = w_1 - w_2$ predicts direction, not magnitude (Appendix C.2).

7 Consequences: Pruning and Selection

Pruning as a consequence test. Pruning is used here as a consequence test because its ranking statistic is a mean rather than a best-of-K maximum; therefore any operator dependence cannot be attributed to the selection-max artifact studied above. Each expert is scored by its mean effect $u(e) - u(\text{native})$ over the audited sites where it was a candidate, under the swap or under forcing, and the lowest-scored fraction in each layer is removed by router masking (Table 1), so both arms remove the same number of experts. The certification covers 16 checkpoint-by-budget cells on three architectures (four checkpoints), each with at least 20 seeded random masks, per-item ARC-Easy, ARC-Challenge, HellaSwag and WinoGrande scores, a paired item bootstrap and Holm correction (Tables 7 and 8). *The rankings differ:* per layer, forcing and swap scores correlate at a mean Spearman $\rho = 0.23$ over 13 layers. *Forcing's ranking reproduces; the swap's does not:* scored on disjoint document halves, forcing's per-layer scores correlate 0.32 with themselves (0.64 on OLMoE-0924, 0.08 on Granite-3.0) and the swap's 0.04. *Position relative to random:* on OLMoE-0924 and OLMoE-0125, forcing-ranked accuracy sits at percentiles 2 and 6 of the random masks and swap-ranked accuracy at 51 and 30; on Qwen1.5-MoE the swap-ranked mask is above nearly every random mask (99), and on Granite-3.0 forcing is the higher (62 against 30). *Endpoints:*

the swap-ranked mask has higher accuracy in 12 of 16 cells (4 Holm-significant; forcing in 0), but on WikiText NLL forcing is significantly better in 11 cells and the endpoints favour the same arm in only 2. *Every criterion against random:* of the 6 criteria scored — our two operators, two published ablation conventions and two router statistics — none beats the median random mask on accuracy in more than half its cells (Table 9). On the two OLMoE releases, forcing therefore yields a partly reproducible ranking that differs from the swap's and falls below the random-mask median in accuracy; there the swap's ranking is close to a random one, so this is not evidence that the swap finds better experts. The pattern is checkpoint-dependent, and why accuracy and NLL disagree is not established: across all masks the two endpoints barely co-vary (median within-cell Spearman -0.06, near zero, over 410 masks).

Selection and allocation. The other decisions we tested show the same null. A frozen identity-only selector, evaluated once on a sealed document split, gains $+0.00086$ nats per site, below the cross-fit floor, and did not replicate on preregistered fresh cohorts; per-site selection of the routed weight *shape* cross-fits to -0.00186 $[-0.00245, -0.00130]$; the aggregation oracle of Li (2026) replicates same-sample but cross-fits to -0.00868; and no headroom ranking allocates a global pruning budget across layers better than uniform allocation outside the seeded random spread (Appendices D.1 to D.3).

8 Limitations

1. **Model scope.** 7 frozen architectures (8 checkpoints; OLMoE contributes two releases; 3.3B–141B total). *Affects:* auxiliary-loss-free and dynamic-bias routers, outside the admitted panel (there, swapping before or after the bias defines different estimands). *Does not affect:* the within-panel comparisons.
2. **Candidate pool.** $K'=4$ committee members outside Mixtral; OLMoE supplies 71 of 202 cells (Appendix B.2). *Affects:* dormant experts, seen only in the out-of-committee slice, where forcing keeps a detectable cross-fitted residual. *Does not affect:* the operator ordering, which holds exhaustively on Mixtral and outside the committee (Appendix C.3). *Open:* that slice at scale.
3. **Resolution.** Cross-fitting covers 115 cells and detects an advantage planted in one candidate with 80% power from 0.0021 nats (swap) and 0.0068 (forcing), from the full preregistered power curve (Appendix B.5). *Affects:* smaller or diffuse persistent headroom, which is below this resolution, not shown absent. *Evidence* against more: the null closely reproduces the observed values.
4. **Numerical precision.** *Affects:* the reported same-sample scale (a matched fp32 re-audit of 13 of 21 planned WikiText cells, Section 5 and Appendix A.3; a four-cell deterministic check agrees). *Does not affect:* the fp32 cross-fit and null reading. *Open:* where in the forward pass it arises, the remaining fp32 cells, and Mixtral-8x22B, whose fp32 weights exceed the largest available node.
5. **Downstream scope.** One-shot pruning without retraining on three architectures (four checkpoints); knowledge-intensive suites, a parameter-magnitude criterion and structured pruning methods were not run (a consequence test, not a pruning benchmark). *Open:* whether the operator dependence survives recovery fine-tuning.
6. **Granularity.** Token-level targets only; a static document-level expert is another estimand (Section 5).

9 Conclusion

Under the evaluated operators and resolution, the positive token-level same-sample maximum is largely reproduced by matched selection noise and does not transfer under independent scoring. Routing-weight geometry sets the scale of that maximum: weight shape sets how strongly an operator perturbs the output, and total mass adds to it on non-renormalizing routers. The operator still changes a decision that uses no maximum, since forcing and the swap induce different pruning rankings. We therefore recommend a **minimum reporting standard** for token-level expert counterfactuals: state (i) the operator and (ii) the scoring horizon, (iii) separate candidate selection from scoring, (iv) report a matched-noise null for the same-sample value, (v) include an identity-preserving control whenever expert identity is interpreted, and (vi) report detection resolution as a power curve rather than a single planted dose. **A counterfactual routing map is interpretable only once its intervention semantics, selection–scoring dependence and numerical resolution are stated.**

USE OF LARGE LANGUAGE MODELS

Generative AI tools (OpenAI ChatGPT, Codex; Anthropic Claude, Claude Code) assisted with ideation, experimental and control design, mathematical derivation, coding, debugging, job orchestration, analysis, literature search, drafting, editing and figures. No dataset was AI-generated. Every number is emitted by code from the experiment registry and gate-checked against the manuscript; released numerical assertions verify the two-expert derivation; AI-assisted code passes the instrument validation suite; nondeterministic cells were rerun deterministically. The authors verified every claim and take full responsibility for all text, code and artifacts; AI tools are not authors.

REPRODUCIBILITY STATEMENT

Models are public checkpoints pinned by revision SHA, and all seven regimes are public benchmarks with recorded content hashes. The appendix gives what is needed to re-implement the intervention and re-run the central measurements; the accompanying code submission carries the rest. It releases the instrument, the ten failure checks, the planted positive control, the per-site intervention records and the analysis pipeline, so every number here is regenerable from a registry artifact with a recorded SHA-256 by running one script.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. arXiv:2209.04836.

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. Multilingual routing in mixture-of-experts. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.04694.

Teng-Ruei Chen. How much of the routing gap is real? Decomposing the router-to-oracle gap into reproducible specialist advantage and single-draw label noise, 2026. arXiv preprint arXiv:2607.03436.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models, 2024.

Leonard Engmann, Christian Medeiros Adriano, and Holger Giese. From observation to intervention: A causal audit of expert importance in mixture-of-experts models, 2026. ICML 2026 Workshop on Philosophy of Science Meets Machine Learning (PhilML), non-archival.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022. arXiv:2110.06296.

Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan Rossi, Trung Bui, Hinrich Schütze, and Nanyun Peng. Steering MoE LLMs via expert (de)activation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.09660.

Saloni Garg and Amit Sagtani. Unsolvability ceiling in multi-LLM routing: An empirical study of evaluation artifacts. *arXiv preprint arXiv:2605.07395*, 2026.

Parvel Gu. Detecting a route flip is easier than knowing whether to fix it: Causal route-mediated damage in quantized mixture-of-experts, 2026. arXiv preprint arXiv:2608.11212.

Wentao Hu, Yanbo Zhai, Xiaohui Hu, Mingkuan Zhao, Shanhong Yu, Xue Liu, Kaidong Yu, Shuangyong Song, and Xuelong Li. Awakening dormant experts: Counterfactual routing to mitigate MoE hallucinations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47257–47271, 2026. doi: 10.18653/v1/2026.acl-long.2187. arXiv:2604.14246.

Haochen Huang, Shuzhang Zhong, Shengxuan Qiu, Zhe Zhang, Shuangchen Li, Cong Li, Dimin Niu, Hongzhong Zheng, Guangyu Sun, Runsheng Wang, and Meng Li. HDA-MoE: Hybrid parallelism and dynamic, adaptive scheduling for mixture-of-experts with 3D near-memory processing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2026. Accepted; arXiv:2609.08682.

IBM Granite Team. Granite 3.0 language models. `https://huggingface.co/ibm-granite/granite-3.0-3b-a800m-base`, 2024. Mixture-of-Experts base checkpoint, 40 routed experts, top-8.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Lie Li, Wen Li, Junxiao Shen, and Guosheng Hu. MAPLE: MoE adaptive plug-and-play layer-wise expert allocation. *arXiv preprint arXiv:2608.15299*, 2026.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2409.09951.

Zongfei Li. Beyond routing: Decoupling expert dispatch and aggregation in sparse mixture-of-experts, 2026. arXiv preprint arXiv:2608.08853.

Zongfang Liu, Shengkun Tang, Boyang Sun, Zhiqiang Shen, and Xin Yuan. EvoESAP: Non-uniform expert pruning for sparse MoE. *arXiv preprint arXiv:2603.06003*, 2026a.

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

Yuetian Lu, Ali Modarressi, Yihong Liu, and Hinrich Schütze. Expert-aware causal tracing of factual recall in sparse MoE language models. *arXiv preprint arXiv:2606.03780*, 2026.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4768–4777, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2202.05262.

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

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

Qwen Team. Qwen3-30B-A3B-Instruct-2507. Model card, `https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507`, 2025.

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 preprint arXiv:2606.25092*, 2026.

Pratinav Seth, Hem Gosalia, Aditya Kasliwal, and Vinay Kumar Sankarapu. CircuitKIT: Circuit discovery, evaluation, and application toolkit for mechanistic interpretability. *arXiv preprint arXiv:2607.19317*, 2026.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, number 28 in Annals of Mathematics Studies, pp. 307–317. Princeton University Press, 1953.

Raghul Sugumar and Amrit Gopinath. COVER: Identifiable evaluation of coalition routing. *arXiv preprint arXiv:2608.28475*, 2026.

Viet-Hoang Tran, Van Hoan Trinh, Khanh Vinh Bui, and Tan M. Nguyen. On linear mode connectivity of mixture-of-experts architectures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2509.11348.

Jiaming Wei, Zekun Wu, Adriano Koshiyama, and Maria Perez-Ortiz. Routing is least learnable where it is most valuable: Bounds on representation routing for web agents. *arXiv preprint arXiv:2608.06171*, 2026.

Danny Wood and James Stringer. Permutation-based stegomalware in large language models: Threats and countermeasures, 2026. arXiv preprint arXiv:2609.16193.

Youngsik Yoon, Siwei Wang, Wei Chen, and Jungseul Ok. When are experts misrouted? Counterfactual routing analysis in mixture-of-experts language models, 2026. arXiv preprint arXiv:2605.07260.

Yu Zhang. Reproducible evaluation of MoE expert caching: Replay semantics, workload contamination, and operating regimes. *arXiv preprint arXiv:2608.07911*, 2026a.

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

A Method and Instrument Validation

A.1 The Identity-Preserving Intervention

The swap is a two-line edit to the gate: exchange the two logits *before* selection and the model's own top-k, softmax, renormalization and capacity logic run unmodified. No labels, no training, no architecture surgery. The twelve reporting fields that the conclusion's six summarize are in the supplementary `docs/reporting_standard.md`. The audit intervenes at one site at a time, so the listing is the single-site form the implementation actually runs: `t` is one flattened token position and `e` one candidate expert. Vectorizing over a batch whose swapped indices differ per row needs advanced indexing (`z[rows, e]`, `z[rows, top1]` gathered and written back), not the shared-column form below; we never rely on that path.

```
def swap(gate_logits, t, e):          # ONE site: token t, candidate e
    z = gate_logits.clone()
    top1 = int(z[t].argmax())         # the incumbent at this site
    if top1 != e:                     # exchange exactly two logits
        z[t, e], z[t, top1] = gate_logits[t, top1], gate_logits[t, e]
    return router_topk_softmax(z)     # native dispatch, untouched
```

Two router patterns, two attachment points. *Logit-accessible* routers (OLMoE, Qwen1.5-MoE, Qwen3-MoE, Granite-3.0) expose the gate linear, so the intervention is exactly the listing above. *Fused or tuple* routers (DeepSeek-MoE) select top-k inside the kernel and expose only the dispatched `(expert_indices, weights)` pair; there we write the candidate into the dominant routed slot, displace the incumbent into the candidate's old slot, and leave the `weights` tensor **strictly untouched**. Both attachment points preserve the routed-weight multiset exactly. HARD has the same two code paths; they coincide where the gate renormalizes and differ in total mass where it does not, hence the † in Table 1. The swap exchanges *two* roles, since its image lies in the orbit of the native routed-weight vector under permutation of expert indices and any permutation moving exactly one index is the identity: no operator reassigns one index while holding the multiset fixed. The published removal conventions use the same attachment points. Ablate-and-renormalize drives e's logit to $-\infty$ so that the router's own softmax renormalizes; on routers that select after the softmax this also admits the next-ranked expert, so there it coincides with router masking (Table 1). Zeroing without renormalization needs write access to the dispatched weights and is run only where the router exposes them (the tuple path). A preregistered bias-corrected (auxiliary-loss-free-style) router extension did not satisfy the prespecified admission criterion and is not included.

A.2 Proof of Remark 1

Proof. g selects some $e_t \in \mathcal{K}_t$ and applies $\mathcal{I}_{\text{swap}}$, so pointwise $u_\ell(g(t)) - u_\ell(\text{native}_t) \leq \max_e [u_\ell(\mathcal{I}_{\text{swap}}(S_t, w_t, e)) - u_\ell(\text{native}_t)] = H^{\text{swap}}_{\ell,t}$, where the swap reference is the unintervened route exactly (Definition 1), by definition of the maximum. Taking expectations and dividing by $\mathbb{E}_t[H^{\text{hard}}_{\ell,t}] > 0$ (both headrooms are nonnegative pointwise because the native candidate lies in $\mathcal{K}_t$, so the division preserves the inequality) gives (1). Nothing in the argument refers to how e_t was chosen. □

The bound is stated over expectations, so it uses a ratio of expectations, $\hat{\kappa} = \sum_c \bar{H}^{\text{hard}}_c / \sum_c \bar{H}^{\text{swap}}_c$ over admitted cells c with $\bar{H}_c$ the cell's site-mean headroom (each cell weighted equally), never a median of cellwise ratios, and it binds the identity-only class alone: router retraining, pruning and changes to gate temperature or to m move the weights and fall outside it. Per-checkpoint ceilings are in the supplementary `statistics/weighting_sensitivity.md`.

A.3 Instrument Validation

Ten failure checks run before any number is admitted; Table 2 reports them, recomputed by code from the released rows over the current admitted registry.

Table 2: Instrument validation over the *current* admitted registry: every hard/swap row of every admitted family (1,235,776 rows in 748 files) and all 202 registry cells. The two row-level and cell-level populations differ by construction: row checks run over all collected rows and the token-set partition over every *pairable* cell, of which the registry admits 202 after the prespecified gates. $*$ marks a check that does not pass everywhere; both are discussed in the text.

check	measured	
forcing changes the output	631,930/632,068 hard rows non-zero Δu (99.98%)	✓
identity no-op is exactly zero	153,229/153,231 swap-native rows exactly 0 (99.9987%)	$*$
oracle exceeds native	$\min H^{\text{hard}} = 0.00359 > 0$ over 202 cells	✓
matched token sets	193/209 cells identical token set ($+4$ equal-count/different-position, $+12$ unequal-count, $+0$ native-filter, $+0$ dropped)	$*$
matched site counts after intersection	202/202 registry cells $n^{\text{hard}} = n^{\text{swap}}$	✓
full commitment is total	632,068/632,068 hard rows carry forced weight $= 1.0$	✓
route sets well formed	1,235,776/1,235,776 duplicate-free top-m sets	✓
router probabilities in range	1,235,776/1,235,776 in $[0, 1]$	✓
planted control recoverable	$\rho = +0.641 / +0.731$ (swap/hard), 128 sites (Appendix A.4)	✓
no target label in any deployable rule	enforced statically at the feature contract	✓

The identity no-op is the sharpest: swapping an expert with itself must move the loss by exactly zero, and it does on 153,229 of 153,231 swap-native rows (99.9987%). The 2 exceptions are localized individually — file, line, checkpoint, layer, regime, sequence, token, candidate, signed delta and artifact hash, in the supplementary `data/noop_exceptions.json`. Both are DeepSeek-MoE layer 11 rows from a nondeterministic collection, and the larger is 2.48 bf16 machine epsilons; we report the rate as 99.9987% rather than rounding it to 100%. Granite-3.0's original collection failed this check at 10^{-3}–10^{-2} nats although the swap hook returned a value-identical logit tensor, so the difference arose in the forward pass, not in the operator: two identical forward passes of Granite-3.0 differ by up to 2.03 in a logit and 0.0041 nats of suffix NLL on average, while OLMoE's differ by exactly 0.0; with deterministic GPU kernels Granite's difference is exactly 0.0, so the failure is run-to-run nondeterminism in its scatter-based expert kernel, not a property of the swap. Every Granite cell in the registry comes from a deterministic re-audit. The second $*$ is the token set, and it partitions exactly: over 209 pairable cells, 193 have an identical token set across the two arms; of the remaining 16, 4 cover different token positions at equal counts, 12 have one arm audited at more sites than the other, 0 lose a site to the native-candidate filter, and 0 fall below the minimum matched-site count. The five categories are assigned in a fixed order, so they are mutually exclusive and sum to 209 by construction (supplementary `data/token_set_partition.json`). Every H is computed on the *intersection* of the two arms' sites, so an unequal audited count changes which sites contribute but never whether the arms are compared on the same ones; the matched

counts are equal in every registry cell by construction. Two further instrument faults were caught by the admission gate and the affected collections re-run: a transformers 5.x regression left the Qwen3 forcing hook inert (candidates differed at only 72.4% of 10,560 sites against a 95% bar; 99.9% over 104,525 after the fix), and Mixtral's full-commitment arm bound on only 85% of sites under a single-device map, binding at 100% when re-run across two devices. The admission criterion is evaluated at the family level over all of a family's rows. Mixtral-8x22B was first held out by it; its non-binding sites are confined to one early collection run under the 5.x environment, its other collections bind at every site, and the family passes the criterion with that run included. Invalid cells were excluded by these prespecified gates; the complete list with reasons is the supplementary `data/registry_excluded.csv`.

Numerical precision. One cell on each of four checkpoints (OLMoE layer 7, Qwen1.5-MoE layer 10, Qwen3-30B-A3B-Instruct layer 24, Granite-3.0 layer 16) was re-scored on the same sequences and token positions under bf16 deterministic kernels, two bf16 non-deterministic replicates and deterministic fp32, and cross-fitted under each. Over 952 shared sites the same-sample ratio moves from 4.38 to 9.69, mainly through the *denominator*: the swap's headroom falls from 0.0078 to 0.0033 nats while forcing's moves only from 0.0342 to 0.0322, both shifts exceeding the bf16 replicate spread, so part of the bf16 swap value is numerical rather than routing. The conclusion does not change: cross-fitted headroom at fp32 is $+0.0005$ $[-0.0001, +0.0013]$ (swap) and $+0.0033$ $[-0.0042, +0.0107]$ (forcing), neither detectably positive, and the null still reproduces the same-sample values (0.0030 against 0.0033; 0.0320 against 0.0322). **Numerical precision changes the apparent magnitude of the oracle without revealing reproducible routing opportunity.** On four cells the forcing interval is an order of magnitude wider than the pooled one in Table 3, so this bounds the precision dependence rather than re-running the 115-cell cross-fit at fp32. *Matched fp32 re-audit (partial coverage).* A preregistered re-audit re-scores each cell's bf16 rows in deterministic fp32 on identical sequences, positions and candidate lists (a site enters only if both arms have every candidate and fp32 keeps the pinned native). At submission 13 of 21 planned WikiText cells and 13 of 21 code cells are complete: OLMoE-0924 L3,7,12; OLMoE-0125 L3,7,12; Granite-3.0 L8,16,24; Qwen1.5-MoE L6,10,18; Qwen3-30B-A3B L8. Pooled over those WikiText cells (3489 swap sites), the swap's same-sample headroom is 0.0078 nats in bf16 and 0.0041 in fp32 and forcing's 0.0423 and 0.0419, so $\hat{\kappa}_{\text{site}}$ moves from 5.33 to 10.06. At fp32 the swap cross-fits to $+0.0004$ $[-0.0002, +0.0010]$, not detectably positive, and forcing to $+0.0017$ $[-0.0025, +0.0064]$; the matched-noise null gives 0.0041 and 0.0446 against the observed fp32 values. The preregistered decision rule is defined on the full planned set and is therefore not applied to this incomplete subset. Mixtral-8x22B is excluded because its fp32 weights need 564 GB (141B parameters $\times$ 4 B), above the 384 GB of the largest single node. Where in the forward pass the dependence arises is not localized, and we draw no causal conclusion from it.

A.4 Planted Ground Truth Cannot Adjudicate

Planted ground truth cannot adjudicate between operators, because fixing a target chooses a winner: the operator whose estimand matches the plant. With single-expert targets, full-commitment forcing computes exactly the planted expert's function and reproduces the planted ranking by construction ($\rho = -1.000$: construction 1's target is a signed utility while the operator scores are oriented as loss increases, so exact reproduction appears as $\rho = -1$); with mixture targets whose truth is defined by *removing* an expert and renormalizing, the removal-shaped conventions win ($\rho = +0.58$, $+0.52$, $+0.51$ over 5 seeds) while the swap scores $+0.23$ and forcing -0.77. Neither plant is more canonical, which is the paper's claim, so the swap is justified by what it holds fixed rather than by benchmark accuracy. A planted *positive* control does validate the instrument, recovering $\rho = +0.641$ (swap) and $+0.731$ (forcing) over 128 sites. A row-permuted negative arm keeps a nonzero ρ ($+0.278$ and $+0.324$), but its routing value collapses to $+0.0026$ and -0.0003 nats, against $+0.0214$ and $+0.0200$ for the planted arm.

B Cross-Fit, Reliability and Horizon

B.1 Cross-Fitted Headroom and the Matched-Noise Null

Both arms select and score the best candidate on the same data, so both maxima are optimistically biased; a maximum cannot be de-biased by permuting candidates within a site, it needs independent outcomes. The two interleaved halves of each site's suffix (even against odd offsets, equal length, identical distributions over causal distance) provide them, for every site collected since the split-half rows were added (7 of the 8 checkpoints; Mixtral-8x22B has no split-half rows). The candidate is chosen on one half and scored on the other; "random" is the value of a uniformly drawn candidate. The matched-noise null gives the full-suffix contrast of every candidate the law of its half-difference, which is exact under exchangeable halves and preserves each candidate's own noise law and the cross-candidate correlation. Intervals are document-clustered bootstraps (2,000 resamples).

Table 3: Same-sample, cross-fitted and null headroom per operator (nats per site). "Cells" is the number of cells with split-half rows for that operator; the forcing-versus-swap comparisons use the 115 cells in which both were collected, which is the common matched population of every such comparison in the paper. The four published conventions were collected on a few cells each, so their rows are indicative only. 80%-power dose: the persistent advantage planted in one candidate that the cross-fit detects with 80% power (Appendix B.5; not estimated for the published conventions, –).

operator	cells	H (full)	H (half)	H^{CF} [95% CI]	random	noise null	80%-power dose
forcing (HARD)	119	0.0253	0.0300	-0.0011 [-0.0022, -0.0001]	-0.0029	0.0291	0.0068
identity-preserving swap	116	0.0069	0.0092	-0.0002 [-0.0006, $+0.0002$]	-0.0005	0.0072	0.0021
ablate + renorm.	3	0.0060	0.0079	$+0.0004$ [-0.0003, $+0.0011$]	$+0.0005$	0.0056	–
fixed-k resample	3	0.0064	0.0088	-0.0001 [-0.0007, $+0.0003$]	-0.0002	0.0067	–
logit force	4	0.0102	0.0136	$+0.0003$ [-0.0004, $+0.0009$]	$+0.0002$	0.0099	–
zero, no renorm.	8	0.0067	0.0090	$+0.0001$ [-0.0003, $+0.0005$]	$+0.0002$	0.0065	–

Detection floor. The analysis detects a persistent advantage planted in a single candidate from 0.0021 nats (swap) and 0.0068 (forcing) at 80% power (Appendix B.5); the earlier single planted doses (0.005, 0.010) are superseded by that power curve. So "no detectable cross-fitted headroom" is a statement at that resolution, not a claim of exact zero. Forcing's cross-fitted value is slightly negative because its reference, the native expert at full routed mass, is better than the average forced candidate. The ratio does not survive as a number: cross-fitted it is a ratio of two near-zero, sign-unstable quantities (5.51, interval $[-51.66, 52.91]$), and the paired difference is -0.0008 $[-0.0018, +0.0002]$. A winner chosen on one half has mean regret 0.0094 nats on the other (swap) against 0.0097 for a random candidate, so selection adds almost nothing over a random pick. Every headroom number in this paper other than Table 3 and its companions is selected and scored on the same sample.

Table 4: **Published conventions disagree with each other.** Every operator against every other on one identical population (41 cells, 24,131 matched sites, 7 checkpoints; Granite-3.0 is absent because its rows for these operators come from the nondeterministic collection). *Upper triangle:* headroom ratio $H_{\mathrm{row}}/H_{\mathrm{col}}$ of expected same-sample headroom (the HARD/SWAP entry is $\hat{\kappa}_{41}$). *Lower triangle:* chance-adjusted argmax agreement, $(\text{agreement} - \text{chance})/(1 - \text{chance})$, with chance the site mean of $1/|\mathcal{K}_t \setminus \{\mathrm{ref}\}|$ (0 is chance, negative is below it). *Diagonal* (—): an operator against itself, omitted. Logit forcing and zeroing without renormalization share no commonly collected cell and are absent. ZERO+RN: ablate and renormalize; RESAMP: fixed-k resample.

	HARD	SWAP	ZERO+RN	RESAMP
HARD	—	4.37	3.13	2.00
SWAP	0.07	—	0.72	0.46
ZERO+RN	0.06	-0.04	—	0.64
RESAMP	0.14	0.01	0.12	—

B.2 Per-Checkpoint Summary and Weighting

Cells within a checkpoint share a tokenizer and a routing implementation, so the generalization unit is the checkpoint and bootstrap intervals resample checkpoints; with 8 clusters the percentile interval

$[2.23, 8.74]$ is a descriptive spread, not a calibrated 95% statement. Architecture- and checkpoint-balanced summaries preserve the direction and OLMoE weighting increases the pooled magnitude (checkpoint-equal median $2.65\times$; without OLMoE 2.73 $[1.81, 3.47]$; the per-checkpoint table and all weighting variants are in the supplementary `statistics/weighting_sensitivity.md`). Forcing does not exceed the swap in 3 cells: DeepSeek layer 17 ood_domain_code, seed 0 (0.68); Mixtral layer 8 ood_domain_code (0.84); Mixtral layer 8 ood_long_context (0.96).

Preregistered reweighting. Each claim was recomputed under cell weighting, checkpoint- and architecture-equal weighting (both OLMoE releases one architecture), each of the eight leave-one-checkpoint-out variants, and without the OLMoE family (checkpoint-cluster bootstrap for $\hat{\kappa}$; document-cluster bootstrap for cross-fitted and null-excess values). $\hat{\kappa}$ is 5.48 $[2.29, 8.77]$ by cell, 3.98 $[1.91, 5.99]$ checkpoint-equal, 3.18 $[2.17, 4.66]$ architecture-equal, 2.65 $[1.70, 3.36]$ without OLMoE, and between 3.45 and 6.09 leaving one checkpoint out; forcing exceeds the swap in 199/202 cells (112/115 without OLMoE). The swap's cross-fitted headroom is -0.0002 $[-0.0006, +0.0001]$ by cell and between -0.0003 and -0.0001 across the leave-one-out variants. Forcing's is -0.0011 $[-0.0022, -0.0001]$ by cell but $+0.0006$ $[-0.0000, +0.0013]$ without the OLMoE family, at the boundary of detection. The preregistered rule also required every null-excess interval to include zero; that fails in 14 variants, all in the same direction: the observed same-sample value lies *below* its matched-noise null (by cell, forcing -0.0039 $[-0.0049, -0.0030]$), so the null over-predicts it there. The null is never exceeded, but an over-predicting null is conservative for detecting persistent headroom, which makes those variants weaker support for our reading than an exact match. No variant reverses the sign of $\hat{\kappa} - 1$, of the forcing-over-swap count, or of the swap's cross-fit; the per-variant table is in the supplementary `statistics/`.

B.3 Per-Token Winner Reliability

At a fixed site the forward pass is deterministic, so a rerun returns a bit-identical argmax; what varies is the estimator, since $u(e)$ is a mean NLL over a finite window. We therefore score each site on two interleaved halves of its own suffix, so their disagreement is sampling noise, and report cross-arm agreement on the identical half-window. Over 19,668 sites on 4 checkpoints with deterministic reruns, forcing agrees with itself 35.2% $[34.5, 36.0]$ and the swap 34.1% $[33.3, 34.8]$, against 33.3% chance; cross-arm agreement on the same half-window is 38.0% $[37.3, 38.7]$. The chance baseline is the site mean of $1/|\mathcal{K}_t \setminus \{\mathrm{ref}\}|$: the reference candidate cannot win its own argmax, so $K'{=}4$ leaves three alternatives. Chance-adjusted, that is $+0.029$ and $+0.011$ within arms against $+0.070$ across them. A *heuristic* Spearman–Brown step-up to the full-length window — heuristic because Spearman–Brown is derived for the reliability of a continuous sum score, not for chance-adjusted agreement on a categorical nomination, so it is an indication of direction and not a calibrated estimate — reaches only $+0.056$ and $+0.022$, so doubling the window does not change the reading. **Neither operator reproduces its own nomination more than a few points above chance, and cross-arm agreement is not smaller than within-arm agreement.** Neither confident sites nor short windows rescue it: in the top 1% of margins, chosen on one half and scored on the other, self-agreement is 36.8% (swap) and 41.4% (forcing), and at $h{=}1$ it is 34.3% and 36.1%. The per-site argmax is therefore not a statistic a decision should consume, under either operator; reported magnitudes average H over thousands of sites.

B.4 Horizon Sensitivity

Every intervention is scored at fixed horizons h from the *same* forward pass, and a horizon is reported only where the full window is available, never truncated.

horizon h (tokens)	sites	H^{hard}	H^{swap}	ratio [95% CI]
1	44610	0.38428	0.04894	7.85 [7.16, 8.53]
2	44371	0.21252	0.03322	6.40 [5.86, 6.93]
4	43515	0.11638	0.02222	5.24 [4.82, 5.67]
8	41690	0.06408	0.01508	4.25 [3.93, 4.59]
16	37956	0.03567	0.01041	3.43 [3.18, 3.69]
32	30211	0.02045	0.00747	2.74 [2.54, 2.96]
64	15353	0.01212	0.00531	2.28 [2.08, 2.50]
full suffix[a]	108323	0.03333	0.00604	5.52 [5.14, 5.88]

[a]Full suffix over every paired site in 7 checkpoints (the horizon rows cover the 6 whose rows record per-horizon losses); same candidate set in both arms; not $\hat{\kappa}$, which is computed per cell over the full registry.

Scope: paired sites of the 6 admitted checkpoints whose rows record per-horizon losses (44,610 sites at h=1); the full-suffix row is not $\hat{\kappa}$. The pooled ratio falls from $7.85\times$ at the intervened token to $2.28\times$ at h=64 and exceeds one at every horizon; on Qwen3-30B-A3B-Instruct it falls from $5.07\times$ to $1.66\times$ and on Mixtral from $1.42\times$ to $1.00\times$, reaching parity. Each horizon defines a different finite-window utility and the same-sample maximum inherits that window's estimation noise, so any reported expert-counterfactual magnitude should state its horizon. Suffix-length strata and the position analysis, which trace the same decay, are in the supplementary `statistics/`.

B.5 Power of the cross-fitted test

Table 3 reports the 80%-power doses from the preregistered power curve below; the earlier single planted doses were diagnostic controls and are not the resolution threshold. The analysis traces the whole curve on the same rows, pairing rule, admission gate and document clusters. In each of 400 document-cluster bootstrap replicates per dose, a fresh non-native candidate per site receives a persistent advantage μ on both interleaved halves, and detection is a one-sided cluster-robust lower bound on H^{CF} above zero ($\alpha = 0.05$). This plants on stored utilities, so it tests the estimator, not a model. Pooled over checkpoints (55,192 swap sites, 56,295 forcing sites), the swap reaches 50/80/90% power at $0.0016/0.0021/0.0026$ nats and forcing at $0.0057/0.0068/0.0072$; false-positive rates at $\mu = 0$ are 0.000 and 0.000. Per checkpoint the swap's 80% dose is OLMoE-0924 0.0041, OLMoE-0125 0.0030, Qwen1.5-MoE 0.0034, Granite 0.0033, DeepSeek-MoE 0.0031, Mixtral-8x7B 0.0028, Qwen3-30B-A3B 0.0047 and Mixtral-8x22B 0.0084. Negative results in the main text are stated at these resolutions. Curves are in the supplementary `statistics/`. *Site density.* Re-auditing every eligible position of the same documents (19,828 swap and 20,984 forcing sites over 14 cells on 3 checkpoints; partial coverage of the preregistered set at submission) gives cross-fitted headroom $+0.0001$ $[-0.0007, +0.0007]$ (swap) and $+0.0010$ $[-0.0002, +0.0022]$ (forcing), against same-sample values of $+0.0060$ and $+0.0223$ and nulls of 0.0059 and 0.0237 (the preregistered rule classifies both operators as bounded in the cross-fit). *Choosing on other documents.* A coarser test chooses one expert per cell (or per native top-1 context, or one router-rank slot) on the other documents of a cell and scores it on a held-out document, so no maximum touches the scoring data; this is a policy-level estimand, not the token-level H^{CF}. Relative to a random candidate from the same set, the per-cell choice gains $+0.0029$ $[+0.0021, +0.0035]$ nats under forcing and -0.0000 $[-0.0002, +0.0001]$ under the swap (56,294 forcing sites). The preregistered rule counts an operator as persistent at the document level if any of its three policies beats a random candidate with an interval excluding zero: forcing qualifies, and the swap qualifies only through its router-rank policy, not the per-cell choice reported here; on fresh documents the rule classifies the swap as not persistent. The per-cell forcing choice matches the held-out same-sample winner at 26.0% against 25.0% chance (forcing). On fresh documents that the registry never saw (6,927 forcing sites, partial coverage), choosing on the registry's documents gains $+0.0015$ $[+0.0003, +0.0031]$ under forcing and $+0.0000$ $[-0.0002, +0.0002]$ under the swap. A coarse, forcing-class component of this size is compatible with the token-level null: forcing's reference is the native expert at full routed mass (Appendix C.3).

C Routing-Weight Geometry

C.1 Shape Versus Mass

The routing state $R = (S, \tilde{w}, M, N)$ is defined in Section 3. The swap changes only the assignment of experts to weights; full commitment also changes $\tilde{w}$ and, on non-renormalizing routers, raises M; ablation changes S (and, with renormalization, $\tilde{w}$ and M). In every measured row the router's own rule computes the weights, so no implemented operator changes N; as conventions, zeroing and explicit rescaling depart from N on routers whose native rule differs. The HARD_NM operator keeps full commitment but rescales the expert branch at the intervened token to the native total mass, which is exact on OLMoE (no shared experts, no renormalization), so $H^{\mathrm{hard_nm}} - H^{\mathrm{swap}}$ is shape-associated and $H^{\mathrm{hard}} - H^{\mathrm{hard_nm}}$ is mass-associated. Over 6 cells and 3,072 matched sites scored under all three operators on OLMoE, $H^{\mathrm{swap}} = 0.00506$, $H^{\mathrm{hard_nm}} = 0.02073$ and

$H^{\text{hard}} = 0.05349$ nats. Weight shape contributes across routers and total routed mass adds a second amplification on non-renormalizing routers; the absolute mass increment follows each router's normalization semantics (OLMoE > Qwen1.5-MoE > Granite = 0), which is the ordering stated before those runs. The increments are descriptive, not an additive causal decomposition. Forcing is also not simply a larger dose of the swap: once checkpoint is conditioned on, the swap's dose raises R^2 by less than 0.001.

Table 5: Forcing at native mass: headroom under the swap, forcing at native mass and plain forcing on the same matched sites, with the shape-associated ($H^{\text{hard_nm}} - H^{\text{swap}}$) and mass-associated ($H^{\text{hard}} - H^{\text{hard_nm}}$) increments and sequence-clustered 95% intervals. Granite-3.0 renormalizes, so its native mass is one and the two forcing arms coincide.

Checkpoint	Sites	H_{swap}	$H_{\text{hard_nm}}$	H_{hard}	Shape incr. [95% CI]	Mass incr. [95% CI]
OLMoE-0924	3,072	0.0051	0.0207	0.0535	$+0.0157$ [$+0.0129$, $+0.0189$]	$+0.0328$ [$+0.0261$, $+0.0395$]
Qwen1.5-MoE	4,096	0.0055	0.0105	0.0246	$+0.0050$ [$+0.0041$, $+0.0060$]	$+0.0141$ [$+0.0119$, $+0.0164$]
Granite-3.0	3,072	0.0069	0.0182	0.0182	$+0.0113$ [$+0.0088$, $+0.0142$]	$+0.0000$ [$+0.0000$, $+0.0000$]

C.2 Scale Versus Location

Scope: the 202 cells with candidate-level dispersion records. Across those cells the dispersion ratio predicts the observed inflation (Spearman 0.92) and the location difference does not (-0.38). Equalizing location leaves the forcing-to-swap ratio at $5.65\times$ ($\hat{\kappa}_{\text{site}}$, site-pooled over the 202 cells, which is why it differs from the cell-weighted $\hat{\kappa} = 5.48$), an algebraic invariance check, while equalizing scale collapses it from $5.65\times$ to $0.50\times$; the scale-equalized residual below one means the dispersion ratio over-corrects by about a factor of two. Fitted on all but one checkpoint, dispersion predicts the held-out one with a worst-case relative error of 29% (OLMoE). Because H subtracts the native candidate's own value under the same operator, an additive concentration offset common to all candidates cancels exactly, so concentration acts as a gain on the identity effects rather than as a shift. Across checkpoints more concentrated routers inflate less ($\rho = -0.79$ over 8 checkpoints), but within checkpoints the cell-level sign reverses in 6 of 8, so routed top-m mass is not a substitute for measuring the swap. A two-expert model solves exactly, and solving it for the estimand we actually measure gives $1/\kappa = w_1 - w_2$, the routed-weight gap between the incumbent and the alternative, with w renormalized over the selected pair. This is exact at m=2 rather than a leading-order limit: over 1946 non-degenerate draws it holds to a relative error of 2.5e-13, in 19711 assertions whose worst absolute discrepancy is 4.9e-11 (`src/two_expert_check.py`). Referencing the forcing arm to the native mixture instead of to the native expert at full routed mass gives a different estimand, with $1/\kappa = (w_1 - w_2)/w_1$; it does not describe the quantity measured here. Both forms predict scale and not location, and neither predicts the panel's magnitude (supplementary `theory/full_two_expert_derivation.md`). The full derivation, its $m \geq 3$ example and the reversal scan are in the supplementary `theory/full_two_expert_derivation.md` and `theory/two_expert_check.md`.

C.3 Candidate Pool and Permutation Controls

Every registry cell of a top-4/6/8 checkpoint uses K'=4 $\leq m$, so its candidates are the model's own committee. Four controls test that this does not drive the result. *Exhaustive pool:* on Mixtral-8x7B ($E = 8$) the audit repeated with $K' = E$ gives $\hat{\kappa} = 1.12$ [1.01, 1.27] over 6 cells, matching its committee value. *Candidate-set size:* on the same exhaustive Mixtral-8x7B rows (2,298 sites), nested candidate sets ordered by router rank show same-sample headroom growing with K' in step with the matched-noise null while the cross-fitted value stays at zero within its interval, for both operators (Table 6; for both operators the preregistered rule classifies the increase as a larger same-sample maximum, not a transferable winner). A larger candidate set buys a larger maximum, not a transferable winner. Forcing does not bind at every Mixtral site; the sensitivity restricted to bound sites is in the supplement. *Outside the committee, same sample:* on OLMoE-0924 layer 7 (480 sites, four arms on the same tokens) $\hat{\kappa}$ is 17.8 inside and 12.2 outside, so the ordering does not depend on restricting candidates to the committee. *Outside the committee, cross-fitted:* a split-half-carrying collection (19 cells, 9,285 sites, Granite-3.0, OLMoE-0924, Qwen1.5-MoE) sits at or below its null same-sample ($\hat{\kappa} = 3.95$ against 4.02); cross-fitted, the swap retains -0.0005 [$-0.0010, +0.0001$] but forcing an outside expert retains $+0.0037$ [$+0.0007, +0.0065$], about 12.1% of its same-sample

Table 6: Candidate-set size on Mixtral-8x7B (all $E{=}8$ experts scored; nested sets by router rank, identical sites). H: same-sample headroom; null: matched-noise null; CF: cross-fitted headroom [95% document-clustered CI]; nats per site.

	forcing			swap		
K'	H	null	CF	H	null	CF
2	0.0039	0.0040	-0.0001 $[-0.0004, +0.0002]$	0.0031	0.0032	-0.0002 $[-0.0012, +0.0009]$
3	0.0056	0.0059	-0.0004 $[-0.0008, -0.0000]$	0.0048	0.0051	-0.0004 $[-0.0017, +0.0010]$
4	0.0068	0.0071	-0.0005 $[-0.0011, +0.0000]$	0.0060	0.0063	-0.0004 $[-0.0019, +0.0011]$
8	0.0091	0.0097	-0.0007 $[-0.0014, -0.0000]$	0.0081	0.0085	-0.0004 $[-0.0020, +0.0015]$

value – the largest positive cross-fitted value in the study (the other, with the intervened token excluded from selection, is also under forcing; Section 5). Its reference is the native expert at full routed mass, so an outside expert at full mass can beat that reference out of sample; this is a statement about the forcing action class, not about the unintervened model. *Widening the identity-preserving class:* admitting every injective assignment of the native routed weights, outside experts included, recovers no detectable cross-fitted headroom on Mixtral (-0.0014 $[-0.0038, +0.0012]$) or Qwen1.5-MoE ($+0.0016$ $[-0.0022, +0.0058]$); on OLMoE-0924 (2 documents) a positive offset appears ($+0.0023$ $[-0.0001, +0.0046]$) that a random reassignment earns equally, the winner's gain over a random member lying between -0.0009 and $+0.0008$ nats. *Outside the committee, at scale:* a preregistered collection gives each site its committee, the four experts just below the selection boundary and eight drawn from the rest (17 cells on 5 checkpoints, DeepSeek-MoE, Granite-3.0, OLMoE-0125, OLMoE-0924, Qwen1.5-MoE; partial coverage at submission). Cross-fitted, forcing a near-boundary expert retains $+0.0048$ $[+0.0017, +0.0081]$ nats and a far one $+0.0028$ $[-0.0013, +0.0069]$; forcing at native total mass retains $+0.0011$ $[-0.0002, +0.0025]$ (boundary) and $+0.0010$ $[-0.0008, +0.0033]$ (far), and the swap -0.0009 $[-0.0016, -0.0003]$ (boundary) and -0.0011 $[-0.0018, -0.0004]$ (far). The preregistered rule localizes the residual to the forcing action class through total routed mass: forcing is detectably positive while the swap and native-mass forcing are not. Raw tables are in the supplementary `geometry/`.

D Downstream and External Checks

D.1 Pruning Certification

Our interventions are per-token, so no single intervention has an end-to-end benchmark delta; a ranking does. Each expert's score is the mean of $u(e) - u(\text{native})$ over every audited site at that layer where e was a candidate – a mean, not the maximum H, so the selection artifact of Section 5 does not apply – nothing is normalized across tokens or layers, and every layer prunes the same number of experts by driving its gate logit to $-\infty$, which is the same mechanism as the audit. Benchmarks are scored by per-option length-normalized log-likelihood. The certification run covers 16 checkpoint-by-budget cells on three architectures (four checkpoints), each with at least 20 seeded random masks, per-item scores, WikiText per-window NLL, a paired item bootstrap and Holm correction across cells. On accuracy, swap-ranked pruning is better in 12 of 16 cells, significantly in 4, and forcing-ranked pruning significantly better in 0; on WikiText NLL, forcing-ranked pruning is significantly better in 11 cells and swap-ranked in 2. The certified result is a statement about the mean aggregation: median, trimmed, positive-part, frequency-weighted and rank scores each keep only part of the mask, and a mean of absolute effects replaces it almost entirely. Table 7 gives every cell. *Reproducibility.* Every certified cell's hard and swap masks are rebuilt from the archived audit rows and compared to the mask the certification recorded: 16 of 16 reproduce exactly (Jaccard 1.00 on both arms). Each certification records its ordered input file list and the resolved file set, so mask reconstruction does not depend on scheduler logs (`downstream/pruning_reproducibility.md`).

Every ranking criterion, against the random masks. Restricting the comparison to our two operators would leave open whether an ordinary criterion does better. It does not. Table 9 places all 6 criteria the certification scored — our two operators, the two published ablation conventions, and the two standard router statistics — in the same cell's random-mask distribution. **None of the six beats the median random mask on accuracy in more than half of its cells** (0 of 6), the best

Table 7: **Certified pruning, swap-ranked minus forcing-ranked**, per checkpoint and budget. Δacc is pooled over ARC-Easy, ARC-Challenge, HellaSwag and WinoGrande (percentage points, positive = swap better); ΔNLL is WikiText per-window NLL (nats, negative = swap better). Intervals are paired item/window bootstraps; p is Holm-corrected across the 16 cells. Random: seeded random masks scored in the same cell.

Checkpoint	Pruned	Δacc, swap$-$hard (pp) [95% CI]	Holm p	Sig.	ΔNLL [95% CI]	Random
Granite-3.0	12.5%	-0.9 [-1.8, $+0.1$]	0.892		$+0.000$ [-0.003, $+0.004$]	20
Granite-3.0	25%	-0.9 [-2.1, $+0.2$]	1.000		-0.016 [-0.022, -0.010]	20
Granite-3.0	37.5%	-0.1 [-1.5, $+1.2$]	1.000		-0.002 [-0.010, $+0.006$]	20
Granite-3.0	50%	-0.9 [-2.2, $+0.5$]	1.000		-0.023 [-0.030, -0.016]	20
OLMoE-0125	12.5%	$+0.1$ [-0.9, $+1.2$]	1.000		$+0.016$ [$+0.010$, $+0.022$]	20
OLMoE-0125	25%	$+1.3$ [-0.1, $+2.6$]	0.715		$+0.026$ [$+0.018$, $+0.035$]	20
OLMoE-0125	37.5%	$+1.5$ [$+0.0$, $+2.9$]	0.698		$+0.009$ [$+0.001$, $+0.018$]	20
OLMoE-0125	50%	$+0.1$ [-1.8, $+1.9$]	1.000		$+0.031$ [$+0.021$, $+0.040$]	20
OLMoE-0924	12.5%	$+2.1$ [$+0.9$, $+3.4$]	0.016	✓	$+0.005$ [$+0.000$, $+0.010$]	20
OLMoE-0924	25%	$+3.4$ [$+1.9$, $+4.8$]	$<$0.001	✓	$+0.020$ [$+0.013$, $+0.027$]	20
OLMoE-0924	37.5%	$+3.0$ [$+1.6$, $+4.6$]	$<$0.001	✓	$+0.022$ [$+0.012$, $+0.032$]	20
OLMoE-0924	50%	$+2.8$ [$+1.1$, $+4.4$]	0.011	✓	$+0.027$ [$+0.017$, $+0.037$]	20
Qwen1.5-MoE	12.5%	$+0.4$ [-0.5, $+1.4$]	1.000		$+0.004$ [$+0.002$, $+0.007$]	20
Qwen1.5-MoE	25%	$+0.8$ [-0.3, $+1.9$]	1.000		$+0.013$ [$+0.010$, $+0.016$]	20
Qwen1.5-MoE	37.5%	$+0.3$ [-0.9, $+1.6$]	1.000		$+0.007$ [$+0.002$, $+0.012$]	20
Qwen1.5-MoE	50%	$+0.4$ [-1.1, $+1.8$]	1.000		-0.003 [-0.009, $+0.003$]	20

Table 8: **One-shot pruning by mean expert effect**, per checkpoint over four budgets. Accuracy is pooled over ARC-Easy, ARC-Challenge, HellaSwag and WinoGrande. Swap better: cells where the swap-ranked mask has higher accuracy (Holm-significant in parentheses). Mask reproducibility: per-layer Spearman of expert scores between disjoint document halves. Percentile: pruned-model accuracy among $\geq$20 random masks (50 = median). $\hat{\kappa}$: the checkpoint's registry ratio of summed cell-mean headroom, forcing over swap (as in Appendix B.2).

		accuracy, swap vs. hard		mask reproducibility		hard–swap	percentile among random	
checkpoint	$\hat{\kappa}$	swap better	$\overline{\Delta\text{acc}}$ (pp)	hard	swap	rank ρ	hard	swap
Granite-3.0	3.02	0/4 (0 sig.)	-0.6	$+0.08$	-0.05	$+0.23$	62	30
OLMoE-0125	11.90	4/4 (0 sig.)	$+1.0$	$+0.39$	$+0.06$	$+0.14$	6	30
OLMoE-0924	11.74	4/4 (4 sig.)	$+2.8$	$+0.64$	$+0.04$	$+0.09$	2	51
Qwen1.5-MoE	3.92	4/4 (0 sig.)	$+0.5$	$+0.22$	$+0.10$	$+0.42$	74	99

median accuracy percentile being 48 for the identity swap; on WikiText NLL 2 do, the best being percentile 32 for routing frequency. This is a consequence test under one fixed protocol, not a pruning-SOTA claim: every arm removes the same experts per layer at the same layers, so the only thing varying is which experts the criterion nominates. A parameter-magnitude criterion was not run and is not estimated.

Table 9: **Pruning by every criterion the certification scored**, over 16 cells (4 checkpoints $\times$ 4 budgets), each against $\geq$20 seeded random masks. $\overline{\Delta\text{acc}}$ and $\overline{\Delta\text{NLL}}$ are means against the unpruned model. Percentiles are the median position of the criterion-ranked mask within that cell's random-mask distribution: higher is better for accuracy, *lower* is better for NLL, and 50 is the median random mask. "Beats random" counts cells above the median random mask on accuracy. Two criteria were collected on 13 cells because Granite-3.0's rows for them come from the superseded nondeterministic collection.

ranking criterion	cells	$\overline{\Delta\text{acc}}$ (pp)	$\overline{\Delta\text{NLL}}$	acc. pct.	NLL pct.	beats random
forcing (HARD)	16	-2.96	$+0.1063$	30	58	6/16
identity swap	16	-2.12	$+0.1149$	48	65	7/16
ablate + renorm. (Hu et al., 2026)	13	-4.10	$+0.3154$	15	100	3/13
fixed-k resample (Yoon et al., 2026)	13	-4.40	$+0.2030$	20	60	2/13
routing frequency	16	-2.47	$+0.0982$	42	32	5/16
mean gate prob.	16	-2.88	$+0.0916$	12	35	3/16

Do the two endpoints order masks alike? The paper reports that accuracy and WikiText NLL disagree about which arm is better and does not claim to know why. Before repeating that, we test the prior question. Each cell scores *every* mask — all six criteria and every seeded random mask — on both endpoints, giving 410 mask-level pairs over 16 cells. The within-cell Spearman correlation

between accuracy change and NLL change, aligned so that positive means the endpoints agree, has median -0.06 and runs from -0.48 to $+0.56$, negative in 11 of 16 cells. The two endpoints therefore do not order masks alike in general, which is the setting in which two criteria disagreeing about them is unsurprising. We report this as a measurement and still do not claim a mechanism. *Across-layer allocation.* Splitting a fixed global budget across layers by any headroom ranking is not a dependable allocator: uniform allocation is best in 3 of 6 (checkpoint, budget) cells, all on one checkpoint, and the dispersion-corrected allocation never won outside the seeded random spread. Per-cell scores, the full mask sets and the layer-allocation tables are in the supplementary `downstream/`.

D.2 Selector Replication

A rule built only from deployment-legal router features (router scores, depth, position, token surface; no hidden states) was frozen with hashes recorded, and a sealed document split (157 groups, held out before any search) opened once. It realizes $+0.00086$ nats per site ($[+0.00025, +0.00152]$) over 31,772 sites, while a random committee member at the *same* sites earns $+0.00050$. On two preregistered cohorts drawn from corpus seeds never used before, each opened once, the confirmed selector gains $+0.00016$ ($[-0.00020, +0.00053]$) and -0.00003 ($[-0.00057, +0.00044]$), both intervals including zero. Most of the sealed gain is therefore earned by identifying intervention-worthy sites; the tested selector does not demonstrate transferable recoverability of the expert choice itself. Granite sensitivity and the search log are in the supplementary `selector/`.

Group geometry: the same null one action class up. Selecting the routed weight *shape* per site — temperature, top-1, top-2-equal or Dirichlet over the native committee, identity and total mass fixed — reports $+0.01750$ nats same-sample over 24 cells and 12,155 sites, and -0.00186 $[-0.00245, -0.00130]$ cross-fitted: below a random action at the same sites (-0.00065) and below an action-class-matched null (-0.00007). No fixed geometry is positive either, down to a 0.1% intervention rate.

A wider selector class, and coarser targets. 190 selector configurations — tabular gradient boosting, MLPs, set transformers and relation-biased graph transformers over the site's candidate set — were trained on development folds and scored on held-out report folds. None clears the materiality rule against the tabular baseline (none of 190); the best realizes $+0.00007$ $[-0.00008, +0.00022]$ nats per site, and the graph transformer does not beat the set model that ignores the graph (-0.000026 $[-0.000113, +0.000059]$). Coarsening the target does not help either: over 154 routing-derived clusterings, the winning *cluster* is no more reproducible across suffix halves than the exact expert once a size-matched random partition of equal granularity is subtracted (none met the prespecified criterion). The geometry oracle covers 24 cells and 12,155 sites on OLMoE-1B-7B-0924 and Granite-3.0, layers 3/7/12 and 16, two regimes and three seeds; its tables are in the supplementary `geometry/cross_fit/`.

D.3 Closest Prior-Work Checks

Engmann et al. (2026) conclude that routing statistics are weak predictors of an expert's interventional importance ($|d| < 0.23$) under zeroing without renormalization. Repeating that question on our instrument, the checkpoint-median $|d|$ runs from 0.022 to 0.046 across the 5 operators and no bootstrap upper bound exceeds 0.072: **their conclusion holds under every operator**, which is what Section 5 predicts for a ranking claim. We use one routing statistic where they use several, so this reproduces the *form* of their inference, not their numbers. Yoon et al. (2026) report that on fragile tokens lower-loss equal-compute routes consistently exist, the best route being chosen from the same routes it is scored on. Reproducing that form with our fixed-k resample operator (24,033 fragile-stratum sites, 6 checkpoints, $G{=}4$ routes per site), the standard route is best at 17.1% of sites against 20.0% under an exchangeable-route null that treats each route as "standard" in turn; the excess is -0.7 points $[-0.8, -0.6]$, negative at every checkpoint. We use fewer routes than their 32, which shrinks every best-of-G gap, and do not test their router-update gains.

Structured aggregation (Li, 2026). Li fixes the selected experts and their total mass and reports that a per-token oracle over 20 reweightings (native, uniform, two temperatures, eight half-mixtures and eight vertices) lowers OLMoE's full-horizon cross-entropy by 0.0160 ± 0.0039 nats. Rerun

on our instrument (3 seeds, 6,076 sites), the same-sample number is of the same size, $+0.01977$ $[+0.01793, +0.02169]$, and so is the rate at which the router's top-1 is the best vertex: 17.3% of sites (theirs: 17.2%). It does not survive selection. Our matched-noise null alone reaches $+0.03420$, and choosing the action on one suffix half and scoring it on the other gives -0.00868 $[-0.01081, -0.00661]$ (L7 -0.00493, L12 -0.01243), worse than a random action at the same sites (-0.00705). The best fixed action is native (Table 10). This null applies to per-token selection within their family; it does not test their trained aggregation head, which we did not rerun. We ran Li's action family unchanged inside the geometry instrument, so the support, native-mass and normalization checks that guard our own geometry screen also gate these rows. Two scope differences matter. Li audits OLMoE layers 3, 7, 11 and 15; we cover 7 and 12. We also score the causally affected suffix of a 128-token sequence, split into interleaved halves, so the raw headroom is comparable to Li's figure only in kind. A null that lets every action keep its own average effect, replacing only the site-specific part with sign-randomized half differences, still leaves no excess (-0.00747 $[-0.00955, -0.00544]$). The two halves pick the same action at 6.7% of sites, against 5.8% if the two choices were independent. Every non-native action runs on the rewrite path, whose numerical offset fails our soft identity check. Where the identity control exists (2,026 sites, seed 0), subtracting it leaves the cross-fit at -0.00663 $[-0.00902, -0.00440]$.

Table 10: **Li's structured aggregation family under our protocol** (OLMoE-1B-7B-0924, 20 actions, dispatch and total routed mass fixed). Values are nats per site, and positive means better than native. *Same-sample* picks the per-site maximum and scores it on the same suffix tokens, the quantity Li reports as 0.0160 ± 0.0039 on their layers. *Noise (N5)* is the same maximum taken over sign-randomized half differences, which is our primary matched-noise null. *Cross-fit* chooses on one interleaved half and scores on the other, averaging both directions. *Random* is a random non-native action. Intervals are 95% document-clustered bootstraps over 96 documents, with both layers of a document resampled together. Sites are pooled by weight. Per-seed cross-fit range: -0.01161 to -0.00684.

		same-sample, full suffix			split halves		
stratum	sites	oracle	noise (N5)	excess	oracle	cross-fit [95%]	random
L7	3,038	$+0.01964$	$+0.02664$	-0.00699	$+0.02629$	-0.00493 $[-0.00672, -0.00322]$	-0.00323
prose	1,519	$+0.02281$	$+0.03308$	-0.01027	$+0.03094$	-0.00735 $[-0.01084, -0.00440]$	-0.00548
code	1,519	$+0.01648$	$+0.02020$	-0.00372	$+0.02165$	-0.00251 $[-0.00390, -0.00113]$	-0.00098
L12	3,038	$+0.01989$	$+0.04176$	-0.02186	$+0.02364$	-0.01243 $[-0.01579, -0.00927]$	-0.01086
prose	1,519	$+0.02596$	$+0.05864$	-0.03268	$+0.03062$	-0.02091 $[-0.02671, -0.01551]$	-0.01789
code	1,519	$+0.01383$	$+0.02488$	-0.01105	$+0.01665$	-0.00396 $[-0.00527, -0.00269]$	-0.00383
pooled	6,076	$+0.01977$	$+0.03420$	-0.01443	$+0.02497$	-0.00868 $[-0.01081, -0.00661]$	-0.00705