CO-ACTIVATION IS NOT COLLABORATION: AN INTERVENTIONAL STUDY OF EXPERT COALITIONS IN MIXTURE-OF-EXPERTS ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Mixture-of-Experts routers score each expert independently and execute the Top-K, so the routed set maximizes a modular set function, yet everything downstream acts on the selected experts' outputs jointly. Whether set utility is additive cannot be read from co-activation, a statistic of the router's own policy. We measure it interventionally on three frozen MoE language models of 1.3B to 14.3B parameters, swapping coalition members at fixed active compute against a candidate pool that includes uniformly random experts, and decomposing pair utility into unary and interaction terms. Non-additivity is real, shallow and modest in cost. On OLMoE the interaction-energy ratio falls from 0.302 at the first MoE layer to 0.0005 at layer 14, about half the non-additive energy at shallow layers lies beyond any two-slot account, and interaction accounts for about a tenth of the router's regret against the best measured equal-compute coalition; the shallow concentration replicates on all three models. Within-layer pairwise co-activation does not track signed interventional synergy, because a pair's synergy barely reproduces. Co-activation is reliable at the sampling that routing logs supply, but a pair's mean synergy on one set of documents does not predict its mean on a disjoint set ($|\rho| \leq 0.054$ over eleven layers of two models), and no co-activation–synergy correlation survives FDR in forty-four tests. The oracle gain is heavy-tailed, and a linear per-token head recovers no detectable share of the reweighting gain at our scale, although the same pipeline recovers a third of a planted signal of the same size.

1 INTRODUCTION

A sparse MoE layer routes each token by scoring all N experts and executing the Top-K:

$$S_{\text{router}} = \arg\max_{|S|=K} \sum_{e \in S} s_e(h),$$

the exact maximizer of a modular (additive) set function of the scores. The layer sums the selected experts' outputs, but everything downstream acts on the sum as a whole. If the counterfactual utility of executing a coalition S is approximately additive in its members, modular routing is the right design; if it is materially non-additive, Top-K maximises an objective other than the one determining the model's output. This is not an impossibility result. A scorer with enough capacity could rank the best coalition's members on top, but such scores would no longer be unary quality, and a signal defined on independent per-expert quality has no reason to produce them. We measure whether existing routers' scores carry that information.

Recent analyses document rich observational structure in expert selection: hierarchical cross-layer co-activation groups, a domain-invariant "standing committee" of heavily routed experts, and co-activation graphs injected into routing itself. But co-activation is a statistic of the router's own policy. It is confounded by score correlation, load-balancing pressure, expert popularity, and hidden-state and weight-space geometry (Wang et al., 2026a; Ahrac et al., 2026) and cannot distinguish functional complementarity from any of these. Separately, counterfactual routing analysis shows that equal-compute alternative routes frequently beat the router's chosen route (Yoon et al., 2026) at the level of individual substitutions, leaving open whether the missed utility has a set-level component.

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.

This paper measures the set-level question directly on frozen models against explicit decision criteria; Appendix I gives every threshold with its outcome and provenance, including the four not met, and states plainly that we hold no external registration. We make three contributions.

1. A measurement framework for expert interaction at equal active compute (§3–4). It combines swap-based coalition counterfactuals preserving the routing budget, a candidate pool containing uniformly random experts (which breaks the censoring induced by measuring only router-preferred experts), a unary/interaction decomposition checked against closed-form ground truth, and a $100\times$ noise-floor criterion that decides which models are measurable at all.
2. A map of non-additivity by depth and its cost to the router (§5.1). The interaction-energy ratio falls monotonically from 0.302 at layer 0 to 0.0005 at layer 14 and survives the entropy, gate-weight and sensitivity controls, and the shallow concentration replicates on granite and a 14.3B shared-expert model. At shallow layers about half the non-additive energy lies beyond any two-slot account. The interaction gap is 10.5% of the router's coalition regret pooled and 22.6% at layer 0.
3. The decoupling result and the fact beneath it (§5.2). Within-layer co-activation does not predict signed interventional synergy because a pair's mean synergy barely reproduces on documents it was not measured on ($|\rho| \leq 0.054$ over eleven layers of two models), which bounds the stable pair-level component at 14.4% of the within-pair standard deviation; we leave the group-level question unresolved. §5.3 adds that the router's oracle gain is heavy-tailed and, at our scale, not recovered by a linear head, and §7 distils eight protocol items from these measurements.

Everything here is measurement; we propose no new router. The audits in §2 lack interaction at equal compute: a unary/interaction split of utility, and pair synergy tested across documents.

2 Related Work

Sparse MoE routing. GShard (Lepikhin et al., 2021), Switch Transformer (Fedus et al., 2022), GLaM (Du et al., 2022), BASE (Lewis et al., 2021), Expert Choice (Zhou et al., 2022), Mixtral (Jiang et al., 2024), DeepSeekMoE (Dai et al., 2024), DeepSeek-V3 (DeepSeek-AI, 2024) and OLMoE (Muennighoff et al., 2025) all rank per-expert scores computed independently of the rest of the set.

Observational analyses of expert structure. "Specialization through Collaboration" (Tang et al., 2026) finds hierarchical cross-layer, sequence-level co-activation groups and prunes with them; "The Illusion of Specialization" (Wang et al., 2026b) reports a domain-invariant standing committee; "A Closer Look into MoE" (Lo et al., 2025) documents norm-driven router preferences. Coalition-attribution pruning, as in Shapley-MoE (Huang et al., 2025), SHAPE (Zhang, 2026) and C-GNN-PRUNE (Li et al., 2026), inherits the premise, and "Co-activation Proposes, Ablation Disposes" (Xu, 2026) finds OLMoE attention-head clusters that are real yet fail ablation closure. §5.2 tests the shared premise within a layer, at pair and group level.

Interventional and counterfactual routing analysis. Yoon et al. (2026) introduced equal-compute counterfactual route comparison on frozen MoEs; we extend it with the unary/interaction decomposition, the random-candidate arm and interaction-specific diagnostics. Salomone et al. (2026) run a pre-registered causal ablation of expert modularity at frontier scale and report that apparent modularity mostly dissolves; the intervention is coarser, but its logic and direction match ours. A causal audit of unary expert importance (Engmann et al., 2026) makes the observational-to-interventional gap explicit on three MoEs including OLMoE; ours extends it to sets. Contemporaneously, "Beyond Routing" (Li, 2026) fixes routed identities and router mass, varies only aggregation weights and, unlike our selection-side attempts, exploits the gap with a 301K-parameter head. App. K gives related work in full.

3 Measurement Framework

Setup. OLMoE-1B-7B has 16 layers, 64 experts per MoE layer, Top-8 token-choice routing and softmax gates without Top-K renormalization; it is frozen throughout. We replicate on granite-3.0-1b-a400m (Top-8, renormalized gates) and Qwen1.5-MoE-A2.7B (24 layers, 60 experts, Top-4, no renormalization, plus a sigmoid-gated shared expert at every token). Renormalization makes a member's weight depend on which partners are in S, so granite's 2,720-site run (layers 2/8/14/20) is reported alongside a `uniform` control under which that partner-dependence vanishes (Appendix C).

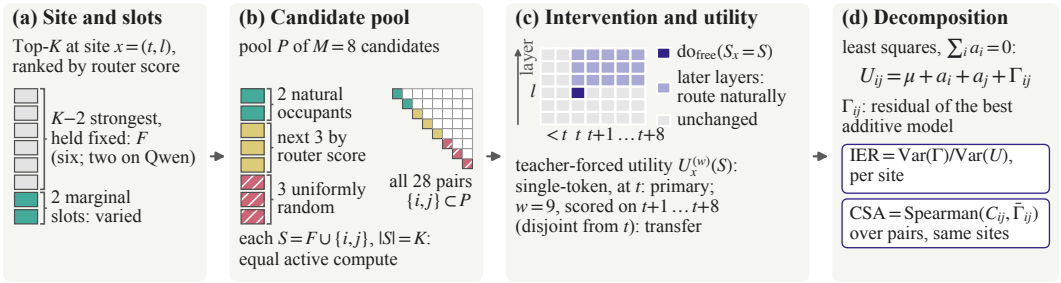

Figure 1: The equal-compute swap protocol (§3): a fixed background F, eight candidates for the two marginal slots, free downstream routing, and the unary/interaction decomposition of pair utility.

Intervention. At a token-layer site $x = (t, l)$, $\mathrm{do}_{\mathrm{free}}(S_x = S)$ executes exactly coalition S there; upstream computation is unchanged and downstream layers re-route naturally. With autoregressive attention only positions $\geq t$ change, from layer l onward, so a single-token coalition costs $(L - l)$ single-position layer passes against the cached clean prefix. No-op replay (natural routing through the intervention path) reproduces the clean forward to $\leq 4 \times 10^{-5}$ nats on the primary sweep (mean $\approx 4 \times 10^{-6}$), our numerical noise floor. A no-op exercises no coalition the sweep measures, so we also check the engine against full recomputation under a non-natural coalition (both slots random, 60 sites, five layers). Agreement is at the float floor and does not degrade with distance from the intervened token (mean $|\Delta| \leq 6 \times 10^{-6}$ nats at every offset; App. A).

A precision floor bounds what we can measure, and it is set by the arithmetic rather than by storage. OLMoE's released checkpoint is stored in bf16, so every measurement here runs fp32 arithmetic on bf16-precision weights; we claim no more weight precision than the checkpoint has. The floor does not compound with recomputation depth (4.01×10^{-6} nats at layer 0, 16 layers recomputed, against 3.80×10^{-6} at layer 14). It does limit which models are measurable: running the same model on the same sites at bf16 arithmetic destroys the measurement.

	replay residual, mean	replay residual, median	effect median	median effect-to-floor
fp32 arithmetic	2.22×10^{-6}	9.0×10^{-7}	3.85×10^{-3}	$4281\times$
bf16 arithmetic	9.16×10^{-3}	2.1×10^{-3}	1.16×10^{-2}	$6\times$

Table 1. Same model, same sites, same bf16-precision weights; only the arithmetic differs (160 OLMoE sites). We adopt a measurement-grade criterion of median effect / median replay floor ≥ 100; fp32 reaches 4.3×10^3, and bf16 misses it by a factor of 18 (unrounded ratio 5.5). Full-scale fp32 runs reach $\approx 4.9 \times 10^3$ (OLMoE), $\approx 5.2 \times 10^3$ (granite) and $\approx 2.0 \times 10^3$ (Qwen).

Because the weights are bf16 in both rows, the difference is attributable to the arithmetic alone, so what the measurement needs is fp32 computation rather than fp32 storage. We implement this by upcasting each weight at the point of use, which holds peak fp32 residency to one weight tensor rather than the model. On OLMoE it reproduces a materialised-fp32 run to 1.1×10^{-6} nats (the float floor, against a 0.11 nat range) at 16.0 GiB peak instead of 28.7. It also admits Qwen1.5-MoE-A2.7B, 14.3B parameters with a shared expert, on one 32GB card, where a six-check gate probe gives 4.0×10^4 against bf16's $2\times$ (App. A).

Equal-compute swap protocol. All coalitions have exactly $|S| = K$ (Figure 1). At each site we hold the $K-2$ strongest routed experts fixed (six; two on Qwen) and vary the two most marginal slots over a candidate pool of $M = 8$: the two natural occupants, the next three by router score, and three uniformly random experts. The random arm makes synergy measurable outside the router's own preference region and de-censors the comparison in §5.2. All $\binom{8}{2} = 28$ equal-compute coalitions are evaluated per site. A top-2-slot variant, which perturbs the strongest slots instead, is measured at 2,720 sites and reproduces the qualitative results with $3.15\times$ the utility spread.

Gate weights and utility. Executing S needs a weighting scheme, and all interaction quantities are relative to it. The primary scheme is `match-natural` (the model's own weights) and the control

is `uniform` $1/K$. $U_x^{(w)}(S)$ is the teacher-forced mean log-likelihood of the true continuation over a window w. Single-token utility is primary, since it is what the router optimises at that position; window utility ($w = 9$, scored on $t{+}1 \dots t{+}8$, disjoint from the selection token) measures transfer. On that window the best coalitions under the two utilities agree at only 6.2% of sites (4.4% if they were independent), so the choice of primary utility changes the answer.

Decomposition. Least squares with $\sum_i a_i = 0$:

$$U_{ij} = \mu + a_i + a_j + \Gamma_{ij}, \qquad (i,j) \in \tbinom{\text{pool}}{2}.$$

Γ is the residual of the best additive model. It reflects non-additivity rather than sampling noise, since utilities are deterministic up to the float floor. Because $K{-}2$ experts are held fixed, the estimand is conditional, and App. A derives exactly what it contains. Writing F for the fixed set and P for the pool, the fit recovers $\Gamma = (I - P_{\text{add}})\,\Delta_i\Delta_j U(F)$, the pool-centred second mixed difference at that background. Here μ absorbs all structure internal to F, and a_i all structure coupling one candidate to F, at any order. Γ, however, carries the pairwise term together with every higher-order term that contains both candidates and part of the background. A term $\theta z_a z_b z_c$ with $c \in F$ lands in Γ in full (IER 0.741), whereas one coupling a single candidate to F is absorbed to machine precision. Higher-order structure can therefore appear here, and Γ bounds what a two-slot design can express rather than the pairwise term itself. The design cannot see interaction among three or more pool members, which is never instantiated since $|S \cap P| = 2$. We call the estimand conditional pairwise interaction at a routed background, using the indefinite article because §5.2 shows Γ is set by token and background jointly. The diagnostics are IER $= \mathrm{Var}(\Gamma)/\mathrm{Var}(U)$ per site, a raw energy ratio with no degrees-of-freedom correction (App. A), and CSA $= \mathrm{Spearman}(C_{ij}, \bar{\Gamma}_{ij})$ per layer, with co-activation counts and synergy means computed on the same sites.

Scale. The OLMoE primary sweep is 8,160 sites (2,040 per layer at layers $\{2, 6, 10, 14\}$) over four domains and three seeds, drawn from 136 documents, which are the clusters of its bootstraps (other runs use 50–516; App. A). All four layers are measured at the identical token positions, so layer comparisons are paired within token. With every control, extension and replication run on the three models, the release totals 79,450 per-site records. App. A gives the full protocol, the statistical plan and the per-run site counts.

4 Estimator Validation

Closed-form ground truth. For coalition output $y(S) = \sum_{i \in S} w_i f_i$ under utility $-\|y(S) - T\|^2$, the pair interaction is exactly $-2w_i w_j \langle f_i, f_j \rangle$, so a designed Gram structure gives the estimator a known answer. Recovery is exact ($r = 1.0000$), but because that utility is additive-plus-pairwise by construction, this only checks the implementation. The rest is the actual test: correct signs on all designed pairs, AUROC 0.95 separating designed from neutral pairs, and both context-dependent sign flips recovered (from -0.543 to $+0.451$). The three-slot estimator returns $< 10^{-10}$ on a surface with exactly zero third-order interaction and detects an injected one.

Compositional toy MoE. A 2.4M-parameter Top-2 MoE on an 8×8 skill-conjunction task, with 12 conjunctions held out, reaches 100% held-in and 34.3% held-out accuracy. On held-out sites, sweeps find median IER 0.169 and 1.15 nats of oracle headroom. Yet the pair predicted by routing-mass skill maps shares an expert with the oracle pair at 31%, against 24.2% by chance, so even here good coalitions cannot be read off routing statistics (App. B).

5 Results

5.1 Non-additivity is material, concentrates in shallow layers, and about half of it lies beyond pairs

Pooled median IER is 0.060 (95% CI $[0.053, 0.066]$), but the dominant structure is a depth gradient that is monotone across all seven measured layers (Table 2). Layers 0, 1 and 2 form a plateau rather than a spike, so "shallow layers" (roughly the first fifth of the network) describes it and "the first MoE layer" does not. The gradient survives three alternative explanations (App. D). Router uncertainty does not explain it: entropy terciles leave layer 2 flat at $0.250/0.288/0.283$, whereas a pooled-tercile analysis would have reported a dramatic "entropy gradient" that is entirely a depth confound. It is

also not a gate-weight artifact, since the uniform control preserves it, or a sensitivity artifact, since median IER against median utility spread across the seven layers ranks at -0.79. Nor is it specific to OLMoE. On Qwen1.5-MoE-A2.7B (60 experts, Top-4, shared expert; 12,240 sites, nine layers), IER falls from a shallow plateau of $0.261/0.269/0.270$ at layers 0/1/2 (OLMoE: $0.302/0.283/0.274$) to 0.0002 at layer 22, monotonically from layer 2 on. As with granite and unlike OLMoE, the very bottom is not strictly monotone, so the concentration and the decline replicate but the shape at the input does not (App. C).

Absolute magnitudes differ. Median RMS $|\Gamma|$ peaks at layer 4 (3.1×10^{-3} nats against a $\approx 4 \times 10^{-5}$ floor), while the ratio is monotone from layer 0. "Material" is a claim about variance share; absolute quantities stay small against a validation cross-entropy of ≈ 2.2 nats (App. D). What this costs the router follows from an exact split of its regret against the best of the 28 coalitions (the marginal-slot oracle gain of §5.3) into a unary and an interaction gap (App. A). The interaction gap is 10.5% of that regret pooled over the primary sweep and 22.6% at layer 0 (App. D), so a unary oracle that ignores interaction already recovers most of it.

Layer	0	1	2	4	6	10	14
IER, 2-slot protocol	0.302	0.283	0.274	0.220	0.144	0.025	0.0005
Energy beyond unary, 3-slot	0.416	0.369	0.333	0.278	0.142	0.034	0.002
Energy beyond pairwise, 3-slot	0.205	0.180	0.168	0.125	0.059	0.013	0.000
Beyond-pairwise share	0.498	0.487	0.501	0.476	0.462	0.358	—

Table 2. Non-additivity by depth on OLMoE. Row 1 is the two-slot protocol (8,160 sites at layers 2/6/10/14 over three seeds; 680 sites per layer at 0/1/4, one seed). Rows 2–4 are an independent three-slot sweep (five strongest experts fixed, three marginal slots varied over $M = 10$, all $\binom{10}{3} = 120$ equal-compute coalitions per site, 250 sites per layer, $w = 8$) that fits nested additive and additive-plus-pairwise models. All rows are raw energy ratios. Utilities are deterministic and every coalition is enumerated, so the fits are exact orthogonal projections with no shrinkage to de-bias (App. A). Document-clustered CIs on row 4 are within ± 0.025 at layers 0–6 and $[0.318, 0.383]$ at layer 10; App. D gives the controls and App. C the granite profile.

Row 1 is second order in the varied slots by construction (§3), and the three-slot sweep bounds what that hides. At layers 0-2 about half the non-additive energy lies beyond any two-slot account (0.498, 0.487, 0.501), falling to 0.358 at layer 10; at layer 2, total non-additivity is $2.0\times$ what a pairwise model expresses. We do not carry that factor to the two-slot IER, which uses a different pool, background and window. For graphs whose edges score pairs additively, this is a structural ceiling rather than an estimation problem. On granite a single-token run gives 0.446 $[0.427, 0.466]$ at layer 2 against OLMoE's single-token 0.494, so the ceiling is not an OLMoE artifact. At normalized depth all three agree: the energy sits in the first fifth and is at the floor by the last measured layer, across two gate conventions and an $11\times$ parameter range.

5.2 Co-activation and interventional synergy are quantities of different kinds

On OLMoE, for every layer, the Spearman correlation between pair co-activation counts within natural Top-8 selections and mean measured synergy $\bar{\Gamma}_{ij}$, on the same sites, is null, both on the full candidate pool and on the de-censored support of pairs containing at least one uniformly random candidate (Table 3). Inference is Fisher-z with a pair-level standard error, $1.06/\sqrt{n-3}$, for the correlation at our sampling depth. Pairs share documents and experts, so we checked it on disjoint document halves: the empirical value is $0.92\times$ it (median, 44 tests) and half-sample 95% intervals agree in 97% of splits (93% at worst), a document bootstrap's in 87% (App. E). Across the full family of forty-four tests (three models $\times$ layers $\times$ two supports) the largest $|\rho|$ is 0.084 and none survives BH-FDR (minimum $q = 0.45$). Two are nominal: Qwen layers 4 and 10 on the random arm ($\rho = -0.065$ and $+0.056$), of opposite sign and each under 0.5% of rank variance. Equivalence against $|\rho| = 0.1$ holds at 32 of 44 after BH: every OLMoE test, 16 of 18 on Qwen, and on granite, whose 406–496 pairs leave intervals as wide as the margin, only layer 0 (a 516-document rerun chosen post hoc). Layers 0 and 1, which carry the highest interaction energy and so make the null hardest to obtain, give -0.009 and $+0.003$.

Layer	0	1	2	4	6	10	14
CSA ρ, full pool	-0.009	$+0.003$	-0.014	$+0.014$	$+0.022$	$+0.002$	-0.003
CSA ρ, random arm	-0.027	$+0.003$	-0.002	$+0.037$	$+0.020$	-0.012	$+0.024$
TOST p vs $\lvert\rho\rvert = 0.1$	6.5e-5	2.4e-5	1.2e-4	1.7e-4	4.9e-4	1.7e-5	1.8e-5
Reliability of synergy	0.014	< 0	0.020	0.027	0.014	0.007	< 0
Reliability of co-activation	0.80	0.75	0.91	0.83	0.94	0.96	0.95
Reliability of co-activation, matched n	—	—	0.32	—	0.45	0.53	0.52

Table 3. Co-activation-synergy alignment and the reliability of both quantities, OLMoE, 1,842-2,016 pairs per layer. Intervals and TOST are Fisher-z with the pair-level standard error $1.06/\sqrt{n-3}$, a 95% half-width on ρ of ≈ 0.047; App. E calibrates it against independent document halves. Reliabilities are Spearman-Brown-corrected over 200 random splits of the same sites and pairs. Row 5 is not at matched per-pair sampling, since co-activation is readable at every position (2,013 per pair) while each synergy observation costs 28 forward passes (median 25-27). Row 6 subsamples it to each pair's own synergy n over 20 draws (sd ≤ 0.03). Layers 0/1/4 have a third of the positions, hence their lower unmatched figures. App. E gives the granite and Qwen profiles, the within-domain tests and the pool sensitivity.

The deeper reason is that $\bar{\Gamma}_{ij}$ barely reproduces across documents. A correlation between co-activation and synergy presupposes that a pair has a synergy stable across the sites where it is measured, and at our sampling it barely does. Table 3 measures both on the same sites and the same pairs, and the comparison must be stated at the right sampling. Unmatched, co-activation's split-half reliability is 0.91-0.96 at the four fully sampled layers against synergy's at most 0.020; at matched per-pair n it is still 0.32-0.53. The cost asymmetry is itself the finding. Co-activation's reliability rises at no marginal cost because the router supplies its observations for free; synergy's does not. A co-activation graph is therefore limited by the size of the pair component its edges would carry rather than by the precision of the statistic building them. This also largely predetermines the correlation tests. Attenuation gives $|\rho_{\text{obs}}| \leq \sqrt{r_\Gamma r_C}\,|\rho_{\text{true}}|$, so at $r_\Gamma \approx 0.02$ even a true correlation of 0.5 would appear as ≈ 0.07, inside our own equivalence bound. The TOST is therefore closer to an arithmetic consequence of the reliability than to independent evidence, and we rest the claim on the reliability itself.

The direct test uses documents a pair was never measured on. Re-measured on 136 new OLMoE and 380 new granite documents, a pair's mean Γ on the original documents predicts its mean on the new ones at $|\rho| \leq 0.054$ over eleven layers, with every 95% interval covering zero. At the upper ends the between-pair standard deviation is at most 14.4% of the within-pair one, or 9.1% under a Gaussian model; this is a bound rather than a zero (App. E). Removing the site scale does not raise it. Dividing Γ by its site RMS, or using its within-site rank, removes the heavy tail but keeps the null (direct bound 12.2% against 14.4%; CSA $0/44$). Interaction magnitude does track co-activation ($|\Gamma|$ over site RMS, $\rho \leq 0.45$), but mostly through the gate weights: on identical OLMoE sites the correlation falls from ≤ 0.33 under the model's weights to ≤ 0.07 under uniform ones (App. E). Context is not all that moves synergy. Holding the token and pool fixed and re-drawing only the held-fixed experts, Γ under the natural background barely predicts Γ under a transplanted one ($r = 0.031$ $[0.006, 0.056]$ at layer 2, 0.003 $[-0.028, 0.034]$ at layer 6). The estimand is thus set by token and background jointly, and $\approx 92\%$ of the transplant arm's within-pair variance is within-token (App. E). At matched sampling the reliability is negligible under all three background rules tested and under uniform gate weights. At the point estimates, expert-interaction edges would need roughly 36–$140\times$ our per-pair sampling on OLMoE to reach even a reliability of 0.5. Figure 2 shows this geometrically.

The expert-group literature acts on communities rather than single edges, so we also tested expert groups. We built a standard co-activation graph (popularity-normalised lift, top-decile edges, greedy modularity) and asked whether pairs inside a community carry more $|\Gamma|$ than a matched random grouping. On the original sampling the excess was nominally positive at layers 2 and 6 ($p = 0.007$, 0.021). Held fixed and scored on 9,520 added sites, the layer-2 partition holds at $+7.0 \times 10^{-4}$ nats ($p = 0.027$) overall and $+1.1 \times 10^{-3}$ ($p = 0.038$) on the 136 documents it was not built from, while the layer-6 one does not ($p = 0.14$). We leave the group-level question unresolved. The layer-2 replication falls just short of BH even within the two layers first flagged ($q = 0.054$; 0.11 over seven); it is specific to the top-decile threshold (top 20%, 15% and 5% partitions do not replicate); and a partition-free test on the strongest co-activation edges finds nothing on the added sites ($p > 0.70$).

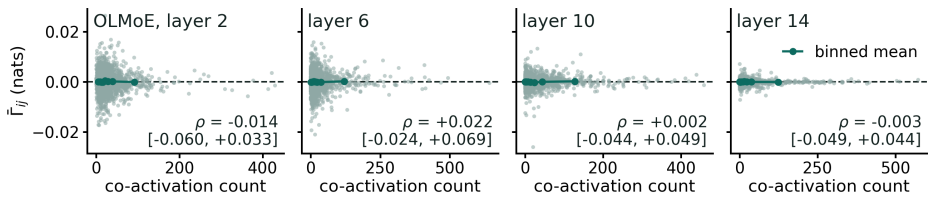

Figure 2: Mean measured synergy $\bar{\Gamma}_{ij}$ (vertical) against pair co-activation count (horizontal), at the four fully sampled layers, with ρ and its 95% interval (Table A6); no structure survives in the plane at any of them. The vertical axis is the weaker of the two: the per-pair synergy mean has a split-half reliability below 0.03 (Table 3), so most of the vertical scatter is within-pair variation across contexts rather than between-pair difference.

The statistic asks whether a partition concentrates pairwise interaction energy, not whether a group has joint function (App. E).

One first-sample departure did not replicate. Granite layer 0 ($\rho = +0.116$, $p = 0.015$ on its first 680 sites) gives $+0.0009$ ($p = 0.99$) on the rerun's 380 unseen documents and -0.013 over the whole rerun; we read the original as sampling variation (App. E).

5.3 The oracle gain is large, heavy-tailed, confined to the token in the mean, and not recovered by a linear head

At the intervened token the router is beatable in hindsight by a wide margin, but the margin sits in a minority of tokens. Reselecting the two marginal slots, reselecting the two strongest, and reweighting the chosen set give mean oracle gains of $+0.0387$, $+0.0898$ and $+0.0871$ nats over 2,720 sites each, against medians of $+0.0109$, $+0.0157$ and $+0.0195$, with 24.3%, 36.4% and 35.3% of sites above 0.05 nats. The mean is 3.6 to $5.7\times$ the median. On a typical token the router is within about 0.011 nats of the best of 28 equal-compute alternatives. The aggregate is carried by a fragile minority, consistent with Yoon et al. (2026), and any summary quoting only the mean overstates how much of the corpus is affected.

In the mean, the gain is confined to the token. Selecting the coalition that maximises token t and scoring it on the eight that follow gives mean changes of -0.00009, -0.00055 and -0.00012 nats, every offset within 2×10^{-3}. This is not an artifact of deep layers having little network left, since at layer 2, with thirteen downstream, the marginal-slot effect is -0.00001. The current-token gain is not paid for later: one intervention's cumulative after-effect over the next eight positions is -0.0007 nats against $+0.039$ at the token itself, though this concerns one intervention against a clean prefix and we do not extend it to a sequence. Per site it is not small: the sum has sd 0.054 and mean magnitude 0.022 nats, over half the token's gain.

Because the oracle requires the label, we tested whether a predictor without it can capture the weighting gain. On the same 2,720 sites we trained a per-token head to choose τ from features a deployed router already has (gate weights over the routed set, their entropy, total router mass, layer) and, in a second variant, from those plus the top 64 principal components of the layer input. Evaluation is five-fold by document, at the intervened token, with gains stated over $\tau = 1$. The relevant baseline is the best single τ chosen on the training documents, because if one global temperature captures the gain there is nothing per-token to learn.

Over natural weighting, the router-visible head gains $+0.00007$ nats (0.1% of the $+0.08707$ nat oracle), the residual-stream head with in-fold PCA -0.00078 (-0.9%), and the best fixed τ itself -0.00071. Paired on the same held-out sites, head minus best fixed τ is $+0.00078$ $[-0.00300, +0.00456]$ and -0.00007 $[-0.00473, +0.00460]$, intervals that include refitting; App. F, Table A9 gives every arm.

Neither head is distinguishable from zero or from the best fixed τ. The best fixed τ is itself 1.0 in four of five folds, so the gap is not a calibration error. The residual-stream head recovers -0.9% of the oracle gain, 95% CI $[-6.0\%, +4.3\%]$ including refits. Two label-free controls make this null interpretable. In a permutation null the target is shuffled across sites within layer, keeping its

marginals and heavy tail; this scores -9.7% (sd 3.2, 200 draws), so the head's -0.9% is above chance ($p = 0.015$), although a head allowed to abstain is not (App. F). In a planted-signal control the target is row-permuted so the best τ becomes a monotone function of a visible residual-stream direction, planted within layer. The signal is recovered at $+34.7\%$ of oracle (95% of 200 plantings: 30–41%), but at $+3\%$ on average with only half the sites planted. The pipeline can turn a strong learnable signal into a third of the gain but recovers no detectable share of the real one.

This negative result is bounded by a linear head, $\approx$2,200 training sites, a one-parameter temperature family and a single-token objective. Contemporaneous work (Li, 2026) reports a gain on this model family from a 301K-parameter head trained on sequence-level cross-entropy over a richer weight space. Their head also reads the already-computed expert outputs, whereas ours sees only the layer-input residual stream. Nothing here contradicts their result, and the contrast localises what would have to be doing the work (input information, capacity, objective or supervision scale) rather than the existence of a gap.

6 Negative results on exploitation

Router order below the Top-K boundary buys no resolvable mean utility. In fitted unary quality, rank-minus-random is $+0.00014$, 95% CI $[-0.0002, +0.0005]$, null at each of ranks 9–11 and on granite (a win-rate edge at OLMoE layer 14 replicates; App. E).

Two further nulls, on different quantities, are given in full in Appendices G and H. Offline, context-conditioned bilinear heads under document-level splits predict per-site Γ_{ij} at held-out $R^2 \leq 0$ for every layer, target and feature set. Trivial baselines do recover a few percent of unary variance at the top slots, so we claim only that the interaction residual transfers at $R^2 \approx 0$. Online, finetuning the router gates (2.1M parameters) with listwise distillation from an equal-compute teacher never reliably beats the LM-only control or a shuffled-label placebo, in three teacher configurations. Four replicates of the control and window arms (identical control runs differ "significantly" under a site-level paired test) give a minimum detectable effect of 0.00057 nats at 80% power, which no informative arm reaches; a leaky positive control moves the endpoint by 4.4 replicate SDs. The teacher substitutes single experts, so this null bounds unary exploitability only.

7 Discussion and a protocol for interventional measurement

For MoE analysis and for routing research. Within-layer co-activation graphs and co-selection statistics describe the router's policy rather than signed pairwise expert interaction. A defender had two openings. For conditioning, we test all four forms available to us: within-domain (null at sixteen tests), community-level (unresolved: one layer, one threshold, §5.2), and token surface class and position quintile (44 more tests, $|\rho| \leq 0.068$, none surviving BH; App. E). For architecture, DeepSeekMoE- and Qwen-style shared experts run at every token by construction, which is the natural place for a yoke to hide. Qwen1.5-MoE-A2.7B has one, and the null holds there: $|\rho| \leq 0.065$ over nine layers, nothing surviving BH, equivalence at 16 of 18, and the two nominal departures opposite in sign. The scope is narrow. The shared expert is in every coalition and is never itself intervened on, as it has no equal-compute counterfactual, so what we show is that a permanently co-activated expert does not yoke co-activation to interaction for the routed experts around it. On the routing side, anyone proposing to exploit counterfactual routing regret should report the oracle's median rather than its mean and beat a single tuned global temperature before claiming a per-token effect. We never trained a coalition-aware router and so do not refute one, but the fixed pair-weight form it would most naturally learn is poorly supported (§5.2), so such proposals must be context-conditioned rather than fixed per pair.

A protocol. These are mostly null and small-effect results, the regime in which measurement discipline decides the answer. Skipping any item below changes a result in this paper (App. L).

1. Measure the numerical noise floor and require effects to clear it by $100\times$. It also identifies the binding resource, here fp32 arithmetic rather than storage, which is what put a 14.3B model on a 32GB card (§3).
2. Stratify within the confound rather than across it. A pooled-tercile analysis reports a dramatic "entropy gradient" that is entirely a depth confound.

3. Do not correct for degrees of freedom when the residual is not sampling noise. With deterministic, exhaustively enumerated utilities, an adjusted ratio is a per-dimension enrichment against an isotropic null and is not a share; read as a share, it overstates the higher-order fraction by $110/75$.
4. Check that a "validation" statistic does not contain the selection statistic. Scoring a coalition on a window that contains the selection token leaves $1/w$ of the same-token gain, which reads as a validated correction. With deterministic utilities the oracle's only optimism is hindsight, and a disjoint window measures transfer instead.
5. Include candidates the system would never choose. Without uniformly random experts in the pool, §5.2 is censored to the router's own preference region and this paper's central result is not measurable.
6. Measure the reliability of a per-unit mean, at matched sampling, before correlating it with anything. Unmatched, the co-activation/synergy contrast comes out up to $3\times$ too large (App. L).
7. Replicate before claiming a difference between trained models. Three pairs of our identically configured runs test as significantly different under a paired test conditioning on the trained model.
8. Give every null a positive control, preferably one that does not see the label. A label-seeing control, App. H's leaky arm included, shows only that the fit works; a planted signal, recovered at $+34.7\%$ (§5.3), licenses the null only for signals that strong (App. J).

8 Limitations

Interaction order and mechanism. Every Γ-based claim is conditional on background, pool and weighting, and Γ includes higher-order terms with fixed experts (§3). Interaction among three or more candidates is outside its support, so a pairwise null is not an interaction null. The three-slot sweep omits Qwen. Without renormalization a layer's output is additive in S, so on OLMoE and Qwen all of Γ arises downstream, from smooth curvature or Top-K flips of later routers, which we do not separate; the depth gradient fits both. §5.2 concerns Γ as deployed, re-routing included.

Support. We study three families: OLMoE-1B-7B (7 layers of 16), granite-3.0-1b-a400m (6 of 24) and Qwen1.5-MoE-A2.7B (9 of 24; 14.3B parameters, shared expert). Intermediate layers are unmeasured. §3's upcast leaves one 32GB card as the ceiling rather than the method, but all three are base models under 3B active parameters, none instruction-tuned, so frontier scale and post-alignment behaviour remain unmeasured. Qwen holds only two experts fixed per site (Top-4 instead of Top-8), so its background is thinner and its Γ less conditioned than the others. The primary protocol varies the two lowest-weight routed slots and scores next-token likelihood; nothing here measures a downstream task or a pruned model.

Statistical scope. granite's 496 candidate pairs, against OLMoE's 2,016 and Qwen's 1,770, put the equivalence bound on the latter two, and the granite run behind the family's layer-0/1 rows was chosen post hoc (the alternatives give 30–38 equivalences and no BH departure; App. E). The fresh-document bound covers eleven layers of OLMoE and granite but not Qwen. The offline null covers rank-16 bilinear heads; the training null resolves 0.00057 nats between four-replicate arms. "Equal compute" means equal FLOPs rather than equal cost. Placement and all-to-all volume depend on which experts run, so some of §5.3's gap is not free in deployment.

Reproducibility Statement

Every number measured on the three models is computed from released per-site records rather than from run summaries, except in Table A10, whose validation perplexities are run-level values and three of whose router-training runs were saved only as run summaries (marked there). The intervention engine, the sweep driver, all analysis scripts, and the 79,450 per-site records they produce are submitted as anonymous supplementary material (excluding two regenerable feature caches, which `extract_features.py` and `extract_agg_features.py` rebuild from the corpus and the released checkpoints); App. J gives the exact model revisions, corpora, sampling seeds and hardware, and App. A gives the protocol and the statistical plan in full. App. I lists every threshold with its observed value, verdict and provenance, including the four that were not met. Two properties of the setup make replication unusually tight: the models are frozen and the engine is deterministic, so effect sizes reproduce to the floating-point floor across platforms, and the no-op replay check (§3) lets any reimplementation verify its own numerical noise floor before comparing numbers with ours.

Ethics Statement

This work measures three publicly released, frozen language models on public corpora. It trains no new generative model, releases no model weights, and involves no human subjects, personal data, or annotator labour. The router-only finetuning arms of §6 modify 2.1M gate parameters of an open checkpoint for a controlled experiment and are not released as a method. The main dual-use consideration is indirect: some compression methods (pruning or grouping guided by within-layer co-activation) rest on the premise that experts which co-activate collaborate. We find that premise unsupported for pairwise interaction in next-token likelihood at the routing slots we perturb, but we ran no pruning or merging experiment and make no claim about how such methods perform; we regard reporting the premise test as risk-reducing. We report the four criteria we failed alongside those we met.

AI Use Statement

In this work, we used generative AI tools as assistants in the following tasks for which disclosure is required: formulating the mathematical claims, proposing and refining hypotheses, designing the experiments, implementing the methods, generating the synthetic datasets, interpreting results, and translation. They also helped clean and reformat the data. We did not use them for proofs, and qualitative data analysis is not applicable to this work. We also used them to help with identifying related work, writing code, creating figures and polishing the text. We reviewed all AI-assisted work: an author checked every number in the paper against the released results, re-ran the analysis scripts on the released per-site records and inspected the intervention engine, and every reference was checked against its primary record. The authors take full responsibility for the contents of this paper, including text, claims and artifacts produced with the aid of generative AI. No LLM is an author.

References

Sagi Ahrac, Noya Hochwald, and Mor Geva. Routers learn the geometry of their experts: Geometric coupling in sparse mixture-of-experts. arXiv:2605.12476, 2026.

Jacob Austin, Augustus Odena, Maxwell Nye, et al. Program synthesis with large language models. arXiv:2108.07732, 2021.

Douglas G. Bonett and Thomas A. Wright. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, 65(1):23–28, 2000.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. arXiv:2107.03374, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, 2024.

DeepSeek-AI. DeepSeek-V3 technical report. arXiv:2412.19437, 2024.

Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. The Shapley Taylor interaction index. In *International Conference on Machine Learning*, 2020.

Nan Du, Yanping Huang, Andrew M. Dai, et al. GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, 2022.

Leonard Engmann, Christian Medeiros Adriano, and Holger Giese. From observation to intervention: A causal audit of expert importance in mixture-of-experts models. arXiv:2606.10703, 2026. ICML 2026 Workshop on Philosophy of Science Meets Machine Learning.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

E. C. Fieller, H. O. Hartley, and E. S. Pearson. Tests for rank correlation coefficients. I. *Biometrika*, 44(3-4):470–481, 1957.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

Granite Team, IBM. Granite 3.0 language models. Technical report, IBM, 2024. URL `https://github.com/ibm-granite/granite-3.0-language-models`.

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

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

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. ISBN 979-8-89176-390-6.

Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Fei Chao, Rongrong Ji, and Liujuan Cao. Discovering important experts for mixture-of-experts models pruning through a theoretical perspective. In *Advances in Neural Information Processing Systems*, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mixtral of experts. arXiv:2401.04088, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, 2021.

Lin Li, Yan Wang, and Zhuopeng Wang. C-GNN-PRUNE: A unified graph-based framework for structure-aware pruning of mixture-of-experts models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22976–22984, 2026.

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

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. A closer look into mixture-of-experts in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4427–4447, 2025. ISBN 979-8-89176-195-7.

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

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, et al. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*, 2025.

Minh-Khoi Nguyen-Nhat, Rachel S. Y. Teo, Laziz Abdullaev, Maurice Mok, Viet-Hoang Tran, and Tan Minh Nguyen. Modeling expert interactions in sparse mixture of experts via graph structures. arXiv:2510.16411, 2025.

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

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

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

Yuanbo Tang, Naifan Zhang, Yan Tang, Meixuan Chen, Shuhan Huang, Tingyu Cao, and Yang Li. Specialization through collaboration: Understanding expert interaction in mixture-of-expert large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2326–2339, 2026. ISBN 979-8-89176-380-7.

Ivan Ternovtsii and Yurii Bilak. Equifinality in mixture of experts: Routing topology does not determine language modeling quality. arXiv:2604.14419, 2026.

Huiyuan Tian, Bonan Xu, and Shijian Li. Beyond geometric complementarity: Coherent overlap in sparse mixture-of-experts routing. arXiv:2607.28308, 2026.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? Interpretable attribution for feature interactions. In *Advances in Neural Information Processing Systems*, 2020.

Xi Wang, Soufiane Hayou, and Eric Nalisnick. The myth of expert specialization in MoEs: Why routing reflects geometry, not necessarily domain expertise. arXiv:2604.09780, 2026a.

Yan Wang, Yitao Xu, Nanhan Shen, Jinyan Su, Jimin Huang, and Zining Zhu. The illusion of specialization: Unveiling the domain-invariant "standing committee" in mixture-of-experts models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14601–14618, 2026b. ISBN 979-8-89176-390-6.

Zihan Wang, Rui Pan, Jiarui Yao, Robert Csordas, Linjie Li, Lu Yin, Jiajun Wu, Tong Zhang, Manling Li, and Shiwei Liu. Chain-of-experts: Unlocking the communication power of mixture-of-experts models. arXiv:2506.18945, 2025.

Yongzhong Xu. Closure-validated circuit discovery in attention heads: Co-activation proposes, ablation disposes. arXiv:2606.09607, 2026.

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

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, 2020.

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

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, 2022.

A MEASUREMENT PROTOCOL AND STATISTICAL PLAN IN FULL

This appendix reproduces the framework of §3 without the compression the page limit imposes, including the statistical plan and the full run inventory.

Setup. OLMoE-1B-7B: 16 layers, 64 experts per MoE layer, Top-8 token-choice routing, softmax gates without Top-K renormalization. The model is frozen throughout.

Intervention. At a token-layer site $x = (t, l)$, $\mathrm{do}_{\mathrm{free}}(S_x = S)$ executes exactly coalition S at that site; upstream computation is unchanged and downstream layers re-route naturally (deployment-faithful semantics). With autoregressive attention, only position t's computation from layer l onward changes, so a coalition evaluation costs $(L - l)$ single-position layer passes against the cached clean prefix; window positions reuse clean states through layer l. No-op replay (natural routing through the intervention path) reproduces the clean forward to $\leq 4 \times 10^{-5}$ nats (max 3.9×10^{-5} on the primary sweep and 5.7×10^{-5} over every other run under the model's own gate weights, except one of the 9,520 OLMoE extension sites at 3.4×10^{-3}, where the 8th and 9th router probabilities tie to float precision and the clean forward took the 9th expert, so the residual is a slot-label swap rather than an engine error; its 28 utilities are valid and the site is kept; mean $\approx 4 \times 10^{-6}$; $\sim 10^{-6}$ on CUDA), which we adopt as the numerical noise floor. No-op replay is a necessary but weak check, though, because it exercises no coalition the sweep measures, and a fault in how the intervened position's key/value are written back would cancel on both sides. We therefore also verify the engine against a full recomputation under a non-natural coalition. A single forward pass over the whole prefix, with the coalition spliced in at (t, l) by hook, is compared position by position with the cached-prefix path at 60 sites across layers 0, 2, 6, 10 and 14, with both varied slots filled by uniformly random experts. Agreement is at the float floor and does not degrade with distance from the intervened token (mean $|\Delta|$ of $4.8, 6.0, 3.4, 6.0, 2.9, 3.4, 4.5, 4.5, 2.1 \times 10^{-6}$ nats at offsets $+0 \ldots + 8$, worst case 3.1×10^{-5}). Such degradation is the signature a mis-propagated cache would produce, and it is absent (`scripts/test_fullforward_vs_cached.py`). Headline effects sit 2-3 orders of magnitude above the adopted floor and 3-4 orders above the mean replay error.

A precision floor bounds what we can measure, and that floor concerns arithmetic rather than storage. One clarification is needed first, because it changes what "fp32" means here. OLMoE's released checkpoint is stored in bf16 (`torch_dtype: bfloat16`; the expert tensors are `BF16` in the safetensors files), so every measurement in this paper runs fp32 arithmetic on bf16-precision weights. We are not claiming access to more weight precision than the checkpoint has.

Two properties of the floor matter. First, it does not compound with the amount of recomputation. The intervention recomputes layers $l \ldots L$, so a residual arising from the clean and intervened paths accumulating in different orders should grow as l decreases. It does not: the mean residual is 4.01×10^{-6} nats at OLMoE layer 0 (16 layers recomputed) against 3.80×10^{-6} at layer 14 (2 layers), and 3.64 versus 3.45×10^{-6} across granite layers 0 and 20. The fp32 floor is a flat property of the arithmetic and does not behave like a divergence between two code paths that accumulates with depth.

Second, and this is what limits the models we could measure, running the same model on the same sites at bf16 arithmetic destroys the measurement. On 160 OLMoE sites,

	replay residual, mean	replay residual, median	effect median	median effect-to-floor
fp32 arithmetic	2.22×10^{-6}	9.0×10^{-7}	3.85×10^{-3}	$4281\times$
bf16 arithmetic	9.16×10^{-3}	2.1×10^{-3}	1.16×10^{-2}	$6\times$

Table A1 (reproduced as Table 1). Same model, same sites, same bf16-precision weights; only the arithmetic differs (160 OLMoE sites). We adopt a measurement-grade criterion of median effect / median replay floor ≥ 100: fp32 reaches 4.3×10^3 and bf16 fails it by a factor of 18 (its unrounded ratio is 5.5, shown as $6\times$ in the table). Full-scale fp32 runs reach $\approx 4.9 \times 10^3$ (OLMoE), $\approx 5.2 \times 10^3$ (granite) and $\approx 2.0 \times 10^3$ (Qwen), on the same definition (median oracle gain over median replay floor); taking each site's range of its 28 coalition utilities as the effect instead gives 1.8×10^4, 1.8×10^4 and 8.4×10^3.

Because the weights are bf16 in both rows, the difference is attributable to the arithmetic alone. This has a practical consequence, since the measurement requires fp32 computation and does not require fp32 storage. A model can be held in bf16 and upcast at the point of use, which bounds peak fp32 residency by the largest weight tensor rather than by the whole network. We implement this in `src/ccgr/fp32_compute.py`. Every weight in these architectures reaches the arithmetic through one of two functional entry points, `F.linear` (for the attention projections, the MLPs, the routers, the per-expert slices of the fused expert tensors and the LM head) and `F.embedding` (for the token table), so patching those two runs every matmul at fp32 while the parameters stay bf16.

One pitfall deserves mention because it is silent. transformers ≥ 5 decorates the fused expert modules so that their `forward` is replaced at load time by a batched-GEMM kernel that indexes the 3-D expert parameter and calls `torch.bmm`; that path never touches `F.linear` and runs the GEMM at the dtype of the weights. The expert arithmetic therefore stays at bf16 while every other matmul in the layer is fp32, and no error is raised; the only symptom is a replay floor three orders of magnitude too high. Routing the experts through their reference implementation fixes it.

Since bf16 widens to fp32 exactly, the upcast path performs the same arithmetic on the same values as the fp32 path and is not an approximation of it. We test this rather than assume it: on 48 OLMoE sites, bf16 storage with fp32 compute reproduces a materialised-fp32 run to a median $|\Delta U|$ of 1.1×10^{-6} nats against a median utility spread of 0.11 nats (a ratio of 9.8×10^4), at a peak of 16.04 GiB instead of 28.69 (`scripts/test_fp32_compute_equiv.py`; `results/fp32_compute_equiv_olmoe.json`; four evenly spaced positions in each of two WikiText-2 and two GSM8K pilot documents, at layers 1, 6 and 10; harder tokens than the sweeps', with a correspondingly wider spread). Here, as in the Qwen ratio of App. C and the cross-stack check of App. J, a site's utility spread is the range (max $-$ min) of its 28 coalition utilities; in App. D and for the top-2-slot variant it is their standard deviation, about a quarter of the range (0.0066 against 0.026 nats at OLMoE layer 2). Table A1's own fp32 and bf16 ratios instead take a site's oracle gain over its natural coalition as the effect. All these ratios are a median effect over a median floor, with one exception: the six-check Qwen probe below uses a median of per-check ratios. The upcast path is what makes Qwen1.5-MoE-A2.7B measurable. Its 14.3B parameters are resident in 26.67 GiB on a single 32GB card, with a replay floor of 1.2×10^{-7} to 4.9×10^{-6} nats and a median effect-to-floor ratio of 3.97×10^4 over six gate checks (two positions of one text at three layers, each check one alternative coalition's $|\Delta U|$ over its own replay residual). This uses a different sample and definition from the full run's 8.4×10^3 (App. C). Run on the same hardware and the same sites at bf16 arithmetic, the identical model gives a floor of 1.6×10^{-3} to 6.0×10^{-2} and a ratio of $2\times$, which is not measurement-grade, and six of eighteen replay checks fail tolerance outright. Salomone et al. (2026) demonstrate interventional testing at frontier scale under a different protocol.

Equal-compute swap protocol. All coalitions have exactly $|S| = K$ (on-distribution; $K = 8$ on OLMoE and granite, 4 on Qwen). At each site we hold the $K - 2$ strongest routed experts fixed (six, or two on Qwen) and vary the two most marginal slots over a candidate pool of $M = 8$: the two natural occupants, the next three experts by router score, and three uniformly random experts. The random arm makes synergy measurable outside the router's own preference region and removes the censoring from the co-activation-vs-synergy comparison. All $\binom{8}{2} = 28$ equal-compute coalitions are evaluated per site. (A top-2-slot variant, perturbing the strongest slots instead, is measured at 2,720 sites and replicates the qualitative results with $3.15\times$ the utility spread.)

Gate weights. Executing S requires a weighting scheme, and all interaction quantities are defined relative to it. The primary scheme is `match-natural` (the model's own weights), and the control is `uniform` $1/K$. Effective-K (weight entropy) is recorded as a covariate.

The three families we measure differ here in a way that matters. OLMoE and Qwen take a raw softmax over all N routed experts without Top-K renormalization (Qwen's configuration sets `norm_topk_prob` to false, as `results/analysis_qwen_fp32_consistency.txt` records), so a member's weight does not depend on which partners are in S, and gate-weight effects are unary-absorbable by construction. Qwen's shared expert is scaled by its own sigmoid gate, a function of the layer input alone; both are recomputed unchanged for every coalition, so the shared expert acts as one more fixed background member and is never a routed one. granite renormalizes over the selected logits, so a member's weight is a function of S. Swapping one partner rescales every incumbent, which injects non-additivity into $U(S)$ even for experts whose functional contributions are exactly additive. The unary-absorbability argument therefore holds for OLMoE and Qwen but

fails for granite, and granite's 2,720-site run (layers 2/8/14/20) is reported alongside the `uniform` control, under which the partner-dependence vanishes by construction (App. C); granite's other runs, its three-slot pilot among them, have no uniform-gate counterpart.

The aggregation arm. The selection-side protocol above asks which experts run. A complementary protocol on the same sites asks how the outputs of a fixed set are weighted. It holds the router's natural Top-K set fixed and varies only the weights over the one-parameter family

$$w_e(\tau) \;\propto\; p_e^{1/\tau}, \qquad \textstyle\sum_{e\in S} w_e(\tau) = \sum_{e\in S} p_e,$$

which contains the model's own weighting exactly at $\tau = 1$, reaches mass-preserving uniform as $\tau \to \infty$ and winner-take-all as $\tau \to 0$, and holds total router mass fixed throughout. As on the selection side, τ is selected on single-token utility and the selected setting is then scored both at the intervened token and on the disjoint window $t{+}1 \ldots t{+}8$. The selection-side and aggregation-side figures are therefore produced by the same procedure and are directly comparable, as oracle gains at t and as transfer effects on the window; neither is a correction to the other. The $\tau = 1$ row runs the model's own weights through the intervention path and must reproduce the clean forward; the observed max residual is 3.7×10^{-5} nats over 2,720 sites.

Utility. $U_x^{(w)}(S)$ is the teacher-forced mean log-likelihood of the true continuation over a window w. Single-token utility is the primary measure, since it is what the router's own objective optimises at that position. Window utility ($w = 9$, scored on $t{+}1 \ldots t{+}8$ so that it is disjoint from the selection token) measures whether an intervention transfers to later positions. The choice between the two matters. On that disjoint window the two utilities' best coalitions agree at only 6.2% of sites (document-clustered 95% CI $[5.3\%, 7.2\%]$; the 2,720 sites of the $w = 9$ run), against 4.4% if the two choices were independent with their observed frequencies (3.6% for a uniform choice among 28), so which one is taken as primary changes the answer (`scripts/window_agreement.py`). At the top slots the agreement, 7.5%, is no higher than its own independence level of 7.0%, because the router's own pair is the single-token best at 23% of those sites. On a window that contains the selection token the same sites agree at 42.8%, so nearly all of that overlap comes from the selection token itself (§7, item 4). The training runs of App. H score such a window (teacher and evaluation windows $w = 8$, positions $t \ldots t{+}7$); on the 896-site $w = 8$ arm of the table below the two utilities' best coalitions disagree at 57.4% of sites, measured on the two-slot protocol rather than on the teacher's single-expert substitutions.

Decomposition and diagnostics. We fit least squares with $\sum_i a_i = 0$:

$$U_{ij} = \mu + a_i + a_j + \Gamma_{ij}, \qquad (i,j) \in \binom{\text{pool}}{2}.$$

Γ is the residual of the best additive model; with deterministic utilities it is non-additivity rather than sampling noise, up to the float floor. Because $K - 2$ experts are held fixed, we need to state precisely which set function this non-additivity belongs to. Write $U(S) = \sum_{T\subseteq S} m(T)$ for the Mobius (Harsanyi) expansion of the true set function, F for the $K - 2$ fixed experts and P for the candidate pool. Every measured coalition is $S = F \cup \{i, j\}$, so

$$U_{ij} \;=\; \underbrace{\textstyle\sum_{T\subseteq F} m(T)}_{\mu_0} \;+\; \underbrace{\textstyle\sum_{T\subseteq F} m(T\cup\{i\})}_{\alpha_i} \;+\alpha_j\; +\; \underbrace{\textstyle\sum_{T\subseteq F} m(T\cup\{i,j\})}_{\gamma_{ij}},$$

which is an identity, with $\alpha_i = \Delta_i U(F)$ the marginal contribution of i to the routed background and $\gamma_{ij} = \Delta_i\Delta_j U(F)$ the second mixed difference there. Since $\mu_0 + \alpha_i + \alpha_j$ lies exactly in the column space of the additive design, the estimator returns $\Gamma = (I - P_{\text{add}})\gamma$, and three exact consequences follow. (i) Every $m(T)$ with $T \subseteq F$ is absorbed into μ, and every $m(T \cup \{i\})$ with $T \subseteq F$ (all orders coupling one varied candidate to the background) is absorbed into a_i. (ii) Γ_{ij} carries the pairwise dividend $m(\{i, j\})$ together with every higher-order dividend $m(T \cup \{i, j\})$, $\emptyset \neq T \subseteq F$, pool-doubly-centred, so higher-order structure remains visible here. A term $\theta\, z_a z_b z_c$ with $c \in F$ enters Γ in full; that single dividend alone returns IER 0.741 in our estimator, while a term coupling only one varied candidate to the background is absorbed to machine precision. Consequently Γ upper-bounds the non-additivity a two-slot design can express, but it does not bound $|m(\{i, j\})|$, which cancellation against the background-coupled terms can make larger. (iii) The design cannot see interaction among three or more pool members. Since $|S \cap P| = 2$ at every site, no $m(T)$ with $|T \cap P| \geq 3$ contributes to any observation; this is a restriction of support rather than absorption. The

estimand is therefore the pool-centred conditional pairwise interaction at a routed background, and every quantity below inherits its dependence on F, on P and on the weighting rule.

The diagnostics are IER $= \mathrm{Var}(\Gamma)/\mathrm{Var}(U)$ per site, reported as a raw energy ratio with no degrees-of-freedom correction (see the note on conventions below); CSA $= \mathrm{Spearman}(C_{ij}, \bar{\Gamma}_{ij})$ per layer, with co-activation counts and synergy means computed on the same sites; and transfer regret, which selects the best coalition on single-token utility and then scores that coalition against the router's pair on the disjoint window $t{+}1 \ldots t{+}8$. Transfer regret is not a debiased version of the same-token oracle. Utilities are deterministic and all 28 coalitions are enumerated, so the selection bias at position t is bounded by the replay floor: at a per-site floor of 3.8×10^{-6} nats, $\sigma\sqrt{2 \ln 28} \leq 1.0 \times 10^{-5}$ nats, or 0.03% of the mean regret. There is nothing of that size to correct, and a disjoint window would not correct it in any case, since it measures a different estimand: whether the intervention's effect reaches later positions.

Regret decomposition. Write S_{r} for the router's marginal-slot pair, $S^* = \arg\max_S U(S)$ for the coalition oracle over the $\binom{M}{2}$ measured pairs, and $S_a = \arg\max_S (a_i + a_j)$ for the unary oracle, the pair preferred by the fitted additive model, which ignores Γ. Then

$$\underbrace{U(S^*) - U(S_{\mathrm{r}})}_{\text{coalition regret}} = \underbrace{U(S_a) - U(S_{\mathrm{r}})}_{\text{unary gap}} + \underbrace{U(S^*) - U(S_a)}_{\text{interaction gap}},$$

an exact identity at every site. The interaction gap is non-negative by construction (S^* maximizes U); the unary gap may be negative at sites where the additive model prefers a pair worse than the router's. Reported shares are ratios of means over sites, interaction gap/coalition regret, so that the two shares sum to one.

Statistical plan. Each claim has a numeric threshold, and every CI comes from a document-clustered bootstrap. Per-site permutation tests are reserved for additive structure, because with deterministic utilities permuting U across pairs cannot test interaction; interaction is judged against the noise floor. Within-layer entropy stratification separates depth from router-uncertainty effects. Winsorization at 1%/99% is applied only to the predictor and target variables of the App. G regressions; no IER, CSA or regret figure anywhere in §5 is winsorized. The three seeds vary which token positions are sampled, not the measurement: the model is frozen and the engine deterministic, so seed-to-seed variation is between-token variation and the seeds are not independent estimates of a fixed quantity. The exceptions to document clustering are the rank correlations computed over pairs (the co-activation-synergy family and the fresh-document test), whose intervals are Fisher-z over pairs with the standard error $1.06/\sqrt{n-3}$, checked against independent document halves (§5.2). Three analyses keep document clustering but not the bootstrap: the conditioning-set sweep and the fresh-document covariance (delete-one-document jackknives, App. E) and the refit-calibrated weighting-head intervals (cluster-robust, §5.3).

Scale and design. The OLMoE primary sweep is 8,160 sites: 2,040 per layer at layers $\{2, 6, 10, 14\}$, over four domains (FineWeb-Edu, MBPP+HumanEval, GSM8K/MATH, OPUS-100 zh/de) and three position-sampling seeds, drawn from 136 documents (34 per domain). Document-clustered intervals cluster on their own run's documents: 136 for this sweep and every run on the same documents (the OLMoE uniform, top-slot, shallow, $w = 9$ and aggregation arms, and granite's natural-gate, uniform-gate and shallow runs); 50 for the three-slot sweeps and pilots on OLMoE and granite; 272 for Qwen and for the OLMoE extension (136 of them new); 516 for granite's extended and deep-layer sweeps (380 of them unseen); 56 for the 896-site $w = 8$ window arm; and 160 for the conditioning-set sweep, whose intervals are delete-one-document jackknives rather than bootstraps. All four layers are measured at the identical token positions, so layer comparisons are paired within token. The 8,160 sites are therefore 2,040 tokens $\times$ 4 correlated measurements, and pooled statistics should be read with that in mind.

Run	Model	Protocol	Weighting	Sites
primary ($\times$ 3 seeds)	OLMoE	marginal slots (ranks 6,7)	match-natural	8,160
weighting control	OLMoE	marginal slots	uniform	2,720
top-slot protocol	OLMoE	strongest slots (ranks 0,1)	match-natural	2,720

Run	Model	Protocol	Weighting	Sites
window arm	OLMoE	marginal slots, w=8	match-natural	896
aggregation arm (§5.3)	OLMoE	routed set fixed, weights varied	τ family, w=9	2,720
leak-free windows (§5.3)	OLMoE	marginal and top slots	match-natural, w=9	5,440
shallow extension	OLMoE	layers 0, 1, 4	match-natural	2,040
third-order pilot	OLMoE	3 slots varied, M=10, w=1	match-natural	500
third-order sweep	OLMoE	3 slots varied, M=10, w=8	match-natural	1,750
replication	granite	marginal slots	match-natural	2,720
gate-convention control	granite	marginal slots	uniform	2,720
shallow extension	granite	layers 0, 1	match-natural	1,360

Total 79,450 per-site records across three model families, counted as the distinct lines of `results/*.jsonl` (the file `phase0_formal_main_merged.jsonl` is a convenience concatenation of seeds 0-2 and is counted only once). They itemize as 33,746 in the table above; 42,180 from the follow-up runs (Qwen 12,240, OLMoE extension 9,520, conditioning-set sweep 9,600, granite extended run (layers 0 and 1, 516 documents) 5,160, granite deep-layer replication (layers 14 and 20, 516 documents) 5,160, granite three-slot 500); and 3,524 retained for provenance (a w=8 aggregation arm whose window contains the selection token, 2,720; the laptop pilots, 756; and a 48-site Qwen adapter probe). An 896-site window arm with the same overlap is likewise retained; the $w = 9$ runs of §5.3 are the ones reported (App. L gives what the overlapping window reports). The third-order pilot evaluates 120 coalitions per site rather than 28. Compute: $\approx$2.5 h wall-clock on one RTX 4090-48G for the primary sweep (fp32, including model load and clean forwards; 1.6 h of recorded sweep time), after a laptop pilot; cross-platform effect sizes reproduce to the float floor.

B Estimator validation in full

Closed-form ground truth. In a system with coalition output $y(S) = \sum_{i \in S} w_i f_i$ and utility $-\|y(S) - T\|^2$, the pair utility's interaction term is exactly $-2w_i w_j \langle f_i, f_j \rangle$. We construct expert vectors with designed Gram structure (redundant pairs with positive inner product, error-cancelling synergistic pairs with negative inner product, near-neutral others, and role assignments that swap between contexts) and run the identical estimator. After projecting the truth into the estimator's residual space, the recovered values correlate with the analytic interaction at 1.0000 (Figure A1). This first number is an implementation check and says nothing about the estimator itself: the synthetic utility is exactly additive-plus-pairwise, so the residual of the additive fit is the projection of the true Γ, and the correlation is an identity up to floating point. The construction does test what follows from it. Signs are correct on all designed pairs, $|\hat{\Gamma}|$ separates designed from neutral pairs with AUROC 0.95, and both context-dependent sign flips are recovered: as the roles swap between contexts, pair $(0, 1)$ moves from -0.543 to $+0.451$ and pair $(2, 3)$ from $+0.451$ to -0.543.

The AUROC deserves a brief comment, because 0.95 understates the estimator rather than bounding it. The four cross pairs, each taking one expert from the redundant role pair and one from the synergistic pair, share the designed direction v by construction and so have non-zero analytic interaction ($|\Gamma| = 0.48$-0.51, larger than the designed synergistic pair's 0.45). Scoring them as negatives places true positives in the negative class. The 0.05 shortfall therefore comes from the labelling and not from the estimator, and Figure A1b separates the three classes so that this is visible.

Compositional toy MoE. We train a 2.4M-parameter transformer whose second block is a Top-2 MoE (16 experts) on an 8×8 skill-conjunction task ($y_t = \pi_b(\pi_a(u_t))$) with 12 conjunctions held out. It reaches 100% held-in but 34.3% held-out answer accuracy, a real composition gap. Coalition sweeps on held-out sites (all 120 pairs) find median IER 0.169 and 1.15 nats of oracle-vs-natural headroom, so choosing the right expert pair substantially repairs unseen compositions. The pair predicted by routing-mass skill maps overlaps the oracle pair at 31% (chance overlap for two pairs drawn from 16 experts is 24.2%; exact pair identity, which this statistic does not measure, would

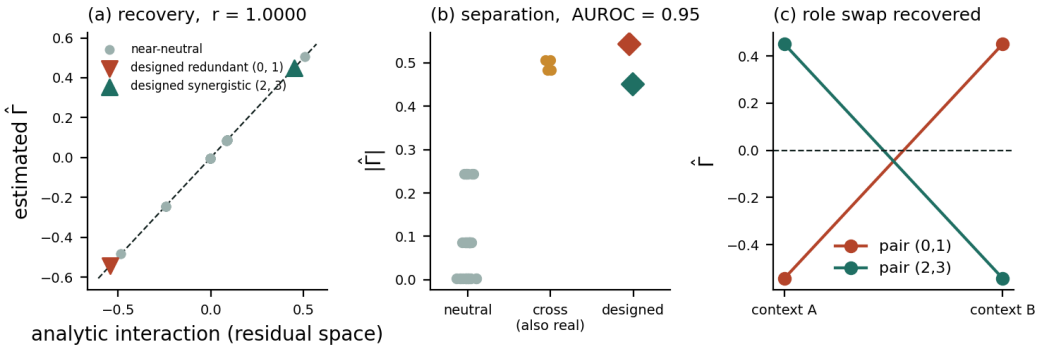

Figure A1: Estimator validation against closed-form ground truth (App. B), pure synthetic system, no model. (a) Recovered $\hat{\Gamma}$ against the analytic interaction projected into the estimator's residual space; $r = 1.0000$. (b) $|\hat{\Gamma}|$ by designed role. The middle class (cross-pairs between the two designed role pairs) has real interaction by construction, which is why the designed-vs-rest AUROC stops at 0.95. (c) The same two expert pairs swap roles between contexts A and B, and the estimator follows the swap in sign and magnitude.

be 0.83%), and its $|\hat{\Gamma}|$ rank is at chance. Even in a controlled compositional system, then, good coalitions cannot be read off routing statistics, which foreshadows §5.2.

C Replication on two further families (granite-3.0-1b-a400m, Qwen1.5-MoE-A2.7B)

Replication. granite-3.0-1b-a400m (1.33B parameters, 24 layers, 32 experts, Top-8, with gates renormalized over the selected logits, a different convention from OLMoE) reproduces every qualitative finding on 2,720 sites at layers $\{2, 8, 14, 20\}$. IER follows a monotone depth gradient (0.216 / 0.134 / 0.035 / 0.002), the interaction share of regret falls with depth (from 18.4% to 2.1%), and CSA is indistinguishable from zero at every layer ($+0.042$, $+0.020$, -0.057, -0.074; $p = 0.41$, 0.67, 0.23, 0.12) and on the random-arm support ($+0.011$, -0.043, -0.050, -0.084; $p = 0.84$, 0.37, 0.29, 0.077). The largest single correlation anywhere, $|\rho| = 0.084$ at the deepest layer, would explain under 1% of synergy rank variance, and multilingual text again shows the largest interaction share (10.3%). Pooled IER median is 0.063, oracle regret 0.036 nats with a 9.6% interaction share, and the router-best rate 7.5%. At the shallow end granite is not monotone, whereas OLMoE is. Extending granite to its two shallowest MoE layers (0 and 1; 1,360 sites) gives the profile $0.184/0.172/0.216/0.134/0.035/0.002$ at layers $0/1/2/8/14/20$, which peaks at layer 2 rather than layer 0. OLMoE declines monotonically from layer 0 (App. D). Monotonicity from the input is therefore not a shared property. The two families share the concentration of non-additivity in roughly the first fifth of the network and the steady decline from layer 2 onward; they differ in the shape at the very bottom. We state the shared finding at that resolution and flag the difference instead of averaging it away; with one seed at granite layers 0-1 we do not attempt to explain it.

Gate arithmetic. Because granite renormalizes over the selected logits (App. A), part of its measured non-additivity could be the gate rescaling that a partner swap mechanically induces rather than expert function, and that artifact would point in the direction of the claim. We therefore repeated the granite sweep under `uniform` weighting at the same 2,720 sites, where the partner-dependence vanishes by construction. The gradient survives, with IER medians of 0.176 / 0.131 / 0.052 / 0.011 at layers 2/8/14/20 against 0.216 / 0.134 / 0.035 / 0.002 under the model's own gates (Table A2). The renormalization confound is real but small: it accounts for roughly a fifth of the layer-2 value and none of the monotone decline. The CSA correlations under uniform weights are -0.047, $+0.051$, -0.035, -0.036 at layers 2/8/14/20, none reaching $|\rho| = 0.084$. With the pair-level standard error of §5.2 their p are 0.36, 0.28, 0.46, 0.46 on the full pool, and none of the eight uniform-gate tests survives BH (minimum $q = 0.53$); the largest is the layer-20 random arm ($\rho = -0.083$, $p = 0.081$), which matches the natural-gate value in layer, sign and magnitude. This rules out gate arithmetic as the explanation of the layer-20 lean, though not sampling, since the uniform sweep reuses the same

2,720 sites and the replication in App. E shows the lean does not survive unseen documents. Pooled IER median is 0.078, and the interaction share of regret is 14.4% / 16.5% / 5.7% / 2.6% by layer.

One asymmetry should be kept in view. granite has 32 experts and therefore 496 candidate pairs against OLMoE's 2,016, so its CSA intervals are twice as wide (half-width 0.093-0.103) and its equivalence test is underpowered (§5.2). granite independently replicates the depth gradient under both gate conventions, along with the direction of the CSA null; the bound on the CSA null rests on OLMoE.

Layer (granite)	2	8	14	20
IER median, model's own (renormalized) gates	0.216	0.134	0.035	0.002
IER median, uniform control	0.176	0.131	0.052	0.011

Table A2. granite IER under both weighting schemes. Under uniform weights a member's weight cannot depend on its partners, so any surviving non-additivity is not gate arithmetic.

Layer	IER, marginal slots	IER, uniform gates	IER, top slots	> 0.1	> 0.3	Entropy terciles (low/mid/high)
0	0.302	—	—	—	—	—
1	0.283	—	—	—	—	—
2	0.274	0.202	0.132	78.5%	45.3%	0.250 / 0.288 / 0.283
4	0.220	—	—	—	—	—
6	0.144	0.099	0.075	59.7%	22.5%	0.125 / 0.141 / 0.169
10	0.025	0.043	0.024	27.8%	8.3%	0.016 / 0.024 / 0.035
14	0.0005	0.012	0.007	7.1%	2.0%	0.000 / 0.000 / 0.001
pooled (2/6/10/14)	0.060	0.063	0.038	43.3%	19.5%	—

Table A3. IER medians per layer under three protocols, with the exceedance fractions and the within-layer router-entropy terciles (primary protocol). The pooled exceedance figures are dominated by layers 2 and 6, which is why we report them per layer. Top-slot figures are seed 0 (2,720 sites); its matched marginal-slot column is 0.281 / 0.149 / 0.027 / 0.0005 (raw, as everywhere else in this paper). Layers 0, 1 and 4 (bold) are a later addition, seed 0, 680 sites each.

Layers 0-1. These were the most likely place for the pattern to break. Layers 0-1 could either continue the climb of layers 2-14 or reveal a first-MoE-layer artifact with a different mechanistic meaning, and neither happens. Across all seven measured layers the profile is

$$0.302 \;/\; 0.283 \;/\; 0.274 \;/\; 0.220 \;/\; 0.144 \;/\; 0.025 \;/\; 0.0005 \quad \text{at layers } 0/1/2/4/6/10/14,$$

monotone throughout, with layer 0 the maximum. Layers 0, 1 and 2 form a plateau (0.302, 0.283, 0.274) rather than a spike, and the decline begins in earnest only from layer 4. The interaction share of regret rises correspondingly, reaching 22.6% at layer 0 against 10.5% pooled. "Shallow layers" is thus the right description and "the first MoE layer" is not, since the effect occupies a region of roughly the first fifth of the network instead of being an input-adjacent special case.

Two caveats apply to these three layers. They are an independent draw of 680 positions per layer from the same 136 documents under the same sampling distribution, and are not the same positions as the four-layer sweep. The candidate-selection rng advances once per (position, layer), so a run with a different layer set does not reproduce the four-layer sweep's positions (27 of the 680 occur in

any of its three seeds). Comparisons of layer medians are therefore between independent samples, which is fine for the profile but means the 0.283 at layer 1 and the 0.274 at layer 2 are not separable. They also rest on one seed, where layers 2/6/10/14 pool three.

A third family. The precision result of App. A makes Qwen1.5-MoE-A2.7B measurable. It has 24 layers, 60 experts, Top-4 routing, 14.3B total parameters (2.7B active), roughly twice OLMoE's size, and a shared expert executed at every token in addition to the routed set. The co-activation literature's architectural bet rests on that last property (§7), and neither OLMoE nor granite has it. We sweep nine layers at 1,360 sites each (12,240 sites, 272 documents, $M = 8$ pool with three uniformly random candidates, all 28 coalitions per site). The replay floor is 1.97×10^{-6} nats median against a median utility range (max − min over the 28 coalitions) of 1.65×10^{-2}, a ratio of 8.4×10^3 (2.0×10^3 with the oracle gain as the effect, as in Table A1); the 3.97×10^4 of App. A is a six-check gate probe.

Two protocol differences matter when comparing the numbers. Top-4 routing means only two experts are held fixed per site, against six on the other two families, so Qwen's conditional background is thinner and its Γ is correspondingly less conditioned. The shared expert is also present in every coalition and acts as an additional always-on background member that no swap can remove.

Layer (Qwen)	0	1	2	4	6	10	14	18	22
IER median	0.261	0.269	0.270	0.210	0.185	0.101	0.047	0.015	0.0002
mean $\lvert\Gamma\rvert$ (10^{-3} nats)	4.16	4.65	4.80	4.74	5.52	4.69	3.82	2.00	0.56
CSA ρ, full pool	$+0.041$	$+0.003$	-0.030	-0.029	$+0.015$	$+0.024$	$+0.010$	-0.009	-0.016
CSA ρ, random arm	$+0.047$	$+0.021$	-0.040	-0.065	$+0.015$	$+0.056$	-0.014	$+0.022$	-0.000
TOST p vs $\lvert\rho\rvert = 0.1$	9.1e-3	5.5e-5	2.6e-3	2.3e-3	3.4e-4	1.3e-3	1.7e-4	1.4e-4	4.2e-4
Reliability of $\bar{\Gamma}_{ij}$	< 0	< 0	0.009	0.038	< 0	< 0	0.032	< 0	< 0
Reliability of co-activation	0.57	0.54	0.52	0.57	0.63	0.69	0.71	0.61	0.60

Table A4. Qwen1.5-MoE-A2.7B, nine layers, 1,360 sites each, 1,767-1,770 pairs per layer. Document-clustered 95% intervals on the IER row are within ± 0.025 throughout. CSA intervals (Fisher-z over pairs, standard error $1.06/\sqrt{n-3}$) have half-width 0.049; every one covers zero except the random arm at layers 4 and 10 ($\rho = -0.065, +0.056$), neither of which survives BH ($q = 0.45, 0.58$). Both reliability rows are Spearman-Brown-corrected over 200 splits of the same run, at 20-21 observations per pair. The co-activation row uses the full sampling the router supplies for free; at matched per-pair sampling it falls to 0.05-0.17, because a pair co-activates rarely when only four of sixty experts are selected.

Three things replicate and one differs. The depth gradient replicates in shape. IER falls from 0.261 to 0.0002, more than a thousandfold, monotonically from layer 2 onward, with a shallow plateau close to OLMoE's in level ($0.261/0.269/0.270$ at layers 0/1/2 against OLMoE's $0.302/0.283/0.274$) and the same collapse at the last measured layer. As in granite and unlike OLMoE, the very bottom is not strictly monotone (layer 2 edges out layer 0), which supports the reading in the granite paragraph above that the shared feature is concentration in roughly the first fifth of the network rather than the shape at the very bottom. We compare the shape rather than the level because of the caveat above: Qwen holds two experts fixed where OLMoE and granite hold six, so its Γ is conditioned on a thinner background and the two levels do not measure quite the same quantity. The plateau nevertheless lands within a few points of OLMoE's, which is worth noting but not worth resting on. The CSA

bound also replicates, with $|\rho| \leq 0.065$ at all nine layers on both supports, nothing surviving BH, and TOST rejecting $|\rho| = 0.1$ at sixteen of eighteen tests; the two random-arm correlations at layers 4 and 10 depart nominally from zero (§5.2). So does the reliability gap: no layer's $\bar{\Gamma}_{ij}$ reaches a split-half reliability of 0.04, while co-activation on the same sites reaches 0.52-0.71. The part that differs is the matched-sampling form of the reliability contrast. At Top-4 over 60 experts a given pair co-activates in roughly 0.7% of selections, so co-activation subsampled to synergy's 20 observations per pair is itself poorly resolved (0.05-0.17). The cost asymmetry that makes co-activation attractive is therefore sharper here, since its reliability is high only because the observations are free. The matched-n ratio that §5.2 quotes for OLMoE is still not the right summary for a Top-4 router, and we do not quote one.

D NON-ADDITIVITY: FULL PER-LAYER RESULTS AND CONTROLS

INTERACTION ENERGY IS MATERIAL AND CONCENTRATES IN SHALLOW LAYERS

Pooled IER has a median of 0.060 (95% CI [0.053, 0.066]). Because IER is a variance ratio, and the absolute scale behaves differently enough, we report the two separately. Median RMS $|\Gamma|$ per site is $1.8/2.1/2.5/3.1/2.3/0.7/0.18 \times 10^{-3}$ nats at layers $0/1/2/4/6/10/14$, against a replay floor of $\approx 4 \times 10^{-5}$. In absolute terms, then, non-additivity peaks at layer 4 and is largest across the middle of the shallow region, whereas the ratio is monotone from layer 0. The deep layers have both a small Γ and a small Γ relative to everything else, but the shallowest layers are ratio-dominant rather than magnitude-dominant. Both readings agree that the deep half is additive and disagree about where the peak sits. When we call non-additivity "material" we mean it in the variance-share sense, and the absolute quantities remain small, a few times 10^{-3} nats against a validation cross-entropy of ≈ 2.2 ($\ln 9.06$, Table A10); 43.3% of sites exceed 0.1 and 19.5% exceed 0.3. The dominant structure is a depth gradient that is monotone across all seven measured layers, with medians 0.302 / 0.283 / 0.274 / 0.220 / 0.144 / 0.025 / 0.0005 at layers 0/1/2/4/6/10/14 (Table A3; layers 0, 1 and 4 are the later shallow extension, and intermediate layers are not measured). Within-layer entropy stratification shows that the gradient is not a disguised router-uncertainty effect. Layer 2 is flat across entropy terciles (0.250/0.288/0.283), layers 6 and 10 show mild within-layer increases (from 0.125 to 0.169 and from 0.016 to 0.035), and layer 14 is null throughout. A pooled-tercile analysis would have reported a dramatic "entropy gradient", which is an artifact of depth confounding that the stratified design catches. The uniform-weighting control also preserves the gradient (0.202/0.099/0.043/0.012), so the interaction is not a gate-weight artifact.

The domain ordering is a difficulty effect, not a language effect. Pooled by domain, the interaction share of regret is largest for multilingual text (11.9%, against 10.4% / 9.8% / 8.2% for English, code and math), but this ordering does not survive matching. OPUS-100 zh/de is simply harder for an English-centric model, with mean clean log-probability -2.68 against -0.98 to -1.89 elsewhere, and half its sites fall in the hardest global tercile. Within matched difficulty terciles the ordering inverts: multilingual has the largest share among hard sites (12.3%) and the smallest among easy ones (8.9%, against 13.0% and 13.1% for English and math). The part that survives is unconditional and unsurprising. Multilingual text has more absolute interaction gap (95% CI $[+0.0028, +0.0051]$) because it has more total regret (CI $[+0.0221, +0.0339]$). We therefore make no claim that language type modulates interaction; the depth profile holds in every domain and is the domain-invariant result.

The gradient is not a link-function or sensitivity artifact. Two mechanisms could produce a depth gradient in IER without any depth-graded functional interaction. First, U is a log-likelihood and log is concave, so experts that contribute additively to the probability of the true token cannot be fitted additively in log space, and the resulting residual grows with effect size. Second, interventions at shallow layers propagate through more remaining layers, so a statistic that merely tracked "how much does intervening here change anything" would fall with depth. We tested both. Re-fitting the identical decomposition on the probability scale ($\exp U$) leaves the medians essentially unchanged at 0.275 / 0.144 / 0.025 / 0.0005, against 0.274 / 0.144 / 0.025 / 0.0005 on the log scale (granite: 0.216 / 0.135 / 0.036 / 0.002 against 0.216 / 0.134 / 0.035 / 0.002). The effects are small enough (a few times 10^{-3} nats) that the link is locally linear over the measured range. IER also does not track intervention magnitude. The median per-site utility spread (standard deviation) across the 28 coalitions is not monotone in depth, and its ordering contradicts the IER ordering at both ends. Layer 2 has the smallest spread of the four OLMoE layers (0.0066 nats) and the largest IER (0.274), while layer 14

has a larger spread (0.0087) and IER 0.0005. granite reproduces the reversal independently (layer 2: spread 0.0047, IER 0.216; layer 14: spread 0.0092, IER 0.035). Within layers the association is weak and inconsistent in sign (Spearman $+0.075$ / $+0.030$ / -0.049 / -0.296; granite $+0.098$ / $+0.067$ / $+0.022$ / -0.068). Layer 14 therefore also serves as an internal negative control. It uses real data and has a utility spread larger than layer 2's, yet the pipeline reports essentially zero non-additivity there, which gives the calibrated zero point that a synthetic validation of positive detection (App. B) cannot supply.

How much does the two-slot protocol miss?

The three-slot sweep. Everything above is second order in the varied slots, because only two slots move (App. A). To bound what that hides we ran a three-slot sweep. We fix the five strongest routed experts, vary the three most marginal over a pool of $M = 10$, evaluate all $\binom{10}{3} = 120$ equal-compute coalitions per site, and fit two nested models, first additive and then additive-plus-pairwise. The residual of the second is third order and above. The estimator is gated against the App. B synthetic system, whose utility is a quadratic form and so has exactly zero third-order interaction; the estimator returns $< 10^{-10}$ there and detects an injected third-order term (`tests/test_triple_decomposition.py`). The sweep covers seven layers with 250 sites each, window utility and a replay floor of 2.7×10^{-5} nats.

Layer	0	1	2	4	6	10	14
Energy beyond unary, E_1/SS	0.416	0.369	0.333	0.278	0.142	0.034	0.002
Energy beyond pairwise, E_2/SS	0.205	0.180	0.168	0.125	0.059	0.013	0.000
Beyond-pairwise share E_2/E_1	0.498	0.487	0.501	0.476	0.462	0.358	—
95% CI (document-clustered)	[.486,.522]	[.471,.507]	[.481,.524]	[.464,.493]	[.442,.480]	[.318,.383]	—
Per-dimension enrichment $(110/75)E_2/E_1$	0.731	0.714	0.735	0.699	0.677	0.525	—

Table A5. Three-slot sweep, OLMoE, $M = 10$, 250 sites per layer, $w = 8$. The unary design is a subspace of the unary-plus-pairwise design (ranks 10 and 45 in the 120-dimensional coalition space). The mutually orthogonal pieces are the pairwise increment, the unary fit and the third-order-and-above residual, so rows 1-3 are an exact orthogonal energy decomposition and take no degrees-of-freedom correction. Utilities are deterministic and all 120 coalitions are enumerated, and on a surface that is unary-plus-pairwise by construction the estimator returns $E_2/E_1 = 0$ to machine precision (`tests/test_triple_decomposition.py`), so the raw share has no bias to correct. The last row divides the share by $75/110$, the value E_2/E_1 takes when U points in a uniformly random direction. It is a per-dimension diagnostic rather than a share, and since it is below 1 at every layer, pairwise directions carry more than their dimensional share of the non-additive energy. Because the two residual dimensionalities are fixed by the design, this is an exact per-site rescaling by $110/75$ and the last row is row 3 times $1.4\overline{6}$ to the digits shown. Shares are medians of the per-site E_2/E_1 (the `high_order_share` field of the release).

Two things follow, and the second is the more useful. First, Γ is an upper bound on the non-additivity a two-slot design can express (App. A), but the two-slot protocol is a lower bound on total interaction. About half of the non-additive energy at the shallow layers is invisible to any two-slot account (0.498,

0.487, 0.501 at layers 0/1/2), and the share falls to 0.358 at layer 10. Within the three-slot sweep this makes total non-additivity $2.0\times$ the part any pairwise model can express at layer 2, $1/(1-0.501)$. We do not carry that factor across to the headline two-slot IER. The two protocols differ in candidate pool ($M = 10$ against 8), fixed background (five experts against six) and utility window ($w = 8$ against 1), so the three-slot energy beyond unary (0.333 at layer 2) is not the same estimand as the two-slot Γ energy (0.274). Our numbers therefore understate non-additivity, by a factor we bound within one protocol instead of extrapolating across two. Second, the share declines gently with depth, from 0.50 to 0.36 across layers 0 through 10, so the two orders scale together closely enough that the depth profile of total non-additivity has the same shape as the pairwise one. Every depth conclusion in this paper keeps its shape under the extension, and its magnitude changes by a factor that is measured and not assumed.

At layer 14 both quantities sit at the noise floor, so their ratio is ill-conditioned and we do not report it (the per-site mean of 0.113 is an artifact of dividing near-zero by near-zero; the median is 0.001). The meaningful statement there is absolute. Energy beyond the pairwise model is 0.000, so the "almost perfectly additive" reading below survives a test at third order as well as at second. An independent run at layer 2 with single-token rather than window utility gives $E_1/\mathrm{SS} = 0.348$ and a share of 0.494, against 0.333 and 0.501 above, so the quantity is stable across utilities and across runs. Its records are in `results/triple_pilot_m10_seed0.jsonl` and are tabulated beside the granite pilot in `results/analysis_granite_triple.txt`, which despite its name holds both pilots. The same single-token pilot on granite (250 sites per layer, 50 documents) gives 0.446 $[0.427, 0.466]$ at layer 2 and 0.372 $[0.315, 0.393]$ at layer 14 of 24, under granite's own renormalized gates with no uniform-gate run. Those two layers are all the three-slot data granite has, and Qwen has none.

One caution carries to §5.2. The co-activation result is a statement about the pairwise component, and a co-activation statistic could in principle track third-order structure we have not tested it against.

Gate mass. One more mechanical explanation deserves closing. OLMoE's gates are a raw softmax over all 64 experts with no Top-K renormalization, so if the two perturbed slots simply carried less gate weight at depth, a vanishing IER would follow trivially. This is not what happens. Mean gate mass on the two varied slots rises with depth (0.047, 0.053, 0.064, 0.066 at layers 2/6/10/14) while IER falls from 0.274 to 0.0005. By this measure the perturbed slots matter more at depth, and they are still additive there.

The same layers sharpen the sensitivity control. Layers 0 and 1 have the smallest median utility spread of any layer measured (0.00465, 0.00467, against 0.0090 at layer 4 and 0.0104 at layer 6) and the largest non-additivity. Across all seven layers the rank correlation between median IER and median utility spread is negative, -0.79 (exact $p = 0.048$, $n = 7$; -0.75, $p = 0.066$, with each site's range in place of its standard deviation; descriptive only at this sample size). Whatever IER tracks, it is not how much intervening at a site changes the output.

Both slot protocols show the same depth gradient, and the strong slots show less relative non-additivity. Perturbing the two strongest routed slots produces $3.15\times$ the utility spread, so the effects are absolutely larger, yet the IER is roughly halved (pooled 0.038 against 0.064 on matched sites) and the interaction share of regret falls from 10.9% to 7.4%. The marginal-slot protocol thus measures the smallest interaction effects in absolute terms and the largest in relative terms. Since the two orderings differ, we report them separately. The monotone decline with depth holds under both.

The uniform-gate control needs one further note, because it reverses at the two deepest layers (0.043 vs. 0.025 at layer 10; 0.012 vs. 0.0005 at layer 14). Uniform weighting is not a small perturbation there: it multiplies the median utility spread by $4.2\times$ at layer 10 and $7.1\times$ at layer 14, against $4.1\times$ at layer 2. On this model uniform weighting is simply a larger intervention at depth, which accounts for the reversal without any claim about where weighting headroom lives.

Where the non-additivity arises. OLMoE and Qwen do not renormalize, so under any coalition the intervened layer's output is exactly $\sum_{i \in S} w_i f_i(h)$ (plus, on Qwen, a shared-expert term that does not depend on S), with every w_i and f_i fixed by the unchanged layer input. Members therefore do not interact inside the layer. On these two models all of Γ is produced downstream, and on granite, renormalization adds a within-layer term that the uniform control removes (App. C). Downstream there are smooth non-linearities (normalization, attention, the expert MLPs, the later gates' softmaxes, the output log-softmax) and the discrete Top-K selection of every later router, which the deployment-faithful semantics leave free (App. A). We do not separate the two, and no record stores downstream

selections. Depth cannot separate them either, since every layer of all three models is an MoE layer and the number of later routers equals the number of later layers. The near-zero transfer of Γ across transplanted backgrounds (§5.2) does not separate them, because it excludes only a downstream map whose curvature does not depend on the background. Logging each coalition's downstream selections would give flip rates at negligible cost per sweep, and a sweep with every later router's selection frozen to its clean route, with weights recomputed or also frozen, would separate the two at about the cost of a sweep. As measured, Γ is the non-additivity of the model's output with re-routing included. That is the quantity a router choosing S at layer l is scored on, and the statistics of §5.2 are properties of that Γ whatever produces it. Two questions are open: whether Γ reflects the experts' joint function or a routing cascade, and whether a routing-frozen Γ has a more reproducible pair component.

What this implies for the "expert specialization" picture

Combining §5.1-5.3 and App. C-F, and stating the claim only at the order and support we measured, we conclude that non-additivity among the most marginal of eight routed experts is material at shallow layers and lies about half beyond any two-slot account, and that its conditional pairwise part has almost no stable pair mean. Three qualifications apply to this conclusion. It concerns pairs at a fixed background and says nothing about eight-member sets. Interaction among three or more candidates lies outside the design's support, and structure coupling a single candidate to the fixed background is absorbed into the unary terms (App. A), so we can neither confirm nor exclude either. It also concerns only the marginal slots; the top-slot protocol shows the same monotone depth decline with $3.15\times$ the utility spread but roughly half the relative non-additivity (App. D). Finally, on OLMoE it covers seven layers of sixteen, including both layers 0 and 1, where the profile plateaus rather than spiking; granite is measured at six of twenty-four.

Within those limits, the picture runs opposite to the usual specialization narrative in one specific respect. Non-additivity is concentrated at shallow layers, where semantic expert identity is least often reported, while deep-layer utility is almost perfectly additive in the marginal slots (although Wang et al. (2026a) report near-identical deep-layer routing across semantically unrelated inputs in five other MoEs). The non-additivity that does exist has almost no stable pair mean, since a pair's average does not persist across the contexts in which the same two experts meet (§5.2). This argues against a fixed weight per pair but leaves room for pair-specific structure that changes with context. Whatever shallow-layer expert combinations are doing, they do not amount to the stable, nameable pairwise collaboration that co-activation graphs and "expert group" analyses posit.

E Co-activation and interventional synergy: full results

On OLMoE, for every layer, the Spearman correlation between pair co-activation counts (within natural Top-8 selections, on the same sites) and mean measured synergy $\bar{\Gamma}_{ij}$ is null: -0.014 / $+0.022$ / $+0.002$ / -0.003, with 95% intervals of half-width ≈ 0.046 over 2,015-2,016 pairs (Fisher-z, standard error $1.06/\sqrt{n-3}$; Table A6). The correlations are equally null on the de-censored support, the pairs containing at least one uniformly random candidate, which only the random arm makes measurable. That support is de-censored rather than unbiased. The random member is uniform over the 53 experts outside the site's top eleven, but its partner, the six fixed experts and the exclusion of the top eleven are all determined by the router. The strictest support available, pairs in which both members are uniformly drawn, is also null: $\rho = -0.000/-0.023/-0.011/-0.004$ at layers 2/6/10/14 over 1,580-1,606 pairs with at least two observations each ($p = 0.998/0.383/0.672/0.891$). Every correlation interval, p-value and TOST in this section uses the Fisher-z transform with the pair-level standard error $1.06/\sqrt{n-3}$ over n pairs, so the validity of that standard error needs checking. It is about 3% wider than $\sqrt{1.06/(n-3)}$, the Spearman variance of Fieller et al. (1957), and about 6% wider than the $\sqrt{(1+\rho^2/2)/(n-3)}$ of Bonett and Wright (2000), which makes it the cautious choice for declaring a departure from zero and the demanding one for declaring equivalence. It treats the per-pair means as independent. They are not exactly independent, since sites within a document share a context and every expert appears in many pairs. These intervals target the correlation observable at our sampling depth, between co-activation counts and the pair means of Γ over the observations each pair actually received, for this model's pairs, across repeated draws of as many documents as we have. They do not target the correlation between population pair means. Noise unrelated to co-activation only shrinks a correlation, so the two share their zero and a test of $\rho = 0$ is valid for both. The

equivalence margin does not transfer. Attenuation gives $|\rho_{\text{obs}}| \leq \sqrt{r_\Gamma r_C}\,|\rho_{\text{true}}|$, and at $r_\Gamma \approx 0.02$ a bound of 0.1 on the observable correlation barely constrains the population one, which is why we rest that claim on the reliability (below) rather than on the TOST.

To check the standard error we use a reference that draws no document twice (`scripts/csa_calibrated.py`): split the documents at random into two disjoint halves, compute ρ on each, and take $\sigma = \text{sd}(z_A - z_B)/\sqrt{2}$ over 2,000 splits per test, which gives the empirical standard error at half size. Over the 44 tests below, σ is $0.92\times$ (median; range 0.67-1.07, above one at three tests) the pair-level value at the same pair count ($0.85\times$ on OLMoE and granite, $0.95\times$ on Qwen). Matching size does not establish coverage, so we also count how often the two halves' 95% intervals agree, $|z_A - z_B| < 1.96\,(\text{SE}_A^2 + \text{SE}_B^2)^{1/2}$, which has probability 0.95 when both standard errors are right and z is normal (`scripts/csa_bootstrap_corrected.py`, which lists every test's σ, pair-level and bootstrap standard errors and agreement rates in `results/csa_bootstrap_corrected.txt`). With the pair-level value they agree in 96.9% of splits (97.2% on OLMoE, 97.8% on granite, 96.2% on Qwen); per test the rate ranges from 93.2% to 99.9% and falls below 95% at five tests, lowest at OLMoE layer 10 on the random arm, where σ exceeds the pair-level value most. Each test's rate is within 0.009 of what a normal z with that test's σ implies, so this check confirms the tails rather than adding a second measure of size; widening those five tests' standard errors until each reaches exactly 95% changes nothing below. Because half size is not full size (each pair has half its observations), we also checked that the calibration does not drift with sample size. When pair means are mostly noise, halving every pair's observations rescales that noise without changing the ranks it produces, so under the null the spread of z is set by the number of pairs and the dependence among them rather than by the number of documents. The data agree: in the 16 tests where halving the documents again barely changes the pair count, the quarter-to-half ratio of σ is 1.02 (median), against $\sqrt{2}$ under document scaling; the ratio of σ to the pair-level value is 0.92 at quarter size as at half (median; quarter over half 0.95-1.17 per test); and agreement at quarter size is again 96.9%. The procedure also passes a test with known truth. In synthetic data built like ours (2,016 pairs of 64 experts, 136 documents, a pair-mean reliability of about 0.02 by design; 8,000 datasets per design, so each coverage carries a Monte Carlo standard error of 0.2-0.4 points), σ from one dataset's halves, carried to the full pair count, gives full-sample 95% intervals that cover 94.9% when pairs are independent and 94.6% when pairs share experts and documents share a context (a design in which the pair-level value covers only 89.1%), and 94.0% with a planted observable correlation of 0.11, beyond the equivalence margin. At full size we compared disjoint document sets from our own data. At nine layers the family's sample is set against documents it never used (136 new OLMoE and 380 new granite documents; at OLMoE layers 0, 1 and 4 the two sets are the same size), and at granite layers 0 and 1 the original 136-document run is set against the 380 documents the extended run added. The standardized difference $(z_1 - z_2)/(\text{SE}_1^2 + \text{SE}_2^2)^{1/2}$ has mean square 0.88 with the pair-level standard errors (1 if calibrated; χ^2 over the 11 layers, $p = 0.56$ on either support) and never exceeds 1.96 in the 22 comparisons. With eleven layers this check is coarse and cannot exclude standard errors off by a factor of 1.5 in either direction, but it is the only real-data check at full size. We call the pair-level interval conservative only in the sense verified here, nominal or better at 39 of 44 tests at half size and 41 of 44 at quarter size and 93.2% at worst, and not at every test.

A document bootstrap is not a valid alternative. In such a bootstrap, documents are resampled with replacement, every drawn copy counts as a separate document so that co-activation and synergy both enter with multiplicity, and both quantities are rebuilt inside each draw (`scripts/csa_bootstrap_corrected.py`). It gives standard errors a median $0.67\times$ the pair-level value and fails every check. Inside the same halves it gives $0.78 \times \sigma$ (median; below one at all 44 tests); its half-sample intervals agree in 87% of splits, below the pair-level rate at every test; across the disjoint document sets its standardized differences have mean square about 2 (χ^2 $p < 0.04$ on both supports); and in the synthetic designs above its full-sample intervals cover only 83–85%. The failure is structural. A bootstrap treats the observed pair means as the truth, but they are already mostly noise. Each draw adds a second, equally large layer of within-pair noise, so in the bootstrap world the reliability of a pair mean is about one half instead of 0.02. Its replicates therefore shrink toward zero, to $0.71\times$ the estimate across the 44 tests (the $1/\sqrt{2 - r_\Gamma} \approx 0.71$ this predicts at $r_\Gamma \approx 0.02$), and spread too little. The document bootstraps behind this paper's other intervals are of site-level means, medians, ratios and gains, which re-average no pair means, so the failure does not carry over to them. We do not use it here, and its narrower intervals are not valid. The release also keeps `scripts/csa_clustered.py`, a variant that counts a document drawn twice only once in co-activation; on

the same draws its standard errors differ from these by under 2% at the median (at most 6%), and none of its intervals is used (App. L). None of the conclusions below depends on the choice. With Bonett and Wright's variance, with Fieller, Hartley and Pearson's, or with the calibrated σ itself, nothing survives BH-FDR (minimum $q = 0.28$, 0.36, 0.44) and equivalence holds at 35, 34 and 36 of 44 tests; Bonett and Wright's slightly narrower interval adds one nominal departure, Qwen layer 0 on the random arm ($p = 0.049$). The document bootstrap alone would turn the two nominal results into BH departures (Qwen layers 4 and 10 on the random arm, $q < 0.04$) and establish equivalence at all but four of the 44 tests, as intervals about 22% too narrow would; rescaled test by test by its own shortfall inside the halves, it gives no departure.

The family is the forty-four tests run under each model's own gate weighting, each on the full candidate pool and on the random-arm support: OLMoE at layers 0, 1, 2, 4, 6, 10, 14 (14 tests; layers 0, 1 and 4 from the one-seed shallow extension, 680 sites per layer, the rest from the three-seed sweep, 2,040), granite at 0, 1, 2, 8, 14, 20 (12; layers 0 and 1 from the extended run, a 516-document rerun with 2,580 sites per layer that re-measures the 136 original documents at a new draw of positions and adds 380 unseen ones, and layers 2, 8, 14 and 20 from the original run, 680 sites over those 136 documents) and Qwen at 0, 1, 2, 4, 6, 10, 14, 18, 22 (18; one run, 272 documents, 1,360 sites per layer). Across all forty-four the largest correlation anywhere is $|\rho| = 0.084$, and none departs from zero after BH-FDR (minimum $q = 0.45$). Two do so nominally, both on Qwen's random arm: layer 4 ($\rho = -0.065$, $p = 0.010$) and layer 10 ($\rho = +0.056$, $p = 0.026$). They point in opposite directions and each explains under 0.5% of rank variance; two raw rejections are what 44 tests at $\alpha = 0.05$ produce by chance, and neither is a pattern a co-activation graph could act on. Equivalence against $|\rho| = 0.1$ is established at 32 of 44 after BH-FDR: all fourteen OLMoE tests, sixteen of Qwen's eighteen (all but the two above), and on granite only layer 0, where the extended run gives $\rho = -0.013$ on the full pool and $+0.001$ on the random arm (TOST $p = 0.034$ and 0.019; $q = 0.046$ and 0.027). Granite's 406-496 pairs give intervals of half-width ≈ 0.1, the size of the margin itself, so its values are consistent with the null but do not establish it. Table A12 lists all 44 tests with the run and sample set behind each.

The choice of granite run changes the equivalence count but leaves the null intact. At granite layers 0 and 1 two different 2,580-site sets exist over the same 516 documents: the extended run, and a pooled set that joins the original run's own 680 sites to the extended run's 1,900 sites on unseen documents (quoted below). On the pooled set layer 0 gives $+0.074$ and $+0.059$ (TOST $p = 0.29$ and 0.19), and equivalence would hold at 30 of 44 tests and at no granite test. The family uses the extended run at layers 0 and 1, and that choice was made after the fact. The extended run was sampled to re-examine the original run's layer-0 result (below), so which granite run fills those rows was not fixed in advance (App. L). The deep-layer replication of layers 14 and 20, run later and on a different stack (App. J), is reported beside the family rather than substituted into it. With the original 680 sites at every granite layer, equivalence would hold at 30 of 44 (largest $|\rho| = 0.116$); with the 516-document runs wherever they exist, at 38; with the pooled construction at layers 0, 1, 14 and 20, at 32, but on the full pool at layers 14 and 20 rather than at layer 0. No choice yields a departure from zero after BH-FDR (minimum q between 0.33 and 0.45; `scripts/csa_family_sensitivity.py`).

Granite's deepest layers leaned negative and did not replicate. At granite layer 20, the deepest layer measured on that model, the original 680 sites give $\rho = -0.074$ on the full pool and -0.084 on the random arm ($p = 0.12$ and 0.077), and layer 14 leans the same way (-0.057). The sign is the opposite of what co-activation-graph methods predict (pairs the router co-activates more often carry slightly less interaction energy), and -0.084 is the largest $|\rho|$ in the family, so we replicated it with granite layers 14 and 20 at 2,580 sites over 516 documents, split at the document boundary. On the 1,900 sites from 380 documents the original run never saw, layer 20 gives $\rho = +0.020$ ($p = 0.68$) and layer 14 gives $+0.015$ ($p = 0.75$), both reversed in sign; pooled over the original and unseen sites, layer 20 is -0.004 ($p = 0.94$). A uniform-weighting control on the original sites reproduces the lean (-0.083), which rules out gate arithmetic but not sampling, since it reuses the same documents. We read both deep-layer leans as sampling variation, as we read granite layer 0 below. The uniform-gate CSA values quoted in App. D are a weighting control and are not in this family; the sixteen within-domain tests below are corrected as their own family.

Everything above is pair-indexed, whereas the expert-group literature does not use its graphs one edge at a time: it partitions experts into communities and acts on those. A pair-level null does not by itself refute an aggregate claim, so we also asked whether the null holds for expert groups. We built a standard co-activation graph on the same models (popularity-normalised lift $C_{ij}N/n_i n_j$ over the

natural Top-8, thresholded at the top decile), partitioned it by greedy modularity, and tested whether measured pairs inside a community have larger $|\Gamma|$ than a matched random grouping. Community labels were permuted within quartiles of expert activation frequency, with 4,000 permutations, stratified by candidate provenance because provenance is itself associated with $|\Gamma|$.

On the original sampling the question is unresolved. On the de-censored support the within-group excess was positive and nominally significant at two layers ($p = 0.007$ at layer 2 and $p = 0.021$ at layer 6, against $p \geq 0.26$ at layers 0, 1, 4, 10 and 14), but it does not survive the specifications we would have had to choose between in advance. Moving the edge threshold from the top decile to the top quintile gives $p = 0.027$ at layer 2 and 0.160 at layer 6; the top 15% gives 0.122 and 0.095; the top 5% gives 0.118 and 0.468 and instead makes layer 0 significant at 0.023. Scoring signed Γ rather than $|\Gamma|$ moves the effect again, to layer 4 ($p = 0.007$) with layer 2 marginal (0.053) and layer 6 null (0.170). Within the primary specification, layer 2 sits at the BH-FDR boundary across the seven layers ($q = 0.046$ with 100,000 permutations) and layer 6 lies above it (0.064). On the strictest support, where both members are drawn uniformly, the minimum p over seven layers is 0.132.

A replication has to hold the discovered object fixed. We added 9,520 sites at the same seven layers (a fresh seed, 136 of them new documents). Rebuilding the partition on the pooled records would test a different partition from the one that produced the discovery: the rebuilt layer-2 partition agrees with the original at an adjusted Rand index of only 0.56, and gives -3.07×10^{-5} nats ($p = 0.565$) on the pooled records and $+1.0 \times 10^{-4}$ ($p = 0.32$) even on the original ones. The test that answers the question fixes the partition G_0 built on the original records and scores it on the added sites (`scripts/group_level_heldout.py`; same statistic, null, 4,000 permutations and seed). At layer 2 the effect holds, with $+6.99 \times 10^{-4}$ nats on the 1,360 added sites ($p = 0.027$) and $+1.13 \times 10^{-3}$ ($p = 0.038$) on the 680 of them drawn from the 136 documents G_0 never saw, which is no smaller than the original $+5.70 \times 10^{-4}$. At layer 6 it does not hold ($+2.86 \times 10^{-4}$, $p = 0.143$; new documents $p = 0.176$). The p values at layers 2 and 6 come from 100,000 permutations, because at 4,000 their Monte Carlo error is as large as their distance from a BH boundary. Layers 0, 1, 4 and 10 are null on both samples; layer 14, null on the original records ($p = 0.86$), gives $p = 0.033$ on the added ones. Even within the two layers the original sampling pointed to, the layer-2 replication falls just short of BH-FDR ($q = 0.054$; 0.076 on the new documents alone), and across all seven layers it is further from it ($q = 0.115$ and 0.267).

We do not call it an effect, for three reasons. First, it belongs to one specification: G_0 rebuilt at the top 20%, 15% or 5% of edges does not replicate on the added sites ($p = 0.54$, 0.79, 0.35), and scoring signed rather than absolute Γ reverses it ($p = 0.96$). Second, a partition-free version of the question, asking whether pairs on the strongest co-activation edges (the top 20%, 10% or 5% of lift on the original records) carry more $|\Gamma|$ than the rest, shows nothing at layer 2 on the added sites ($p > 0.70$; ≥ 0.62 on the new documents alone). Third, it is small, since 7×10^{-4} nats is about a tenth of the mean $|\Gamma|$ at that layer (6.7×10^{-3}). We therefore report the group-level question as unresolved. At one layer and one threshold, a co-activation partition concentrates slightly more pairwise interaction energy on documents it was not built from, and nothing about it carries across specifications. Two scope limits apply to that reading. The statistic is the mean $|\Gamma|$ of measured pairs inside a community against a matched random grouping, so it tests whether a co-activation partition concentrates pairwise interaction energy rather than whether an expert group has joint function. Group-level utility could live in unary contributions, in third-order and higher terms, or across layers, and none of those would have to show up as larger within-community pair residuals at the same layer. The pair-level results of this section are unaffected either way. They concern whether a single pair has a reproducible synergy, and a group-level excess averaged over many pairs is compatible with per-pair means that do not reproduce.

The first-sample departure at granite did not replicate. On the original 680 sites, the smallest p among granite's tests was at granite layer 0, the shallowest layer of any model measured: $\rho = +0.116$ over 496 pairs, raw $p = 0.015$, 95% CI $[+0.023, +0.207]$, an interval that excludes zero. It was the only granite test with raw p below 0.05 on the original sampling, and it was positive, in the direction co-activation-graph methods would predict. On that sample it was bounded in three ways. It does not survive BH-FDR; with the original run's rows in the 44-test family its q is 0.33 (`scripts/csa_family_sensitivity.py`). On the de-censored random-arm support the same layer gives $+0.088$ with CI $[-0.006, +0.179]$. That interval is consistent with the full-pool estimate, so the subset does not independently reach significance, which is different from the signal failing to reproduce. And it rests on granite's 496 pairs rather than OLMoE's 2,016. This called for more sampling at

that site, so we ran it. The extended granite sweep measures layers 0 and 1 at 2,580 sites (3.8× the original 680) over 516 documents, and we split it at the document boundary. On the 1,900 sites drawn from 380 documents the original run never saw, $\rho = +0.0009$ with $p = 0.99$ (95% CI $[-0.092, +0.094]$). Redrawing the positions within the same 136 documents (174 of the 680 coincide with original ones) already gives -0.038 ($p = 0.43$), and the extended run as a whole, which is the set the family uses, gives -0.013 ($p = 0.78$). A different 2,580-site set, the original run's own 680 sites pooled with the 1,900 unseen ones, gives $+0.074$ ($p = 0.12$; TOST $p = 0.29$, not equivalent). The departure does not replicate, and we read the original $+0.116$ as sampling variation. The extension sharpens the reliability statement more than the correlation. On the extended run, at a median of 138 observations per pair, the split-half reliability of $\bar{\Gamma}_{ij}$ at granite layer 0 is -0.024, while co-activation at that same per-pair sampling reaches 0.892. Four times the data leaves the synergy side where it was. Because failure to reject is not evidence of absence, we also report an equivalence test. Against $|\rho| = 0.1$, TOST rejects an observable correlation that large at every OLMoE layer ($p < 0.006$) on both supports; this bounds the observable correlation rather than the population one (above). On granite the same procedure is underpowered, with 406-496 pairs rather than 2,016 and an interval half-width of 0.093-0.103, and equivalence is established only at layer 0 of the extended run (two of twelve tests; the pooled set above would give none). The granite values are consistent with the OLMoE null but do not independently establish it.

The four layers above were chosen before the depth profile was known. Layers 0 and 1, which carry the highest interaction energy (App. D), were added later and are equally null, with $\rho = -0.009$ and $+0.003$ over ≈1,975 pairs (95% CI half-width 0.047), and $+0.014$ at layer 4. These are the layers where the test matters most, since a null is otherwise easiest to obtain where there is least to find. At layer 14 the interaction energy is at the numerical floor, so a null there is nearly vacuous; on OLMoE the null holds at the shallow layers where non-additivity is largest.

Conditioning the test does not recover an association. Pooling four domains is precisely the design that would hide a domain-conditioned association, so we test for one directly. Recomputing the correlation within each domain (co-activation counted inside the domain, synergy averaged inside it) gives 16 tests (4 layers × 4 domains, 1,594-1,901 pairs each) with $|\rho|$ never exceeding 0.032, a mean of -0.001, and a minimum BH-adjusted q of 0.948. None is significant at any threshold, so co-activation does not track within-domain synergy either. Stratifying instead by token surface class and by position quintile within the document gives a separate family of 44 tests over the seven OLMoE layers, with $|\rho| \leq 0.068$, a smallest raw p of 0.068 and nothing surviving BH-FDR (`scripts/token_conditioned_csa.py`; p from scipy's Spearman test, which the pair-level standard error would only enlarge).

The result also holds when co-activation is normalised for expert popularity or estimated from more positions. The pair-level tests above use raw counts C_{ij}, which rank pairs partly by how often each member is routed at all; on the full pool, Spearman$(C_{ij}, n_i n_j)$ is 0.35–0.60 on OLMoE, 0.39–0.84 on granite and 0.10–0.26 on Qwen. Replacing C_{ij} by the lift $C_{ij}N/n_i n_j$ of the group-level test above (PMI is its logarithm, so its rank correlations are identical) leaves the family null. Over the same 44 tests $|\rho| \leq 0.087$ (granite layer 20, random arm; the lean that did not replicate above), with three raw $p < 0.05$ (Qwen layers 0, 4 and 10 on the random arm: $+0.050$, -0.063, $+0.056$), none after BH-FDR (minimum $q = 0.55$), and equivalence at 31 of 44 (granite layer 0 drops out, layer 8's full pool enters). With the original run at granite layers 0 and 1, lift gives $+0.125$ at layer 0 on the full pool ($p = 0.009$, against $+0.116$ for counts) and still no BH departure (minimum $q = 0.28$). Normalised PMI, the ϕ coefficient, Jaccard and a shrunk lift $(C_{ij} + 5)/(E_{ij} + 5)$, $E_{ij} = n_i n_j/N$, give $|\rho| \leq 0.113$ (ϕ at granite layer 2, random arm; two to five nominal tests per measure) and no BH departure (minimum q 0.40–0.54 per measure, 0.55 over all 220). Popularity alone, $n_i n_j$, gives $|\rho| \leq 0.036$, and on pairs whose both members are uniformly drawn, lift gives $|\rho| \leq 0.020$ at layers 2/6/10/14. The co-activation side is also not short of positions. Lift is reliable at our sampling (split-half over positions, 100 random splits: 0.87–0.95 at OLMoE layers 2/6/10/14, 0.63–0.75 at 0/1/4, 0.73–0.99 on granite and 0.61–0.80 on Qwen, whose Top-4 of 60 leaves a median of two to four co-activations per tested pair). Re-estimating co-activation from every distinct position in the release's formal runs at that layer, which gives 2.5–3.3× the positions on OLMoE, up to 4.9× on granite and essentially none added on Qwen, moves the pair ranking little (Spearman 0.85–1.00 with the same-site lift, whose positions it contains) and leaves the family null ($|\rho| \leq 0.079$, none after BH, equivalence at 32 of 44). Neither choice changes r_Γ, and the attenuation bound below turns on r_Γ (`scripts/csa_normalised.py`, `scripts/csa_pooled_positions.py`).

The deeper reason is that $\bar{\Gamma}_{ij}$ barely reproduces across documents. A correlation between co-activation and synergy presupposes that a pair has a synergy that is stable across the sites where it is measured, and at our sampling this barely holds. Splitting each pair's observations at random into halves and correlating the two half-means across pairs (Spearman-Brown corrected, 200 random splits) gives a reliability of 0.020 / 0.014 / 0.007 / < 0 at layers 2/6/10/14, despite a median of 25-27 observations per pair; granite's highest is 0.105, on the random-arm support at layer 20 (`scripts/pair_reliability.py`). Very little of the variance in a pair's mean measured synergy is stable across the sites at which that pair is measured. This has two consequences, and we state the second because it limits our own evidence. First, the contrast with co-activation has to be made at matched sampling. Measured on the same sites and over the same pairs, co-activation's split-half reliability is 0.906 / 0.938 / 0.958 / 0.947 at layers 2/6/10/14, against synergy's 0.020 / 0.014 / 0.007 / < 0. Those two rows differ in per-pair sample size, since co-activation is readable at every position (2,013 observations per pair) while each synergy observation costs 28 forward passes (median 25-27 per pair). Subsampling co-activation to each pair's own synergy sample size and averaging over 20 independent draws gives 0.319 ± 0.028 / 0.451 ± 0.018 / 0.525 ± 0.019 / 0.518 ± 0.018. That is still far above synergy's at every layer, but well short of the factor the unmatched comparison suggests. We state the contrast as a difference rather than a ratio, because synergy's reliability is too close to zero for a ratio to be a stable number. Each synergy figure is a mean over 200 random splits, and the same estimator on `scripts/reliability_sensitivity.py`'s own random splits gives 0.020 / 0.020 / -0.001 / -0.056, so the layer-10 entry is indistinguishable from zero and a sampling multiplier computed from it is a point estimate rather than a bound. We regard the cost asymmetry as a finding in its own right. The router supplies co-activation observations for free, so co-activation's reliability rises to 0.91-0.96 at no marginal cost, and synergy's does not. The same asymmetry explains the shallow-extension layers, where co-activation reaches only 0.800 / 0.754 / 0.835 at layers 0/1/4 on 680 positions while synergy there does not rise. Co-activation is a stable property of a pair and synergy is not. Second, this largely predetermines the correlation tests above. Classical attenuation gives $|\rho_{\text{obs}}| \leq \sqrt{r_\Gamma\, r_C}\, |\rho_{\text{true}}|$, so at $r_\Gamma \approx 0.02$ even a true correlation of 0.5 would appear as ≈ 0.07, which lies inside our own equivalence bound. The TOST result is therefore close to an arithmetic consequence of the reliability and is not independent evidence about co-activation and synergy. We keep it as a descriptive bound on what is observable and rest the claim on the reliability itself.

We also state the claim at the strength the data support. Reliability is a function of the number of observations per pair: $r = \sigma_b^2/(\sigma_b^2 + \sigma_w^2/n)$ with $r \approx 0.02$ at $n \approx 25$ implies a between-pair standard deviation of roughly 3% of the within-pair (context) standard deviation. That is very small but not zero, and roughly 10^3 observations per pair would resolve it. What we can defend is that the pair-level component is estimated at a few percent of the context-level variation and bounded at 14.4% of it (below); we do not claim that the quantity does not exist.

The direct test uses documents the pair was never measured on. A near-zero split-half estimate cannot serve as an upper bound on its own, and splitting the same documents cannot say whether a pair's mean reproduces on new text. Later runs measured the same pairs on documents outside the original runs, namely 136 new OLMoE documents (the 9,520-site extension) and 380 new granite documents (the extended and deep-layer sweeps), so we correlate each pair's mean Γ on the original documents with its mean on the new ones (`scripts/reliability_fresh_documents.py`). We use the Spearman correlation over pairs with at least two observations on each side, with Fisher-z intervals based on the same pair-level standard error. Shuffle, relabelling and document-half checks put the empirical standard deviation at 0.83-$1.11\times$ that error; where a check exceeds one we widen the interval by that factor, which moves no rank bound below by more than 0.3 points and neither maximum. Over the eleven layers with fresh documents the cross-sample correlation is at most $|\rho| = 0.035$ on OLMoE and 0.054 on granite, every interval covering zero, with upper ends of $+0.071$ and $+0.146$ (Table A11). Turning that into a variance ratio takes a model, and the two we can state give different answers. Write a pair's mean on one document set as $\bar{\Gamma}_{ij} = \theta_{ij} + \bar{\varepsilon}_{ij}$, with σ_b^2 the variance of θ across pairs and σ_w the root-mean-square within-pair standard deviation; the within-pair standard deviation varies across pairs with a coefficient of variation of 0.53-1.05, so σ_w is 1.3-$2.2\times$ the median pair's. The first model is direct. The two sets share no document, so a pair's errors on them are independent and the covariance across pairs of its two means is unbiased for σ_b^2 whatever the correlation scale, the pair-specific within-pair variances or the counts n_{ij}. A pair recurs only 1.03-1.19 times in a document it appears in, so the document-clustered variance of each pair mean is 0.99-$1.01\times$ the independent-observation value; the dependence that does matter, between pairs

measured at the same sites, is carried into the interval by a stratified delete-one-document jackknife (`scripts/reliability_direct_covariance.py`). That jackknife is conservative, since deleting documents from each set in turn counts the product of the two sets' noise twice, and disjoint document halves put the empirical standard deviation at 0.58-0.75× its value (2,000 splits). We do not shrink it. In simulations that inject a known pair component into the real records (Gaussian, confined to 2% of pairs, or proportional to each pair's own standard deviation), the shrunk interval under-covers and misses the truth in up to 11% of draws, while the jackknife's upper end fell below the truth no more often than the nominal 2.5% of draws in every setting (at most 2.2% of the 2,000 draws in any of 180 settings). At the per-layer upper ends it bounds the between-pair standard deviation at 14.4% **of the within-pair one on OLMoE and** 7.5% **on granite**, both at layer 14 (16.0% and 8.2% with a Bonferroni correction over each model's layers); every interval covers zero, and the point estimates run from −5.0% to +8.6%. The second model goes through the attenuation identity. If every pair shared one within-pair variance and the pair effects were Gaussian, two independent means would correlate at $\tau^2/\sqrt{(\tau^2+1/n_A)(\tau^2+1/n_B)}$ with $\tau=\sigma_b/\sigma_w$ and n_A, n_B the harmonic-mean observations per pair; solving at the Spearman upper ends gives 9.1% **on OLMoE and** 5.4% **on granite** (8.2% and 5.3% with the median n; 9.4% and 5.9% with the harmonic n divided by the D of Table A11; up to 10.4% and 5.9% with $\overline{s^2}/\overline{s^2/n}$, which allows pair-specific variances, in place of the harmonic n). That figure is conditional on the model, which the data do not satisfy. The identity is for a Pearson correlation, and the pair means have excess kurtosis of 2-190. In the same simulations the rank-based bound held for Gaussian components unrelated to the pair's own variance up to an injected τ of 0.15 (worst 0.2% of draws; 4.0% in one of 60 Gaussian settings, at $\tau=0.25$), but missed the truth in up to 100% of draws when the component was confined to 2% of pairs and in up to 94% when it scaled with the pair's own standard deviation. It cannot exclude a sparse set of strongly interacting pairs, whereas the direct bound showed no such blind spot in the same simulations. Variance-scale correlations are fragile here for the same reason: at OLMoE layer 14 the Pearson correlation is +0.078 against a Spearman +0.025, and it falls to +0.016 without three of the 1,984 pairs, each of which owes its deviation on the new documents to a single document. Qwen, and granite layers 2 and 8, have no fresh documents and are not covered by either bound. We also computed split-half reliability with whole documents assigned to each half, over 200 random splits of the same documents (hence different from Table A6's observation-level figures): 0.013 / −0.020 / 0.012 / 0.034 / 0.017 / −0.004 / −0.057 at OLMoE layers 0/1/2/4/6/10/14 (`scripts/reliability_interval.py`). These values are descriptive only. Their 2.5-97.5 percentiles describe the estimator's dispersion over splits of fixed documents rather than a population interval, and their upper ends (0.109 / 0.105 / 0.160 on OLMoE / Qwen / granite) are not a bound (App. L). Resampling documents with replacement before splitting is worse still, because identical observations land in both halves and inflate the reliability to $[+0.41, +0.59]$. An expert-interaction graph would need roughly 36-140× our per-pair sampling on OLMoE before its edges reached even a reliability of 0.5. The factor is $(1-r)/r$ over Table A6's split-half reliabilities and is therefore set by the layer with the highest reliability, which is why granite's most favourable estimate ($r=0.105$) would need only ≈8× its own. These are point estimates. At the upper ends above, a reliability of 0.5 would need 49 observations per pair at OLMoE layer 14 (direct bound; ≈2× the median 25) or 120 at layer 4 (rank bound; ≈13× the median 9). Interaction here is a property of a combination of pair, context and conditioning set; any pair-level component of the signed mean is too small to resolve at our sampling.

Model	Layer	Pairs	n_A / n_B: median (harmonic)	D_A / D_B	Spearman ρ [95%]	Pearson ρ	τ rank (%)	τ rank, n/D (%)	$\hat{\tau}$ direct (%)	τ direct (%)
OLMoE	0	1,987	9 / 9 (7.6 / 7.6)	1.09 / 1.05	−0.009 [−0.055, +0.038]	+0.024	7.2	7.5	+5.8	12.7
OLMoE	1	1,994	9 / 9 (7.7 / 7.5)	1.03 / 0.95	−0.005 [−0.051, +0.042]	−0.018	7.6	7.6	−4.7	11.3
OLMoE	2	1,996	27 / 9 (24.7 / 7.4)	0.99 / 0.99	−0.035 [−0.081, +0.012]	−0.029	2.9	2.9	−5.0	5.4
OLMoE	4	1,990	9 / 9 (7.3 / 7.3)	0.98 / 1.10	+0.011 [−0.036, +0.058]	+0.031	9.1	9.4	+6.0	11.4
OLMoE	6	2,007	26 / 9 (23.9 / 7.3)	1.00 / 0.84	−0.002 [−0.049, +0.044]	+0.005	5.9	6.0	+2.0	7.1

Model	Layer	Pairs	n_A / n_B: median (harmonic)	D_A / D_B	Spearman ρ [95%]	Pearson ρ	τ rank (%)	τ rank, n/D (%)	$\hat{\tau}$ direct (%)	τ direct (%)
OLMoE	10	1,984	25 / 8 (22.3 / 6.9)	0.99 / 1.09	-0.013 $[-0.060, +0.033]$	-0.010	5.3	5.4	-3.1	6.9
OLMoE	14	1,984	25 / 8 (23.0 / 7.0)	0.94 / 1.06	$+0.025$ $[-0.022, +0.071]$	$+0.078$	7.8	8.0	$+8.6$	14.4
granite	0	496	38 / 102 (34.9 / 99.0)	1.32 / 1.05	$+0.054$ $[-0.040, +0.146]$	$+0.037$	5.4	5.9	$+3.0$	6.0
granite	1	406	42 / 120 (38.2 / 110.2)	1.08 / 1.01	-0.001 $[-0.104, +0.102]$	$+0.007$	4.2	4.3	$+1.1$	5.2
granite	14	496	37 / 103.5 (33.7 / 96.0)	1.19 / 0.96	-0.025 $[-0.118, +0.068]$	$+0.094$	3.6	3.8	$+4.0$	7.5
granite	20	496	35.5 / 98 (33.4 / 95.8)	1.03 / 0.91	$+0.013$ $[-0.080, +0.107]$	$+0.033$	4.6	4.7	$+2.4$	5.4

Table A11. The fresh-document test by layer: rank columns from the fresh-document script, direct columns from the direct-covariance script (both named in the text). A is the original run's documents, B the later runs' documents outside A; pairs need at least two observations on each side. n: observations per pair on each side, median (harmonic mean). D: variance of the pair means over the prediction $\overline{s^2} \cdot 1/n$ for independent observations with one common variance; it absorbs σ_b^2 and unequal within-pair variances, so it is not a clustering effect (the document-clustered design effect of each pair mean is 0.99-1.01), and the n/D column divides each side's harmonic n by $\max(D, 1)$. Spearman intervals are Fisher-z with standard error $1.06/\sqrt{n-3}$. The Pearson correlation is shown without an interval, because the Fisher interval does not apply under these tails and the direct columns carry the variance-scale inference. The two document sets hold 136 / 136 documents on OLMoE and 136 / 380 on granite; `scripts/reliability_direct_covariance.py` writes these rows, with the document counts, to `results/reliability_table11.md`. τ is the between-pair standard deviation as a percentage of the within-pair one at the upper 95% end. For the rank columns it is the Spearman upper end inverted through the attenuation identity, conditional on the model in the text (one common σ_w); for the direct columns it is the covariance of the two sets' pair means with a stratified delete-one-document jackknife interval, over the root-mean-square σ_w. $\hat{\tau}$ is the direct point estimate $\mathrm{sign}(\hat{\sigma}_b^2)\sqrt{|\hat{\sigma}_b^2|}/\sigma_w$.

Does the candidate pool explain the low reliability? Γ_{ij} is the residual of an additive fit over all 28 coalitions at a site, hence a function of the whole pool, and the pool is never repeated: across the sites where a given pair is measured, the other six candidates overlap at a mean Jaccard of 0.06 with no exact repeats. Pool variation is therefore fully confounded with context in the reliability figure, and the objection that the low number is an artifact of redrawing the pool deserves a direct answer. We give two. Algebraically, the $m = 8$ residual is exactly the average of the $m = 5$ residuals over all twenty five-member sub-pools containing the pair, $\Gamma_{ij}^{(8)} = \frac{10}{7}\,\mathbb{E}_{\text{sub-pool}}[\Gamma_{ij}^{(5)}]$, so the estimator has already averaged over pool choice. Empirically, our three seeds redraw the random arm independently and revisit 108 sites at the same document, position and layer with identical fixed background and identical deterministic candidates, which isolates the pool component exactly. Over the 1,202 matched (site, pair) observations this yields, Γ_{ij} reproduces across independent pool draws at $r = 0.903/0.937/0.962/0.986$ at layers 2/6/10/14, with a pool-induced variance share of 2.1% / 3.0% / 2.0% / 0.4% against the full-sweep variance (9.7% / 6.7% / 4.0% / 1.4% against the replicate subsample's own smaller variance, the conservative reading). Substituting into $r = \sigma_b^2/(\sigma_b^2 + \sigma_w^2/n)$, removing pool variation entirely moves the layer-2 reliability from 0.020 to 0.020, and the conservative share moves it to 0.022. The candidate pool therefore does not explain why $\bar{\Gamma}_{ij}$ is unreliable. These records cannot separate context from conditioning set, since the protocol derives the six fixed experts from the token. The top-slot variant is informative here: it measures the same rank-8/9/10 pairs at the same 2,720 sites with two of the six fixed experts replaced, and the two measurements of the same pair at the same token correlate at $\rho = +0.004/-0.026/+0.033/+0.078$. A sweep that holds the token fixed while randomising which experts are held fixed separates the two under its own sampling rules (see the third point below), and we ran it (`scripts/conditioning_set_sweep.py`, 9,600 records at layers 2 and 6). At each site the token, the layer and the whole candidate pool are held fixed and all 28 coalitions are re-measured under several backgrounds: the natural one, three transplanted from other sites at the

same layer (so the conditioning set is randomised over backgrounds the protocol itself produces rather than uniform ones the model never sees), and two uniform redraws. The transplant library is the natural top six at 100 positions (the first 20 documents, all FineWeb-Edu, five positions each; 98 distinct sets at layer 2 and 100 at layer 6). Each draw makes up to 40 uniform attempts over the library and is accepted when it is disjoint from the pool, and no draw failed. Seventeen of the 4,800 transplant draws happen to equal the site's own natural set. Only expert identities are transplanted, and every coalition's weights are the router's probabilities at the current token. Intervals below are delete-one-document jackknife intervals, checked against the spread between disjoint document halves (`scripts/conditioning_set_analysis.py`).

We state the result at the strength the design supports. The design directly identifies sensitivity; it does not identify a variance share of the natural design. In the natural protocol the background is a deterministic function of the token, $F = f(x)$. Transplanting backgrounds changes the joint $p(x, F)$, so components estimated under the transplanted joint do not transfer to the natural one, and subtracting the two arms' variances would silently assign every context-background interaction and covariance term to the background. We therefore report three things, and the first needs no such assumption.

First, the background sets the estimand jointly with the token. At the same site and pair, Γ under the natural background predicts Γ under a transplanted one at $r = +0.031$ $[+0.006, +0.056]$ at layer 2 and $+0.003$ $[-0.028, +0.034]$ at layer 6 (Pearson over $\approx$67,000 site-pair cells per layer, excluding the 196 and 280 cells whose transplanted set is the site's own natural one and so reproduces Γ exactly; with them, layer 6 gives $+0.010$). Uniform redraws agree with the natural arm no better ($+0.034$ $[-0.006, +0.074]$ and -0.016 $[-0.054, +0.022]$). Because the comparison is scale-free, it does not depend on how the arms' spreads compare. It is also a direct observation under the transplant rule and needs no independence of background and token (see the third point): weighting each site by its acceptance probability gives $+0.031$ and $+0.002$, and within every tercile of that probability the point estimate stays below 0.05 in absolute value.

Second, a spread ratio mostly reflects scale. Setting the standard deviation with which swapping the background moves Γ at layer 2 (2.55×10^{-2} nats) against the entire within-pair spread of the natural design (1.82×10^{-2}) gives a ratio of $1.40\times$ $[1.10, 1.71]$, which is not evidence that the background moves the estimand more than everything the natural design varies. The ratio factors into a difference of scale between arms and a within-arm term. The within-pair spread of Γ under transplanted backgrounds is $1.44\times$ $[1.12, 1.77]$ that under the natural one, and inside the transplant arm, changing the background at a fixed token moves Γ $0.975\times$ $[0.945, 1.005]$ as much as changing token and background together does. At layer 6 neither departs from one (ratio 0.997 $[0.72, 1.27]$, scale 1.05 $[0.76, 1.34]$). A transplanted set is not the token's own top six, so it plausibly carries different gate mass. The records store no router probabilities, so for layer 2 this remains a hypothesis rather than a measurement, and we do not quote the ratio as a finding.

Third, what the within-arm share measures. The library is fixed before the sweep, but a draw is kept only if it is disjoint from the site's pool, so the accepted background follows $q(F \mid x) \propto q_{\text{lib}}(F)\,\mathbf{1}\{F \cap P(x) = \varnothing\}$, with q_{lib} uniform over the library's entries, and still depends on the token through its pool. A site admits 25 to 69 of the library's 100 entries at layer 2 and 11 to 73 at layer 6 (mean 47 at both), and the token's two slot occupants and ranks 9–11, five of the pool's eight members, alone explain half the variance of that acceptance rate (the three random members a third, the token's domain 1–3%). Uniform redraws also depend on the token, though only through their support: they are six of the 56 experts outside the pool, with every such set equally likely at every site. Everything below is therefore a sensitivity under these sampling rules and is not a property of backgrounds drawn independently of the token (`scripts/conditioning_set_support.py`). Within the transplant arm, re-drawing the background at a fixed token accounts for 92.9% $[87.1, 98.8]$ of the arm's within-pair variance at layer 2 and 91.9% $[87.7, 96.0]$ at layer 6 (98.6% and 95.2% under uniform redraws, whose Wald intervals reach past 100%). This within-token share is a law-of-total-variance split of the arm's realized joint law of token and background, so it needs no independence. The estimator is $\hat\sigma^2_{F|x}/(\hat\sigma^2_{F|x} + \hat\sigma^2_x)$, with $\hat\sigma^2_{F|x}$ the mean over (pair, site) cells of the variance across the site's three backgrounds and $\hat\sigma^2_x$ the mean over pairs of the variance of site means less $\hat\sigma^2_{F|x}/3$, clamped at zero. Under a crossed model $\Gamma = \mu + g_x + g_F + g_{xF}$ with the background independent of the token, it would estimate $\mathrm{Var}(g_F) + \mathrm{Var}(g_{xF})$ rather than a background main effect. Separating the two is exploratory. Library backgrounds recur, so the same pair meets the

same background at different sites. Entering the half mean-squared differences of such couples (15,389 and 14,903), of couples that share the site and of couples that share nothing into balanced crossed-design formulas gives the interaction 104% $[76, 131]$ and 92% $[78, 106]$ of the variance, the background main effect -12% $[-39, +15]$ and $+7\%$ $[-8, +21]$, and context alone $+8\%$ $[-5, +22]$ and $+1\%$ $[-6, +8]$. Those formulas assume that token and background are crossed independently and that the three kinds of couple differ only in what they share. The acceptance rule grants neither, since a background is observed only at sites whose pool it avoids and only 32% of the all-different couples could have exchanged backgrounds, and a functional ANOVA with dependent inputs needs its own identification (Hooker, 2007). The direction survives each correction we tried. Weighting each site by its acceptance probability, keeping only couples that could have exchanged backgrounds, both together, dropping draws of a site's own set, or giving every background equal weight puts the interaction at 104–109% and 84–99% and the background main effect between -14% and $+11\%$. The one subset that is fully crossed, balanced and in support by construction (362 and 302 cases in which two sites in different documents drew the same two backgrounds for the same pair) gives interaction 84% $[52, 115]$ and 84% $[44, 124]$ and background main effect $+6\%$ $[-5, +18]$ and $+6\%$ $[-1, +13]$. The exact shares are therefore model-dependent, and a background main effect of a fifth of the variance is not excluded. The data show no detectable fixed effect of a background on a pair, and in every variant most of what a background does to a pair changes from token to token. Context, the token-level mean over the arm's backgrounds, is small by the single-arm estimator (standard deviation 7.0 and 4.9×10^{-3} nats, a quarter to a third of the background term), though at layer 6 its interval excludes zero.

None of this rescues the pair. At matched per-pair sampling, with eight observations from eight distinct sites over the same 1,506 and 1,505 pairs in every arm, the split-half reliability of $\bar{\Gamma}_{ij}$ is -0.015 / $+0.004$ / $+0.021$ at layer 2 and $+0.042$ / $+0.002$ / -0.011 at layer 6 for the natural, transplant and uniform arms, every jackknife interval within ± 0.06 of its estimate. Drawing the eight observations without the distinct-site rule lets a transplant pair's halves share sites and inflates its figure to $+0.046$ / $+0.051$. None of the three background rules raises the pair mean's reliability at matched sampling.

The gate weighting does not rescue it either. Under the model's own weights a random candidate executes with a median gate weight of 0.0098 against 0.0219 at ranks 9–11 (below), so the random arm is de-censored in which expert is drawn but not in the weight it runs with, and a pair met with different provenance at different sites runs with different weights at each. The uniform-gate control (App. D) re-measures the first seed's 2,720 OLMoE sites (same documents, positions, fixed experts and candidates) with every member at $1/K$, which removes both (`scripts/uniform_csa.py`). Co-activation and $\bar{\Gamma}_{ij}$ stay unrelated, with $\rho = -0.021/-0.013/-0.007/-0.001$ on the full pool at layers 2/6/10/14 and $|\rho| < 0.02$ on the random arm; the smallest p of the eight is 0.38 (BH $q = 0.98$), equivalence against $|\rho| = 0.1$ holds at all eight after BH, and disjoint document halves put the empirical standard error at 0.76–$0.91\times$ the pair-level one. The pair mean is not more reliable either. At the same median of nine observations per pair, its split-half reliability is $-0.003/+0.040/-0.026/+0.005$ under uniform weights against $+0.001/+0.011/-0.016/+0.021$ under the model's own, and every full-pool paired difference's delete-one-document jackknife interval covers zero, and still does when narrowed to the spread between disjoint document halves. Among pairs of the router's own five candidates (the two slot occupants and ranks 9–11, median weights 0.027 and 0.022), uniform weights give $+0.024/+0.039/-0.108/+0.020$ against $+0.107/+0.016/+0.038/+0.118$ at five to seven observations per pair; the two natural values near 0.11 shrink on the three-seed sweep's same support (0.015 and 0.071 at ten) and do not carry to the 136 new documents ($\rho = -0.033/+0.025/+0.014/+0.016$, every interval covering zero). Provenance does leave a trace under the model's weights. At every layer the share of within-pair variance lying between a pair's provenance classes lies 2–4 points above a within-pair permutation null (permutation $p = 0.03$–0.13, below 0.05 at layers 6 and 14 only), and the root-mean-square Γ of two slot occupants is 1.5–$2.1\times$ that of any other provenance at layer 10 and 1.7–$3.6\times$ at layer 14 (at most $1.3\times$ at layers 2 and 6). Uniform weights remove both (-1.0 to $+1.5$ points; at most $1.3\times$), and standardizing Γ within provenance class leaves the reliability at $-0.005/+0.008/-0.024/+0.016$. The weight mix adds a few percent to the within-pair variance but is not what makes $\bar{\Gamma}_{ij}$ unreliable. Like granite's, the eight CSA tests here are a weighting control and are not in the 44-test family.

The site's scale does not explain it either. If $\Gamma_{ij}(x) = s(x)\,\gamma_{ij}(x)$ with a heavy-tailed site scale s, a stable pair component of γ could be buried in $\bar{\Gamma}_{ij}$. The scale is real (the site RMS of Γ has a

coefficient of variation of 1.9–3.4 over the eleven fresh-document layers), so we re-ran the fresh-document estimators and the 44-test CSA family on Γ divided by its site RMS and on its within-site rank (`scripts/gamma_scale_reliability.py`). Both remove the tail (pair-mean excess kurtosis at most 2.2, against 2–190) and neither raises the bound. The direct upper ends (not Gaussian-model figures) are 12.2% and 5.4% (OLMoE, granite) standardized and 12.0% and 4.8% for ranks, against 14.4% and 7.5%, every covariance interval covering zero. The cross-document Spearman stays within $|\rho| \leq 0.092$; two OLMoE layers are nominally positive ($+0.059$ and $+0.060$ at layers 0 and 14, standardized), and neither survives BH over the eleven layers ($q = 0.069$). The split-half estimator above gives at most 0.045 (standardized) and 0.065 (ranks) on OLMoE, against co-activation's matched 0.32–0.53, and no CSA test survives BH on either version, with the pair-level Fisher p or an expert-relabelling permutation p.

Interaction magnitude behaves differently. $|\Gamma|$ over its site RMS reproduces across documents, with direct $\hat{\tau}$ up to 18.2% on OLMoE and 17.5% on granite and intervals excluding zero at 4 of 7 and 4 of 4 layers (still at 2 of 7 and 3 of 4 with additive expert effects removed). It also tracks co-activation, with ρ up to $+0.449$, with 34 of 44 tests surviving BH (30 of them positive; 29 with the permutation p), against 17 of 44 for raw $|\Gamma|$ and for Γ^2. The link runs mostly through the executed gate weights. On the same 2,720 first-seed sites under uniform gate weights (`scripts/gamma_scale_controls.py`), OLMoE's full-pool correlation falls from $+0.139/+0.141/+0.308/+0.325$ to $-0.004/+0.048/+0.032/+0.071$ at layers 2/6/10/14 (BH over the eight tests: from 7 to 0 with the permutation p, from 7 to 1 with the Fisher p), and granite's two positive ones, $+0.304$ and $+0.242$ at layers 2 and 14, fall to -0.008 and $+0.038$. Co-activation rank-correlates with a pair's mean $\log w_i w_j$ at $+0.26$ to $+0.72$, and regressing the standardized magnitude on log gate weights and provenance leaves 1 of 44 tests with the permutation p (5 with the Fisher p, three of them negative) and pair-specific reproducibility at 1 of 11 layers. Little survives uniform weights: raw magnitude at OLMoE layers 10 and 14 ($\rho = +0.07$ to $+0.12$), and granite layer 20's negative association, which is expert-level. Under uniform weights granite's raw $|\Gamma|$ keeps some pair-level reproducibility (split-half up to 0.18 with expert effects removed), which co-activation does not track (0 of 8). Whatever co-activation predicts about interaction magnitude therefore comes mostly from the router's own weighting.

The inference is therefore narrow. Interaction is a property of a (pair, context, conditioning set) combination, the estimand is set by token and background jointly, and none of the three background rules tested, nor uniform gate weighting, makes the signed pair mean reproducible. No arm holds one background fixed across many contexts, so a modest reliability under a fixed background is not excluded, though pairs that meet the same background at two sites agree no detectably better than pairs that do not. The inference is corroborated independently in App. G, where shrunken per-pair means transfer across documents at $R^2 \approx 0$.

For the literature, the consequence is stronger than a null correlation and lies elsewhere. Co-activation is a perfectly stable pair-level statistic, since it is a count over sites and more data makes it more precise; interventional synergy is not. Expert-interaction graphs built from co-activation are therefore limited by the quantity their edges carry rather than by the estimator. Per-site Γ reproduces across independent candidate pools at $r \geq 0.90$, but a pair's mean barely reproduces across documents at any affordable sampling. Figure 2 of the main text shows the same fact geometrically: the co-activation axis is well resolved while the synergy axis is dominated by within-pair variation, and no structure survives in the plane.

Layer	ρ full	95% CI	ρ random	TOST p	reliability
0	-0.009	$-.056, +.038$	-0.027	6.5e$-$5	0.014
1	$+0.003$	$-.043, +.050$	$+0.003$	2.4e$-$5	< 0
2	-0.014	$-.060, +.033$	-0.002	1.2e$-$4	0.020
4	$+0.014$	$-.032, +.061$	$+0.037$	1.7e$-$4	0.027
6	$+0.022$	$-.024, +.069$	$+0.020$	4.9e$-$4	0.014
10	$+0.002$	$-.044, +.049$	-0.012	1.7e$-$5	0.007
14	-0.003	$-.049, +.044$	$+0.024$	1.8e$-$5	< 0

Table A6. Co-activation-synergy alignment per layer (OLMoE), on the full candidate pool and restricted to pairs containing at least one uniformly random candidate (the de-censored support the

random arm makes measurable); layers 0, 1 and 4 in bold are the later shallow extension, one seed. Intervals and TOST use the Fisher-z transform with the pair-level standard error $1.06/\sqrt{n-3}$, which §5.2 checks against independent document halves (half-sample 95% intervals agree in 96.9% of splits, 93.2% at worst). Equivalence against a margin of $|\rho| = 0.1$ is a two-one-sided test on that scale; every layer establishes it. The last column is the Spearman-Brown-corrected split-half reliability of the per-pair mean synergy, which bounds any correlation that quantity could exhibit with anything. Pair counts are 1,965-2,016 per layer on the full pool and 1,842-2,016 on the random arm; the interval shown is for the full-pool correlation.

Below the Top-K boundary, router score order buys no reliable mean utility. This is a unary result, separate from the decoupling above, and we do not claim it as a contribution. Fitted unary quality by candidate provenance is $+0.0003$ for slot occupants, -0.0000 for next-by-rank and -0.0002 for uniformly random candidates ($n = 16{,}320/24{,}480/24{,}480$), with a rank-minus-random difference of $+0.00014$, 95% CI $[-0.0002, +0.0005]$, which includes zero. Broken out by individual rank rather than pooled, the difference is $+0.00032$ at rank 9, $+0.00004$ at rank 10 and $+0.00006$ at rank 11, and every document-clustered interval contains zero (rank 9: $[-0.00030, +0.00096]$). Because the rank-9 point estimate is $2.3\times$ the pooled one, we state the claim at the rank where it is weakest as well as on average. An expert the router ranks 9th is not measurably a better substitute than one drawn uniformly at random from the 53 outside the top eleven. The ranking carries information up to the Top-K boundary and no resolvable mean benefit below it. The result bears only on which expert refills a vacated marginal slot. Under the model's own gate weights the next-ranked expert is, on average, no better a refill than a random one; the comparison says nothing about refilling against dropping the token, about pruning experts globally, or about re-ranking within the Top-K. The one exception, a win-rate edge at a single layer, is described below. On granite the same contrast is -0.00066 (95% CI $[-0.00158, +0.00020]$), also including zero but with the opposite sign; its uniform-gate control gives $+0.00016$.

The router's marginal-slot pair also coincides with the counterfactual-best pair at 7.8% of sites against a uniform rate of 3.6%; among the ten coalitions drawn only from its own top eleven (the two slot occupants and ranks 9–11), at 11.7% against 10.0%; and among the seven of those that keep at least one slot occupant, at 12.2% against 14.3%. These argmax rates do not measure preference. The router's pair is the only coalition that swaps nothing, and the others differ in how far they can move. A random candidate carries a median gate weight of 0.0098 against 0.0219 at ranks 9–11, so under match-natural weights the three coalitions that pair a random candidate with the same partner are near-duplicates (median within-site spread 0.0017 against 0.0045 nats) and rarely hold the maximum; per coalition, slot-plus-random pairs are best at 2.6% of sites against 6.0% for slot-plus-rank pairs. In head-to-head comparisons that redundancy does not matter. The router's pair beats any of its 27 alternatives in 51.6% of comparisons (document-clustered 95% CI $[50.8, 52.5]$; 51.7% against a slot-plus-random coalition, 51.0% against a slot-plus-rank one), and a random substitute is worse than a rank 9–11 substitute for the same partner in 51.1% $[50.5, 51.6]$. That edge is neither spread over depth nor shared across models (`scripts/rank_vs_random.py`): it is 49.1%, 49.5% and 50.0% at OLMoE layers 2, 6 and 10 against 55.7% $[54.5, 57.0]$ at layer 14, and 49.9% $[49.1, 50.7]$ pooled on granite and 49.5% $[49.1, 50.0]$ on Qwen. That layer's edge recurs on 136 held-out documents (55.7% $[53.7, 57.8]$) and falls to 52.3% $[50.3, 54.2]$ under uniform gates. At layer 14 the raw mean advantage, which a few large-margin sites dominate, does not reproduce on the held-out documents ($+0.00020$, $[-0.00080, +0.00115]$). The converse case occurs once. On granite's 516-document replication run, layer 20 shows a raw mean advantage ($+0.00142$, $[+0.00060, +0.00235]$) with no win-rate edge (50.6%), a 5%-trimmed mean that includes zero ($+0.00041$, $[-0.00009, +0.00093]$), and the opposite sign on the original 680 sites (-0.00045). Below the boundary, rank order buys a win-rate edge at one layer of one model and occasional tail-driven mean gains, but no mean benefit that holds across samples.

F Oracle gain, propagation and recoverability: full results

At the intervened token the router is beatable in hindsight by a wide margin, but the margin is concentrated in a minority of tokens.

Intervention	mean	median	p90	p99	sites > 0.05 nats
reselect, marginal slots	$+0.0387$	$+0.0109$	$+0.104$	$+0.338$	24.3%
reselect, top slots	$+0.0898$	$+0.0157$	$+0.253$	$+0.856$	36.4%
reweight, set fixed	$+0.0871$	$+0.0195$	$+0.246$	$+0.848$	35.3%

Table A7. Oracle gain at the intervened token, defined as the best of the measured equal-compute alternatives minus the router's own choice, on single-token utility. 2,720 sites per row. These are oracle quantities, since choosing the maximiser requires the label.

The mean is 3.6 to $5.7\times$ the median. On a typical token the router is close to optimal among the alternatives we measure (about 0.011 nats from the best of 28), and the aggregate is carried by a fragile minority, consistent with the fragile-token headroom reported by Yoon et al. (2026). Any summary that quotes only the mean overstates how much of the corpus is affected.

In the mean, the gain is confined to the token intervened on. Below, the coalition that maximises token t is selected and then scored on the eight tokens that follow.

Offset	+0	+1	+2	+3	+4	+5	+6	+7	+8
reselect, marginal slots	$+0.0390$	$+0.0009$	$+0.0003$	-0.0007	-0.0010	-0.0001	$+0.0001$	-0.0004	$+0.0000$
reselect, top slots	$+0.0905$	-0.0011	-0.0012	-0.0010	-0.0020	$+0.0002$	$+0.0008$	-0.0002	$+0.0002$
reweight, set fixed	$+0.0874$	-0.0013	$+0.0003$	$+0.0007$	$+0.0009$	-0.0001	-0.0006	-0.0009	$+0.0001$

Table A8. Mean gain by distance from the intervened token, $w = 9$. Each arm sweeps 2,720 sites; rows use the subset with a full nine-token window, which drops 32 sites for the two reselection arms and 48 for the reweighting arm, all at document ends. The largest off-token entry, -0.0020 (top slots, $+4$), is -0.00199 unrounded (-0.00200 if the sites with a partial window are kept at the offsets they reach).

Averaged over $t{+}1 \ldots t{+}8$ the three interventions give -0.00009, -0.00055 and -0.00012 nats. This holds layer by layer, so it is not an artifact of deep layers having little network left to propagate through; at layer 2, with thirteen layers downstream, the marginal-slot effect is -0.00001 and the top-slot effect -0.00137.

The current-token gain is not paid for later. One intervention's cumulative after-effect over the eight following positions is -0.00073 nats (marginal slots; -0.00440 at the top slots) against $+0.039$ at the token itself. This is a statement about one intervention, and we do not extend it to a sequence. Every site here is measured against a clean prefix, one intervention per forward pass, so whether many simultaneous interventions add is untested, and App. D is this paper's own demonstration that a system can be sharply non-additive in the dimension one assumes away. Three gaps matter in particular. First, the after-effects are near zero in the mean but not per site; the per-site sum over $t{+}1 \ldots t{+}8$ has a standard deviation of 0.054 nats and a mean absolute value of 0.022, over half the same-token gain. Second, nothing beyond offset $+8$ is measured, whereas under joint intervention a token inherits a contribution from every earlier intervention rather than from the previous eight. Third, the coalition chosen for t is the argmax under the clean prefix, not under one perturbed at every prior position. A sequence-level number would require a joint-intervention run, which we did not do. Counted by sign, wins and losses on $t{+}1 \ldots t{+}8$ are within three points of even once exact ties are set aside. Where the single-token selection returns the router's own choice (its pair, or $\tau = 1$), both sides of the comparison are the same forward pass and the difference is exactly zero. Wins / ties / losses are 49.1 / 7.2 / 43.7% at the marginal slots, 38.5 / 22.7 / 38.8% at the top slots and 47.0 / 5.4 / 47.6% for

reweighting, so a win share should be read against half the untied sites (46.4%, 38.7% and 47.3%). Over untied sites the chosen intervention wins at 52.9% (document-clustered 95% CI $[51.0\%, 54.9\%]$, document sign-flip $p = 0.005$), 49.9% ($[47.6\%, 52.1\%]$, $p = 0.92$) and 49.7% ($[47.7\%, 51.7\%]$, $p = 0.80$), and the untied means, -0.00010, -0.00071 and -0.00013 nats, each have an interval containing zero. The marginal-slot excess comes from small wins, since the average win is smaller than the average loss ($+0.0027$ against -0.0032 nats), and where the difference exceeds 10^{-3} nats the win rate is 51.0%. By layer the top-slot win rate runs from 44.1% at layer 2 ($[40.0\%, 48.2\%]$; Holm-adjusted $p = 0.03$ over the arm's four layers, 0.08 over all twelve layer-by-arm cells) to 54.0% at layer 10, and the marginal slots reach 56.0% at layer 14 (adjusted $p = 0.06$ over the twelve). The layer-2 top-slot mean, -0.00137 nats, has an interval containing zero ($[-0.0033, +0.0005]$), so we read no downstream cost into it (`scripts/transfer_sign_test.py`).

Because the oracle above requires the label, we also tested directly whether a predictor without it can recover any of the weighting gain, which is the estimand the rest of this paper only approaches indirectly. On the same 2,720 sites we trained a per-token head to choose τ from features a deployed router already has (the gate weights over the routed set, their entropy, the total router mass, the layer) and, in a second variant, those plus the top 64 principal components of the layer-input representation. Evaluation is five-fold by document, at the intervened token, and every gain in Table A9 is over the natural weighting ($\tau = 1$). The relevant baseline is the best single τ chosen on the training documents, and the two paired rows score each head against it on the same held-out sites (`scripts/weighting_head_paired.py`). If one global temperature captures the gain, there is nothing per-token to learn.

	mean gain (nats)	95% CI	positive at
natural weighting ($\tau = 1$)	0	—	—
best fixed τ (chosen on train)	-0.00071	$[-0.00189, +0.00021]$	9.5%
best fixed τ per layer	-0.00217	$[-0.00499, +0.00018]$	30.0%
learned head, router-visible features	$+0.00007$	$[-0.00287, +0.00332]$	41.7%
learned head, + residual stream (transductive PCA, comparison only)	-0.00325	$[-0.00753, +0.00136]$	42.9%
learned head, + residual stream (PCA fitted in-fold)	-0.00078	$[-0.00476, +0.00326]$	42.0%
router-visible head − best fixed τ (paired)	$+0.00078$	$[-0.00216, +0.00392]$	42.2%
in-fold head − best fixed τ (paired)	-0.00007	$[-0.00397, +0.00408]$	42.3%
planted-signal control, $\alpha = 1.0$	$+0.03025$	—	—
planted-signal control, $\alpha = 0.5$	$+0.00610$	—	—
permutation null (target shuffled within layer)	-0.00843	$[-0.01372, -0.00315]$	—
leaky control (features contain the label)	$+0.08646$	$[+0.07657, +0.09756]$	87.9%
oracle (uses the label)	$+0.08707$	$[+0.07718, +0.09821]$	94.6%

Table A9. Recoverability of the aggregation-side oracle gain, five-fold by document, scored at the intervened token. Every row is a gain over $\tau = 1$ except the two paired rows (head minus best fixed τ, same held-out sites; "positive at" is the share of sites where the head wins). CIs are document-clustered bootstraps (2,000 resamples of 136 documents, in fold order), except the permutation null's, which is the 2.5-97.5th percentile of its 200 shuffles. Every interval above holds the five fitted heads fixed. Refitting the whole pipeline inside disjoint document halves puts the variances of the two paired contrasts at $1.47\times$ and $1.32\times$ their cluster-robust values (100 splits, themselves imprecise; $1.42\times$ and $1.33\times$ over 2,000 splits, `scripts/weighting_head_refit.py`). Taking the larger of each contrast's ratios, calibrated standard errors are 21% and 15% wider, and widening

the cluster-robust intervals ($[-0.00234, +0.00390]$ and $[-0.00411, +0.00398]$) accordingly gives $[-0.00300, +0.00456]$ and $[-0.00473, +0.00460]$, which changes neither conclusion. "Positive at" on the paired rows is the win share; ties are 14.9% and 13.3%. On the other rows it counts strict wins over $\tau = 1$, which tie exactly wherever $\tau = 1$ is chosen, so it should not be read against 50% (the best fixed τ is 1.0 in four of five folds). Planted-signal rows express the recovered gain as nats; as a fraction of their own oracle they are $+34.7\%$ and $+6.8\%$, one planting draw each ($+35.8\%$ and $+3.3\%$ averaged over 200 draws; App. J). The permutation null is the mean over 200 within-layer target shuffles.

Neither head is distinguishable from zero, and neither improves on the best fixed τ by more than sampling noise (paired: $+0.9\%$ and -0.1% of the $+0.08778$-nat headroom over that fixed τ). The fold partition used here is if anything favourable to the heads: over 20 random partitions the paired means average -0.00040 and -0.00107 (ranges $[-0.00220, +0.00096]$ and $[-0.00284, +0.00066]$). The best fixed τ is itself 1.0 in four of five folds, so the model's own temperature is already the best global setting and the gap is not a calibration error either. The head recovers -0.9% of the oracle gain, 95% CI $[-6.0\%, +4.3\%]$ once refitting varies ($[-5.5\%, +3.7\%]$ holding the five fitted heads fixed). Refitting the whole pipeline inside disjoint document halves puts the variance of the head's gain over $\tau = 1$ at $1.227\times$ its cluster-robust value over 2,000 splits and $1.253\times$ over 1,000 further splits with independent seeds, and the full-sample fold-partition spread alone implies $1.252\times$; the interval uses the largest (`scripts/weighting_head_refit.py`). The ratio grows with sample size ($1.06\times$ at a quarter of the documents), so the half-sample figure may understate full-sample refitting; adding the partition variance on top of it gives $[-6.5\%, +4.7\%]$. Averaged over 20 fold partitions the same head recovers -2.2% (range -4.2% to -0.3%); the partition used here is the fourth most favourable of the 20.

Where the PCA is fitted matters, because it moves the headline number. An SVD of the residual-stream features fitted on every site before the fold split is not label leakage, but it is not inductive cross-validation either, since each training fold then sees the test folds' feature distribution. This transductive fit gives -3.7% of the oracle, against -0.9% with the PCA fitted inside each fold. The conclusion is the same and the in-fold interval is narrower, so we report the in-fold number and show the transductive one in Table A9 for comparison (App. L).

Three checks make that null interpretable, and the first of them is needed because the obvious positive control is too weak. A control that appends the ten true gains to the feature vector recovers 99.3% of the oracle, but it regresses D on a design matrix containing D and so only verifies that ridge-plus-argmax can read a column it was handed. It licenses "the pipeline is not the limitation" and nothing more. We therefore add two controls that never see the label. The first is a permutation null. Shuffling the target across sites within layer keeps its exact marginal distribution and heavy tail while unlinking it from the features, and gives -9.7% of the oracle (sd 3.2, 200 draws). The head's -0.9% sits above that null at one-sided $p = 0.015$ (0.004 over the 1,000 draws of the selective-head null below), so the head does something better than chance and still does not beat a single global temperature. The second is a planted-signal control. We row-permute the target so that the best τ becomes a monotone function of a residual-stream direction the head can see; labels never enter the features, and the signal is planted within layer so the layer indicator offers no shortcut. The identical pipeline then recovers $+34.7\%$ of the oracle at $\alpha = 1$ ($+35.8\%$ averaged over 200 independent plantings, 95% of them within $[+30.0\%, +41.0\%]$; at $\alpha = 0.7$ and 0.5 the averages are $+14.2\%$ and $+3.3\%$, and the paired interval, refits included, excludes zero in 87% and 14.5% of plantings; App. J). This feature set, capacity and training scale can therefore turn a strongly structured learnable signal of the same size into a third of the available gain, though not reliably one planted at only half the sites, and they recover essentially none of the real one. Second, the head has not simply learned to do nothing. It uses all ten temperatures, with the in-fold head choosing $\tau \to \infty$ at 26.8% of sites, $\tau = 0.2$ at 12.2%, and the natural $\tau = 1$ at only 13.5% (`scripts/weighting_head_paired.py`). Its choices are varied and confident, and they are no better than the model's own.

Allowed to abstain, the head does no better, and the features do not predict the direction of the best deviation. Because the head never abstains, one may ask whether the forced argmax spends a usable signal on bad deviations. We tested this after the result above was known, with a selective head (`scripts/selective_weighting_head.py`, `scripts/selective_head_followup.py`) that uses the same folds, features and ridge but keeps $\tau = 1$ unless the predicted gain of its argmax over $\tau = 1$ exceeds a threshold δ. The threshold is chosen by nested five-fold cross-validation over the training documents from a grid that includes never deviating ($\delta = 0$ is the head

above). The chosen δ runs from 0.006 to 0.20 nats, with never deviating in one router-visible fold. The selective heads keep $\tau = 1$ at 86.8% and 93.2% of sites (router-visible, in-fold; the oracle deviates from $\tau = 1$ at 94.6%) and recover -0.04% and -2.3% of the oracle gain, 95% CIs $[-2.2\%, +2.2\%]$ and $[-5.1\%, +0.5\%]$ including refits (conditional variances inflated $2.49\times$ and $1.54\times$, the larger of the fold-partition and document-half refit ratios); over 20 fold partitions they average -0.45% and -0.61%. Paired against the best fixed τ they differ by $+0.00067$ $[-0.00168, +0.00303]$ and -0.00130 $[-0.00412, +0.00153]$ nats, with intervals widened the same way. Neither is distinguishable from zero or from the best fixed τ, and neither beats its own permutation null (null means -0.75% and -0.45%; one-sided $p = 0.27$ and 0.96, 1,000 draws). The argmax head's margin over its null (-0.9% against -9.7%) thus shows that the features carry some information about the gains; it does not show that they hold a gain the forced argmax wastes. A head fitted to shuffled targets pays for deviating on noise, and once both may abstain the margin disappears. The predicted gain also fails to rank the head's own choices. Where the argmax head deviates, its Spearman correlation with the realized gain is $+0.003$ and -0.052, and the in-fold head's most confident decile loses 0.019 nats, although perfect abstention over the same picks would reach 20% and 23% of the oracle. The features cannot tell sharpening from flattening either. Of the 2,571 sites whose best τ lies strictly on one side of 1, 55.5% are best below it; an in-fold ridge on the same features separates the two at AUC 0.529 and 0.524 (95% CIs $[0.500, 0.558]$ and $[0.494, 0.553]$ including refits), against 0.517 from the layer indicators alone and 0.520 and 0.518 under a within-layer label permutation ($p = 0.21$ and 0.32, 20,000 draws); within layer the AUCs are 0.487 and 0.490, against null means of 0.477 and 0.484. Knowing exactly where to deviate and in which direction, with one fixed τ per side (0.2 and $\tau \to \infty$), would be worth 73% of the oracle.

This is a bounded negative result, and its bounds are specific. It covers a linear head on ≈2,200 training sites, a one-parameter temperature family, and a single-token log-likelihood objective. Li (2026) reports a gain on this model family from a 301K-parameter head trained against sequence-level cross-entropy over a richer weight space, using the already-computed expert outputs and expert identities, whereas our head sees only the layer-input residual stream. Nothing here contradicts that result. The contrast localises what would have to be doing the work, namely input information first and then capacity, objective and supervision scale, rather than the existence of a gap.

The transfer measurement has limits of its own. Tables A7 and A8 measure where an intervention's effect lands rather than whether it can be predicted; the recoverability question is the one Table A9 answers. Utilities here are deterministic, so the oracle is not inflated by measurement noise. It is still a hindsight bound, conditioned on the realized next token that no router sees, and there is no second independent draw of the utility at position t against which it could be validated. A disjoint window measures transfer to other positions, which is a different quantity. §6 asks the same question of expert selection, namely whether a predictor that never sees the label can choose better than the router at the same position; App. G answers negatively offline and App. H negatively online, both at our scale and with the qualifications stated there. Contemporaneous work reaches the opposite answer for a related quantity: the aggregation head of Li (2026) lowers cross-entropy on this model family, so nothing here should be read as showing that gains of this kind cannot be learned.

G Offline predictability of the interaction residual

Can the measured interaction be predicted? Prediction is the prerequisite for exploiting it online. We train context-conditioned bilinear heads $\widehat{W}_x(i,j) = q_i^\top \mathrm{Diag}(c(x))\, q_j$ (with a joint unary term; rank 16; Huber loss, weight decay, early stopping) on the formal sweep data, under document-level train/val/test splits, to predict (a) per-site residual interaction Γ_{ij} and (b) site-centered utility ΔU_{ij}, from three feature sets: the layer-input representation h (re-extracted for 8,159/8,160 sites by deterministic corpus replay, with router-probability verification), h concatenated with the router's full probability vector, and the router probabilities alone.

The result is a consistent null. Held-out $R^2 \leq 0$ holds for every layer, target, and feature set (down to -0.28 for the h-only variant), within-site rank correlation is $\leq +0.04$, selection by predicted scores underperforms the router's own pair (the selected-coalition regret, i.e. the utility of the head's top-ranked pair minus the router's, is negative), and the oracle hit rate is at chance. Shrunken per-pair means also transfer at $R^2 \approx 0$ across documents, so the pair-identity component of interaction is unstable across contexts, consistent with §5.2. A ridge probe from the identical features does predict

a site-level scalar, router entropy from h (held-out R^2 $+0.46$ to $+0.60$ across layers; `scripts/probe_features.py`), so the features are not uninformative. As the next paragraph explains, however, this is no positive control for an expert-indexed target.

The null is not specific to interaction, and it is weaker than a boundary. Router entropy is a site-level scalar and a far easier target than anything indexed by expert identity, and its success does not show that an expert-indexed quantity is learnable here at all. Predicting the fitted unary effect a_i from the same features under the same split gives held-out R^2 of 0.000, 0.001, -0.032, -0.292 at layers 2/6/10/14, which is no better.

Because that result reflects on the pipeline as much as on the target, we ran three trivial baselines for a_i on the same splits. At the marginal slots all three sit at zero (provenance indicators ≤ 0.0004, per-expert train mean < 0, router probability ≤ 0.0003), so neither a static per-expert mean nor these two features carries transferable signal there. The top slots differ. Router probability alone reaches $R^2 = 0.020/0.000/0.028/0.033$ at layers 2/6/10/14 and provenance indicators reach $0.013/0.000/0.007/0.014$. These values are small but positive, and above what our rank-16 bilinear heads achieved on the same data; a two-feature linear model that beats the head is evidence about the head.

We therefore state App. G at the scope it supports. With the features tested (the layer input, the router probabilities, and both), the predictors tested (rank-16 context-conditioned bilinear heads for Γ_{ij} and ΔU_{ij}; a ridge probe from h and three trivial baselines for a_i) and $\approx$1.3k training sites per layer, no expert-indexed target reaches a held-out R^2 above 0.001 at the marginal slots; at the top slots a few percent of unary variance is linearly recoverable, and our heads did not capture it. This is a transfer null for these predictors at this scale. It is neither a property of the target nor a predictability boundary, since a context-conditioned component can average to zero over every expert or pair and still be readable from x (App. H). Only a narrower claim survives, namely that the interaction residual transfers at $R^2 \approx 0$ across documents under every predictor tested, even where a unary term is partly recoverable. The null also persists under the top-slot protocol (perturbing the two strongest routed slots; 2,720 sites with $3.15\times$ the utility spread), with held-out $R^2 \leq 0$ and selection below the router baseline for every layer and feature set.

H Online counterfactual distillation

Unary counterfactual supervision is not exploitable at this scale

What the teacher supervises. The experiment below distills from a teacher that performs equal-compute single-expert substitutions at sampled sites. It measures how good each candidate would be in the slot, and it does not measure how well a pair of candidates works together. Its score for candidate i is therefore $U(F \cup \{i, j_0\}) = \text{const} + \alpha_i + \gamma_{ij_0}$, a conditional marginal that contains one row of the interaction matrix. This is not exactly the fitted a_i of App. A, though the two agree closely (median Spearman 0.964 over the primary sweep). By construction the teacher cannot deliver pair structure between two slots varied at the same time, so the Γ_{ij} of App. G lies outside its signal. We chose this channel deliberately for three reasons. It carries 89.5% of the single-token oracle gap (one minus the pooled 10.5% interaction share of regret, App. D); it is the quantity the counterfactual-routing literature proposes to use; and it can be measured at training cost, whereas an interaction teacher would need $\binom{M}{2}$ evaluations per site rather than M. As a result, this experiment can bound unary exploitability and says nothing about coalition-aware training, which we did not run.

The experiment. On frozen OLMoE we finetuned only the router gates (2.1M parameters, all 16 layers) for 3000 steps with (a) LM plus load-balance loss alone, (b) the same plus listwise distillation from an online counterfactual teacher (equal-compute single-expert substitutions at sampled sites, measured on the current weights), and (c) the same plus the identical loss on teacher labels shuffled across candidates. The last arm is a placebo that keeps every gradient-scale property of the loss while removing its information content. We ran three teacher configurations: a single-token teacher at a small distillation weight, the same teacher at a weight comparable to the LM term, and a window-utility teacher that scores exactly the quantity the evaluation scores (Table A10). In no configuration does the informative arm show a benefit over the LM-only control beyond run-to-run noise. At the small weight, where all three runs were saved only as run summaries, the two distillation arms have identical validation perplexity to four decimals. The informative arm's oracle regret there

sits 0.30 replicate standard deviations below the four-replicate control mean, 0.00009 nats above the control run saved with it, and 0.00030 nats (1.2 replicate standard deviations) below its own placebo, which is another single summary-only run; two single runs cannot resolve a difference of this size (minimum detectable effect 0.00115 nats, `scripts/mde.py`; Table A10). At the larger weight the informative arm is $+0.00046$ nats *above* the four-replicate control mean — 1.9 pooled between-run standard deviations in the worse direction. It is also 0.00085 nats above its own placebo on the same sites (paired 95% interval $[+0.00012, +0.00180]$), a gap about as large as the one separating identically configured control replicates 1 and 3 (below). For the reason given in the replicate analysis below, every arm is compared against the four-replicate control mean rather than any single run. Except for the three rows Table A10 marks as summary-only, all oracle-regret figures are recomputed from the released per-site arrays rather than from run summaries. The window teacher is the configuration that removes the target-misalignment objection below, and with it we ran four replicates (Table A10). Their mean sits $+0.00039$ nats *above* the four-replicate control mean ($t = 2.30$, $p = 0.061$, between-run), and they show no reliable advantage over their own placebo. The replicate mean matches the placebo's 0.00538 to 4×10^{-6} nats, the replicates lie between 5×10^{-5} below and 9.6×10^{-5} above it, and every paired 95% interval spans zero.

Positive control. A null result cannot be interpreted unless the apparatus is shown to be able to move the endpoint. The offline experiment above has only a partial check of this kind (its ridge probe predicts a site-level scalar, a far easier target than an expert-indexed one; App. G), and a training experiment needs its own. We therefore added a deliberately leaky positive control, an arm whose teacher trains on the evaluation sites themselves and scores exactly the quantity the evaluation scores. It is not a method proposal. It is the strongest supervision the rig could receive, and it sets the ceiling against which a null becomes meaningful. It moves the endpoint decisively: oracle regret 0.00391 against a control mean of 0.00499, a shift of -0.00108 nats or **4.4 run-to-run standard deviations**, with router-is-best rising from $\approx$21% to 33.5% and validation perplexity unchanged (9.0320). Paired intervals exclude zero against three of the four control replicates, so the rig can move oracle regret. By the standard of §7 this is the weaker kind of control, since it sees the evaluation labels, and we have no label-free control for the online null. It shows that the endpoint can move but not how small an effect the design resolves; the minimum detectable effect below addresses that.

Detectable effect size. Training is not bit-reproducible on GPU, so we ran four replicates of the control and four of the window-teacher arm under identical configurations. The pooled between-replicate standard deviation is 0.000242 nats (the standard error of a four-run arm mean is 0.000121). This is the same order as every between-arm difference we measured, and site-level paired intervals cannot see it because they hold the trained model fixed. Three separate pairs of identically configured control runs are "significantly different" under the site-level paired test (control replicate 1 vs. 3: $+0.00082$, 95% CI $[+0.00027, +0.00146]$, $p = 0.008$; 2 vs. 3: $p = 0.022$; 3 vs. 4: $p = 0.029$). A training-arm comparison at this effect size that does not replicate is therefore measuring run-to-run noise. A fifth control run, saved only as a summary (0.00482, Table A10), is not in the replicate set. Adding it puts the window arm $+0.00043$ nats above the control mean (Student $p = 0.028$, Welch $p = 0.031$), in the harmful direction, so leaving it out hides no benefit.

With four replicates of the control and window arms and six residual degrees of freedom, the design detects a true difference between them of 0.00057 nats at 80% power (a normal approximation would give 0.00048 and overstate the resolution); a four-replicate arm against a placebo run once needs 0.00091 (`scripts/mde.py`). The two arms' replicate variances are also unequal. The control replicates have a standard deviation of 0.00034 against the distillation arm's 0.000065, a fivefold difference we cannot explain and simply report, so we quote the between-run comparison under both assumptions: Student $p = 0.061$, Welch $p = 0.099$. In terms of quantities this experiment measures directly, the detectable difference is 11% of the frozen model's endpoint (0.00526 nats) and about half the shift the positive control produces (0.00108). Effects smaller than 0.00057 nats are not detectable here, and no informative arm shows a benefit that large. Between runs the window teacher differs from the control by $+0.00039$ nats, a positive difference that means slightly worse routing, although at Welch $p = 0.099$ it cannot be distinguished from zero. It also has no reliable advantage over its shuffled-label placebo: two replicates fall below it, by at most 5×10^{-5} nats, about a fifth of a replicate standard deviation. The single-token placebo's apparent advantage over control replicate 1 (-0.00077 nats) is likewise within replicate noise.

Arm (3000 steps, gates only)	Teacher	Val. PPL	Oracle regret (nats)	Router-is-best
frozen (no training)	—	9.0641	0.00526	16.8%
control, replicate 1	none	9.0333	0.00536	20.0%
control, replicate 2	none	9.0294	0.00502	20.8%
control, replicate 3	none	9.0306	0.00454	21.5%
control, replicate 4	none	9.0324	0.00503	21.0%
control, summary-only	none	9.0321	0.00482	20.5%
counterfactual distillation, summary-only	single-token, small weight	9.0311	0.00491	19.3%
placebo (shuffled labels), summary-only	single-token, small weight	9.0311	0.00521	19.5%
counterfactual distillation	single-token, large weight	9.0349	0.00545	20.3%
placebo (shuffled labels)	single-token, large weight	9.0297	0.00459	22.8%
counterfactual distillation	window (w=8), rep. 1	9.0332	0.00535	18.0%
counterfactual distillation	window (w=8), rep. 2	9.0321	0.00538	19.8%
counterfactual distillation	window (w=8), rep. 3	9.0325	0.00547	19.5%
counterfactual distillation	window (w=8), rep. 4	9.0323	0.00533	20.5%
placebo (shuffled labels)	window (w=8)	9.0323	0.00538	22.0%
positive control (leaky)	evaluation sites themselves	9.0320	0.00391	33.5%

Table A10. Router-only finetuning arms, each evaluated on the same 400 held-out sites (lower oracle regret is better). Replicates share a configuration and differ only through GPU training nondeterminism. The pooled between-replicate standard deviation is 0.000242 nats; it sets the scale below which arm differences cannot be interpreted and gives the design a minimum detectable effect of 0.00057 nats at four replicates per arm, and 0.00091 for a four-replicate arm against a placebo run once. Rows marked summary-only (`results/stageB_olmoe_control.json`, `results/stageB_olmoe_unary.json`, `results/stageB_olmoe_placebo.json`) were saved before per-site arrays were logged, so they enter neither the replicate analysis nor any paired test. Validation perplexity is a run-level value in every row; oracle regret and router-is-best are recomputed from the per-site arrays wherever they exist (`scripts/router_training_summary.py`). The positive control trains on the evaluation sites themselves. It is not a method, but it shows the apparatus can move this endpoint by about twice the minimum detectable effect.

What these two results do and do not establish

What this does and does not claim. Three results bear on recoverability, and they agree. §5.3 trains a linear head to exploit the weighting gap at the intervened token and recovers none of it; App. G cannot predict the interaction term offline; App. H cannot distil the unary one online. §5.3's transfer measurement is a different quantity, and we do not offer it as an explanation for any of them. §5.3 does bound the scope: whatever a per-token method recovers is realised at the token it acts on, and it neither compounds nor is cancelled over the following eight.

This is narrower than a claim about aggregation methods in general. Contemporaneous work (Li, 2026) reports a gain from a learned aggregation head on this model family. Our measurement is an oracle on a one-parameter temperature family at fixed router mass, evaluated on next-token log-likelihood over an eight-token window. A trained head optimising a different objective over full sequences measures a different quantity, and our null does not refute it. We can say only that in this family a per-token oracle gain exists at the intervened token ($+0.08707$ nats), that it does not carry to

the eight tokens that follow (-0.00012, Table A8), and that neither of our heads recovers a resolvable part of it beyond one global temperature.

The obvious objection to a single-token teacher is that it optimizes a target which disagrees with these runs' window-utility metric ($w = 8$, positions $t \dots t{+}7$, selection token included) at 57.4% of sites (App. A; measured on the two-slot protocol's 896-site $w = 8$ arm rather than on the teacher's single-expert substitutions). This was the one live alternative explanation, and we test it directly. Batching all M candidate evaluations into one pass (verified against the sequential path to 1.3×10^{-5} nats, and 20-26$\times$ faster) makes a window teacher cost about what the sequential single-token teacher did, so we ran the aligned-target configuration under otherwise identical settings. The outcome is unchanged. Over four replicates the informative arm's mean oracle regret matches its placebo's to 4×10^{-6} nats, so target misalignment was not what made the counterfactual signal inert.

Two results, stated separately. App. G is an offline transfer null on the interaction residual specifically. We do not state it as a boundary on expert-indexed prediction generally, because trivial baselines recover a few percent of variance on the unary term at the top slots where our heads recovered none. At offline-measurement scale ($\approx$1.3k training sites per layer, orders of magnitude below the 10^5-10^7 an online teacher would generate at pretraining scale), no head we trained predicts per-site coalition utility differences on held-out documents, whether from the representation or from the router's own view. Three caveats apply: the marginal-slot protocol measures the smallest interaction effects (the top-slot variant has $3.15\times$ the utility spread), rank-16 bilinear heads may under-express the interaction, and pair means are unreliable (§5.2). That last caveat limits only a predictor that assigns each pair one fixed weight; it does not limit one conditioned on context. An interaction $\Gamma_{ij}(x) = b_{ij}\, z(x)$ with $\mathbb{E}[z] = 0$ has pair means near zero yet is predictable wherever $z(x)$ can be read from x. The context-conditioned heads of App. G are built to capture this kind of structure, and at this scale they did not find it. App. H is a null on the unary channel at a stated resolution. A positive control establishes that the apparatus moves the endpoint by 4.4 replicate standard deviations, and four replicates of the control and window arms give a minimum detectable effect of 0.00057 nats between them, about half the shift the positive control produces. Counterfactual distillation delivers no benefit that large, and its trend is slightly negative.

Neither result closes the other's question. A proposal to exploit coalition structure online must contend with App. G's transfer null (or show that a stronger predictor crosses it) and must supervise interaction directly, which no experiment here does. A proposal to exploit unary counterfactual signal at this scale has to explain why a placebo-controlled, positively-controlled experiment found no benefit larger than 0.00057 nats.

I Decision criteria and outcomes

Each claim below carries a numeric criterion. The criteria come from our internal study plans, which carry no timestamps and were updated as the work proceeded, and we hold no external registration (no OSF entry or timestamped hash). The table therefore records our own criteria and cannot independently show when each was fixed. The provenance note below says which criteria the plans record and which they do not, and readers should weigh the table accordingly. Four rows (P5, P6, P7, P9) record criteria we failed.

#	Claim	Criterion	Observed	Verdict
P1	Measurement is above the numerical floor	median effect / median replay floor ≥ 100	4.9×10^3 (OLMoE), 5.2×10^3 (granite)	met
P2	bf16 is not measurement-grade	same criterion applied to a bf16 probe	ratio of order 1	met (bf16 arithmetic excluded; App. A)
P3	Interaction energy is material	pooled IER median > 0.05	0.060 (0.081 under the df adjustment the criterion was stated with)	met

#	Claim	Criterion	Observed	Verdict
P4	Depth gradient is not a router-uncertainty effect	within-layer entropy terciles do not reorder layers	no reordering	met
P5	Co-activation predicts synergy	$\|\rho\| > 0.1$ at any layer, BH-corrected	max $\|\rho\| = 0.084$ over 44 tests, min $q = 0.45$	not met
P6	Interaction is predictable offline	held-out $R^2 > 0$ for any layer/target/feature set	$R^2 \leq 0$ everywhere	not met
P7	Counterfactual distillation beats control *and* placebo	both paired 95% CIs exclude zero, same sign	no informative arm; the one paired interval that excludes zero favours the placebo (large weight; App. H)	not met
P8	The training rig can move the endpoint	positive control shifts oracle regret > 2 replicate SD	4.4 SD	met
P9	Equal-compute headroom transfers to later tokens	transfer regret CI excludes zero on a window disjoint from the selection token	-0.00009, CI $[-0.00039, +0.00020]$	not met

Provenance. The plans record P3's threshold, stated on degrees-of-freedom-adjusted IER and on a window utility containing the selection token, which the plan named primary (App. A gives why single-token utility is primary here); P4's within-layer stratification; and P7's verdict rule. They do not record P1's $100\times$ ratio or P2's gate; P5's margin of 0.1, which tightens a planned 0.3; P6's criterion, held-out $R^2 > 0$, the natural null for any predictor; or P8's positive control, added after the training arms had returned their verdict. P9 is a post hoc formalisation. The plan read "select on single-token utility, evaluate on the independent window utility", and the disjointness requirement was made explicit only after the overlap was found. P5's BH family was not fixed in advance either; it grew as layers and models were added (App. L gives its stages). The reported q values are for the final 44-test family, in which no test departs from zero after BH under any choice of granite run (minimum q 0.33-0.45, App. E). P2 is a protocol gate rather than a scientific claim and is listed only for completeness. P5's equivalence form (TOST against $|\rho| < 0.1$) is reported in §5.2, but as that section notes, it is close to an arithmetic consequence of the measured reliability and so is not an independent test.

J CONFIGURATION

Exact artefacts. Checkpoints: `allenai/OLMoE-1B-7B-0924` at revision `6d84c48581ece794365f2b8e9cfb043c68ade9c5`, `ibm-granite/granite-3.0-1b-a400m-base` at `d91cbed802d85eb1b32374623331f6ab2b37403a`, and `Qwen/Qwen1.5-MoE-A2.7B` at `1a758c50ecb6350748b9ce0a99d2352fd9fc11c9`. The Qwen runs hold the checkpoint at its released bf16 and upcast at the point of use (App. A), so the stored weights are bf16 and all arithmetic is fp32; the other two families materialise fp32 from the same bf16 sources. The runs called `from_pretrained` without an explicit `revision=`, a defect we record (the released engine takes the revision as a parameter, `--revision` in `scripts/run_phase0.py`). We give the hashes because the GPU host that produced the CUDA sweeps and the machine used for analysis both resolve the OLMoE and granite checkpoints to exactly these snapshots, so those two pins are verified after the fact rather than asserted; the Qwen pin has no second machine to confirm it. The corpora, with the exact paths the loader uses, are FineWeb-Edu shard `sample/10BT/000_00000.parquet` (first 4,096 rows), MBPP plus HumanEval, GSM8K plus MATH (algebra subset), and OPUS-100 zh/de, with 34 documents per domain, 136 in all, truncated to 512 tokens. No site

manifest file is needed, since every released record carries `domain`, `doc_id`, `t` and `layer`. The exact set of measured sites, and with it the document sample, the position sample and the per-layer pairing, can therefore be recovered from the data itself, whether or not the corpus loader reproduces bit-for-bit. The software versions were Python 3.11.15, PyTorch 2.13.0 (CUDA 13.0), transformers 5.15.0, tokenizers 0.22.2, safetensors 0.8.0, datasets 5.0.1, numpy 2.4.6, scipy 1.17.1 and networkx 3.6.1. We did not freeze a lockfile at run time; these are the versions pinned in the environment that produced the runs, read back from it.

Sources and licences. The models are OLMoE (Muennighoff et al., 2025, Apache-2.0), granite-3.0-1b-a400m (Granite Team, IBM, 2024, Apache-2.0) and Qwen1.5-MoE-A2.7B (Qwen Team, 2024, Tongyi Qianwen licence). The corpora are FineWeb-Edu (Penedo et al., 2024, ODC-By), MBPP (Austin et al., 2021, CC BY 4.0), HumanEval (Chen et al., 2021, MIT), GSM8K (Cobbe et al., 2021, MIT), MATH (Hendrycks et al., 2021, MIT), OPUS-100 (Zhang et al., 2020, whose dataset card states no licence) and, for the precision pilots, WikiText-2 (Merity et al., 2017, CC BY-SA 3.0). We use them for research only, release no model weights and no corpus text, and the released per-site records hold document identifiers, token positions, expert indices and computed quantities only.

One run used a different stack. The granite deep-layer replication of App. E (layers 14 and 20, 2,580 sites over 516 documents) was produced on an Apple M3 Max under Metal rather than CUDA, with Python 3.13.2, PyTorch 2.12.0, transformers 5.12.0, numpy 2.2.6 and scipy 1.16.2, because the cloud instance was unavailable when that run was needed. When stacks are mixed inside a replication argument, provenance should be checked rather than asserted, so we checked it. We re-evaluated released granite records on the second machine, reading the recorded (document, position, layer, fixed set, pool) and forcing the same 28 coalitions so that the comparison depends on neither run's random number stream. This reproduces them to a median $|\Delta U|$ of 1.6×10^{-5} nats against a median utility spread of 6.6×10^{-2}, a ratio of 4.2×10^{3}. That difference sits at the level of granite's own replay floor (1-4×10^{-5} nats, App. A), so the two stacks agree to within the noise the measurement already carries, and the engine gate passes on the second machine at a median effect-to-floor ratio of 1.9×10^{3}. Every other number in this paper comes from the CUDA stack above.

Hardware. We used rented cloud GPUs with non-reference memory configurations, and state this explicitly because the capacities do not match NVIDIA's published specifications for these model names. An RTX 4090 board carrying 48GB (reference: 24GB) ran the formal sweeps, since fp32 OLMoE does not fit in 24GB. An RTX 4080 SUPER board carrying 32GB (reference: 16GB) ran the engine consistency check of App. A and every follow-up run; there `nvidia-smi` reports 32,760 MiB and fp32 OLMoE fits with roughly 4GB to spare. The released outputs of the GPU gates (`scripts/gpu_gates.sh`) record the board as the driver names it, NVIDIA GeForce RTX 4080, with the same 32,760 MiB and the software stack above. Nothing in the measurement depends on the board. The capacity determines which models can be materialised at fp32, and App. A's upcast is what removes that dependence. The laptop pilot is CPU-only and is the source of no reported number.

Measurement engine. fp32 arithmetic throughout, on checkpoints released in bf16 (App. A); OLMoE-1B-7B (16 layers, 64 experts, Top-8, raw-softmax gates) and granite-3.0-1b-a400m (24 layers, 32 experts, Top-8, gates renormalized over the selected logits). Documents truncated to 512 tokens; positions drawn uniformly from $[32, n-1)$, 5 per document, 34 documents per domain over four domains. Candidate pool $M = 8$ (two slot occupants, three next-by-rank, three uniformly random drawn without replacement from the 53 experts outside the pool); $M = 10$ with three random for the three-slot sweep.

Oracle regret has two uses in this paper, and they are different quantities. In §5 it is $\max_S U(S) - U(S_{\text{router}})$ over the $\binom{M}{2}$ measured pairs, on single-token utility. In App. H and Table A10 it is the endpoint of the training experiment, taken over the $M = 8$ single-expert substitutions at the weakest routed slot, on the $w = 8$ window utility, at 400 fixed held-out sites. The App. H endpoint therefore includes the intervened token in its window. Because every arm is scored on identical sites with identical windows, this is common to all arms and cancels in the between-arm differences that section reports, but it means the absolute level in Table A10 is not comparable to the §5 oracle gaps.

Aggregation family. $\tau \in \{0.2, 0.35, 0.5, 0.7, 1.0, 1.4, 2.0, 3.0, 5.0, 10^6\}$, with $w_e(\tau) \propto p_e^{1/\tau}$ renormalized to preserve $\sum_{e \in S} p_e$. $\tau = 1$ reproduces the model's own weights exactly.

Router-only training (App. H). Gates only, 2.1M parameters, all 16 layers trainable, everything else frozen. AdamW, lr 10^{-5}, 200-step warmup, cosine decay, weight decay 0, gradient clip 1.0,

batch 2 sequences of 1,024 tokens, 3,000 steps, load-balance coefficient $\beta = 0.001$. Distillation weight $\alpha = 1.0$ for the large-weight single-token and window arms of Table A10, and $\alpha = 0.1$ for its small-weight single-token arms, which were saved only as run summaries (App. H). Loss: listwise KL from $\mathrm{softmax}(u/\tau_u)$ with $\tau_u = \mathrm{std}(u)$ onto the gate logits restricted to the M candidates.

Teacher budget. 4 sites per step $\times$ 8 candidates $\times$ 3,000 steps $= 96{,}000$ counterfactual evaluations per arm; at $w = 8$ that is 768,000 position-evaluations. The teacher consumed 56% of wall-clock. This is the "scale" at which App. H's null holds, and it is small, since an online teacher at pretraining scale would generate several orders of magnitude more.

Evaluation sites (App. H). 400 sites, drawn from a held-out FineWeb-Edu shard only (not the four-domain corpus), fixed across arms by seed so that all comparisons are paired over identical sites; window $w = 8$.

Weighting head (§5.3, Table A9). Ridge regression, $\lambda = 1$, predicting the gain of each of the ten τ settings and taking the argmax; no early stopping, no hyperparameter search. Features: the eight gate weights of the routed set sorted descending and mass-normalised, their entropy, top-1 and top-2 weight and their difference, total router mass, router entropy, $\log$ top-1, and four layer indicators (19 dimensions); the second variant appends the top 64 principal components of the layer-input representation, centred and unit-scaled, fitted inside each training fold (83 dimensions). A transductive variant, with the PCA fitted before the fold split, is shown in Table A9 for comparison only (-3.7% of oracle against the in-fold -0.9%; `scripts/run_weighting_head.py`). Five-fold cross-validation uses folds formed over documents, so no document appears in both train and test. The positive control is a planted signal rather than a leaked target: it is built by moving rows of the gain matrix D, and no gain is recomputed. Within each layer we rank sites by their projection on the leading principal component v of the layer input (computed once over the whole sample and pooled over layers, where it agrees with every fold's own leading component at $|r| \geq 0.999$; 81% of its variance lies between layers, and the planting uses only its within-layer order) and rank the rows of D by the index of their best τ among the ten candidates, ties broken at random; with probability α a rank slot receives its rank-matched row, and otherwise a random row of the same layer. At $\alpha = 1$ this is a within-layer permutation of D with the same ten discrete candidates, the same gains and the same oracle ($+0.08707$), and its best τ is monotone in v (mean within-layer Spearman 0.974); at $\alpha = 0$ it is a pure within-layer shuffle; in between, rows can repeat (Spearman 0.683 and 0.489 at $\alpha = 0.7$ and 0.5; planted oracles $+0.08834$ and $+0.08942$). Labels never enter the feature matrix, and because the planting is layer-stratified the layer indicator offers no shortcut. The identical pipeline recovers $+34.7\%$ of the planted oracle at $\alpha = 1.0$, $+16.4\%$ at 0.7, $+6.8\%$ at 0.5 and -8.8% at $\alpha = 0$ (no planted structure). α is a mixing probability rather than a strength, and each value above is one planting draw (generator seed 7), which fixes the tie-breaking among rows with the same best τ (the only randomness at $\alpha = 1$) and, below $\alpha = 1$, which slots keep their rank-matched row and which rows fill the rest. Re-planted with 200 independent seeds on the same folds and head (`scripts/weighting_head_planted_seeds.py`), the recovery averages $+35.8\%$ at $\alpha = 1$ (sd 2.8; 95% of draws within $[+30.0\%, +41.0\%]$, range $+29.2\%$ to $+43.8\%$), $+14.2\%$ at 0.7 (sd 4.0; $[+6.4\%, +22.2\%]$), $+3.3\%$ at 0.5 (sd 3.8; $[-4.6\%, +10.2\%]$) and -9.8% at 0 (sd 3.2). The last of these is a within-layer shuffle and reproduces the permutation null of §5.3. The seed-7 draws sit at the 37th, 70th, 81st and 62nd percentiles, so $+34.7\%$ is typical and $+6.8\%$ was favourable. Flipping the sign of v moves the $\alpha = 1$ mean by $+0.8$ points ($+36.6\%$, sd 2.9), and re-drawing the fold partition (20 partitions crossed with 20 plantings) leaves the $\alpha = 1$ spread unchanged (sd 2.8; partition means $+34.4\%$ to $+36.4\%$). Refitting the whole pipeline inside disjoint document halves (300 splits, pooled over 20 plantings) puts the variances of the planted contrasts at $1.47\times$ (gain over $\tau = 1$) and $1.64\times$ (head minus fixed τ) their cluster-robust values at $\alpha = 1$. With that inflation the seed-7 draw recovers $+34.7\%$ $[+23.1\%, +46.4\%]$, and with each α's own ratio the paired head-minus-fixed-τ interval on the planted target excludes zero in 100% of draws at $\alpha = 1$, 87% at 0.7 and 14.5% at 0.5. The control therefore shows that the pipeline detects a planted signal about as structured as $\alpha = 0.7$ (within-layer Spearman 0.68 between v and the best τ) in most draws, and one as weak as $\alpha = 0.5$ (0.49) rarely. The label-appending control (99.3% of the oracle) is implemented in the same script. Targets and evaluation are the single-token utility at the intervened position.

A note on the IER convention. All IER and energy-share figures in this paper are raw ratios of sums of squares, without a small-sample degrees-of-freedom adjustment. That adjustment is inappropriate here for a reason specific to this design, which we state because it is easy to get wrong. The utilities

are deterministic and every coalition in the design is enumerated, so the fits are exact orthogonal projections and the residual reflects actual non-additivity rather than sampling noise; there is no mechanical shrinkage to de-bias. Algebraically, the adjusted quantity equals the observed residual energy divided by the energy a uniformly random utility direction would leave. It is therefore a per-dimension enrichment relative to an isotropic null and cannot be read as a share. Because both designs are fixed, the adjustment is a single constant per protocol ($27/20 = 1.35$ for the two-slot sweep at every site, $110/75$ for the three-slot sweep). Under either convention every gradient, stratification, control, cross-layer comparison and null in this paper has the same shape, and only absolute levels differ. Where the enrichment reading is useful we report it separately and under that name (Table A5, last row). The released records carry both conventions (`ier` and `ier_adj`, `high_order_share`), so either can be recomputed.

Statistics. Document-clustered bootstrap, 2,000 resamples, seed 0, for every CI unless stated. Spearman intervals via Fisher z with the pair-level standard error $1.06/\sqrt{n-3}$ (§5.2). Winsorization at 1%/99% applied only to the predictors and targets of the App. G regressions.

K Related work in full

Sparse MoE. GShard, Switch Transformer, and GLaM established large-scale token-choice Top-K routing; BASE formulated routing as balanced assignment; Expert Choice reversed the selection direction; Mixtral, DeepSeekMoE, and DeepSeek-V3 brought sparse MoE (with fine-grained and shared experts, and bias-based balancing) into open LLM practice. All of them rank independent per-expert scores, and the inductive bias they share in doing so is our object of study.

Observational expert-group analyses. "Specialization through Collaboration" (Tang et al., 2026) finds hierarchical cross-layer co-activation groups by dictionary learning; "The Illusion of Specialization" (Wang et al., 2026b) reports a domain-invariant standing committee; A Closer Look into MoE (Lo et al., 2025) documents norm-driven router preferences. These works establish group structure in routing statistics. The first also uses its groups to guide expert pruning, an intervention that tests compression rather than interaction. Our §5.2 tests one implication of the shared premise, namely that token-level, same-layer co-activation predicts conditional pairwise interaction, and finds that it does not. Cross-layer, sequence-level groups and the compression they enable are outside what we test (App. E's scope limits). "The Myth of Expert Specialization in MoEs" (Wang et al., 2026a) suggests why the premise is weak: because MoE routers are linear maps, hidden-state similarity explains expert-usage similarity on five pre-trained MoEs, so specialization is a property of the representation space rather than of the routing architecture. The same holds for a pair at a fixed layer. Whether two experts are co-selected is decided by that layer's router weights and its input, and neither expert's output enters the decision, whereas Γ_{ij} is what the two outputs do together downstream. Co-activation could still track synergy if training aligned router and experts (the router–expert weight coupling of Ahrac et al. (2026), below, is one such channel), so whether it does is an empirical question, which §5.2 tests for same-layer pairs. Cross-layer co-activation differs in kind, since an upstream expert's output does enter a downstream routing decision. Convergent evidence from a different intervention appeared while this work was in progress. "Co-activation Proposes, Ablation Disposes" (Xu, 2026) clusters attention heads (not experts) by co-activation and validates the clusters by ablation. In OLMoE, head co-activation shows structure only after examples are stratified by their expert-routing pattern, and the recovered head community is statistically real against a random-partition null yet fails closure (ablating it lowers loss); its dense-model communities pass. That result and ours point the same way but differ in object and method: it studies attention heads and ablates discovered communities, while we study experts, compare equal-compute alternatives and decompose their utility into unary and pairwise terms.

Pre-registered causal tests of modularity. Closest in method and opposite in scale, Salomone et al. (2026) run a pre-registered causal ablation test of expert modularity on a frontier MoE (Command A+) and report that apparent modularity mostly dissolves, with only one of six hypothesised expert families showing modularity that survives independent validation. Their intervention is coarser than ours, since they ablate expert families instead of swapping members at equal compute, and their unit is a family rather than a pair; the logic of intervening instead of observing and the direction of the result are the same. Their scale also bears on our App. A claim that precision bounds what can be

measured. A frontier-scale causal test is evidently possible under a different protocol, and our bound applies to our own protocol and implementation rather than to interventional MoE analysis in general.

Coalition attribution for pruning. Shapley-MoE (Huang et al., 2025) and SHAPE (Zhang, 2026) attribute performance to experts (and coalitions) for offline pruning; C-GNN-PRUNE (Li et al., 2026) prunes with a co-activation-plus-similarity graph. These methods are post hoc and model-level. Our measurements are per-token and equal-compute, and our results caution that co-activation-derived edges do not reflect interaction. A causal audit of unary expert importance (Engmann et al., 2026) makes the observational-to-interventional gap explicit on three MoEs including OLMoE. It finds that no observational metric predicts interventional importance and uses a positive control to show the null is not a power failure, the role our planted-signal control plays for the weighting head (§5.3). We extend that gap to sets: that audit asks whether a single expert matters, and we ask whether the combination does.

Online interaction-aware routing. SymphonySMoE (Nguyen-Nhat et al., 2025) refines gate scores with an EMA co-activation "social graph"; CoR (Hu et al., 2026) redistributes activations at inference using unary counterfactual expert impact at constant activation count. The first is observational and the second is unary. (Chain-of-Experts, Wang et al., 2025, changes the execution model to sequential intra-layer communication while leaving the selection rule alone, and is orthogonal to our question.) Our decoupling result (§5.2) directly questions the premise of co-activation-graph routing, and our offline transfer null (§6) quantifies what supervision set-level extensions would need.

Counterfactual routing analysis. "When Are Experts Misrouted?" (Yoon et al., 2026) introduced equal-compute counterfactual route comparison on frozen MoEs, the executed-route-only supervision problem, and fragile-token headroom. We build on that measurement protocol and extend it with the unary/interaction decomposition, the random-candidate arm, the disjoint-window transfer measure, and the interaction-specific diagnostics. Equifinality results at small scale (routing topology underdetermining LM quality; Ternovtsii and Bilak, 2026) concern overall quality rather than whether set utility is additive; our own data speak to additivity, which fails materially through the middle of the network and is close to exact only at the deepest layers. "Routers Learn the Geometry of Their Experts" (Ahrac et al., 2026) shows router-expert weight coupling, a mechanism by which co-activation can arise without functional coupling, which is consistent with our findings. ("Coherent Overlap," Tian et al., 2026, reports that actual routes explain token representations better than matched alternatives. That is a representation-level claim, whereas our equal-compute comparisons are at the loss level, where alternatives do beat the router at many sites; the two coexist because representational fidelity and next-token utility are different objectives.) A contemporaneous paper isolates the complementary axis of the same problem. "Beyond Routing" (Li, 2026) holds the routed identities, the expert computation, and the total router mass fixed on the same model and varies only the within-set aggregation weights. It reports a comparable router-suboptimality rate (its top-scored expert is best at 17.2% of sites) and, unlike our selection-side attempts, exploits the gap successfully with a 301K-parameter aggregation head. §5.3 and App. H state the boundary this draws.

Interaction decompositions. Our decomposition is a fixed-cardinality functional-ANOVA restriction (Hooker, 2007) with sum-to-zero identification; at fixed $|S| = K$ the pairwise term is identified only up to unary-absorbable components, which our estimator removes by construction. The cooperative-game literature formalizes the same object more generally. Grabisch and Roubens (1999) axiomatize interaction indices over all coalition sizes, the Shapley–Taylor (Dhamdhere et al., 2020) and Faith-Shap (Tsai et al., 2023) indices give faithful finite-order truncations, and Archipelago (Tsang et al., 2020) detects interactions in a model-agnostic attribution setting. We use the restricted least-squares form instead because, under an equal-compute constraint, every evaluable coalition has $|S| = K$ (in the two-slot design, the $\binom{M}{2}$ completions of a fixed background of $K - 2$ routed experts), whereas those indices integrate over the full subset lattice. Estimating a Shapley-style index here would require evaluating coalitions of other sizes, which changes the active-compute budget and therefore changes the quantity being measured. Under the current two-slot support, Γ is all the non-additivity the design can express. It is not the largest interaction notion identifiable at equal compute, since varying more slots at the same $|S| = K$ identifies higher orders, as the three-slot sweep does ($\binom{10}{3} = 120$ equal-compute coalitions per site, App. D). The price is that it is an upper bound relative to an additive model class rather than an axiomatically unique index.

L Analysis revisions made during the study

Several analyses in this study were first computed one way and then replaced. The body reports only the analysis used; this appendix collects every such revision in one place, giving what was first computed and what it gave, what was wrong with it, and what replaced it, so that no result above has to be read next to a figure it does not use. The first column gives the §7 protocol item a revision illustrates, where there is one; items 5 and 7 have no row, because their evidence is stated with them in §7. The bf16 probe of the first row is released as `results/qwen15_moe_bf16_replay.txt`, the miscounting CSA bootstrap as `scripts/csa_clustered.py`, and the same bootstrap counting every drawn copy as `scripts/csa_bootstrap_corrected.py`; `results/pair_reliability.json` records the 16-test stage of the BH family.

Item	First computed, and what it gave	What was wrong	What replaced it, and what it gives
1	Qwen1.5-MoE-A2.7B probed at bf16 arithmetic: effect-to-floor ratio of order 1, so the model was first excluded	bf16 *arithmetic*, not bf16 storage, was the binding limit, and the whole model at fp32 did not fit one card	fp32 arithmetic on bf16 storage, upcast at the point of use: effect-to-floor 3.97×10^4, Qwen measured (App. A)
2	router-entropy terciles pooled over layers: a dramatic "entropy gradient"	a depth confound: layers differ in both entropy and IER	terciles within layer: flat at layer 2, 0.250 / 0.288 / 0.283 (App. D)
3	energy ratios with a small-sample degrees-of-freedom adjustment: every two-slot IER $\times 27/20$, the beyond-pairwise share $\times 110/75$, read as a share	utilities are deterministic and every coalition is enumerated, so the adjusted ratio is a per-dimension enrichment, not a share	raw ratios (Table A3 and Table A5; the enrichment reported under its own name)
4	the coalition selected on token t scored on a window containing t ($w = 8$, 896 sites): headroom 0.0048 nats, read as transfer; the single-token and window utilities' best coalitions agreeing at 42.6% of sites	the window contains the selection token, so $1/w$ of the same-token gain leaks into it	the window $t{+}1 \ldots t{+}8$: -0.0001 nats (§5.3); agreement 6.2%, against 4.4% for independent choices (App. A)
6	co-activation and synergy reliabilities at unmatched per-pair sampling	co-activation had $2{,}013$ observations per pair against synergy's 25-27, so the contrast came out 1.8-$2.8\times$ too large (layers 10 to 2)	matched per-pair sampling: co-activation 0.32-0.53 (§5.2)
6	percentiles of split-half reliability over splits of the same documents: upper ends 0.109 / 0.105 / 0.160 (OLMoE / Qwen / granite) used as a bound	dispersion over splits of fixed documents is not a population interval	disjoint documents: $\tau \leq 14.4\%$ / 7.5% directly; 9.1% / 5.4% under a Gaussian model (§5.2, App. E)
6	background-arm reliability without the distinct-site rule, and unmatched figures: $+0.046$ / $+0.051$ at layers 2/6, read as holding the conditioning set fixed raising reliability several-fold	a transplant pair's two halves could share sites, and the unmatched figures are not reproduced by the released records	eight distinct sites per pair at matched sampling: $+0.004$ / $+0.002$ (App. E)
6	the spread of Γ across backgrounds against the natural design's within-pair spread: $1.40\times$ $[1.10, 1.71]$ at layer 2, read as the background moving the estimand more than everything the natural design varies	the ratio is mostly a difference of scale between the arms ($1.44\times$)	the cross-background correlation: $r = 0.031$ / 0.003 (§5.2)
—	the background sweep's variance split by subtraction: the transplant arm's context term subtracted from the natural arm's within-pair variance, and the remainder read as the conditioning set's share of the natural design	in the natural protocol the background is a function of the token, so the subtraction assigns every context-background interaction and covariance term to the background	the transplant arm's within-token share, 92.9% / 91.9%, stated as a sensitivity under the transplant rule, beside the cross-background correlation (App. E)

Item	First computed, and what it gave	What was wrong	What replaced it, and what it gives
—	the CSA standard error $1.06/\sqrt{n-3}$ described as Bonett and Wright's	it is about 6% wider than Bonett and Wright's at $\rho = 0$	named the pair-level standard error, with Bonett and Wright's and Fieller, Hartley and Pearson's reported beside it (App. E)
—	a document bootstrap for the CSA that counted a document drawn twice only once in co-activation: standard errors a median $0.67\times$ the pair-level one and 0.73-0.78$\times$ the empirical σ inside document halves, i.e. 22–27% too small; the granite layer-20 lean at $q = 0.043$ within granite's 12 tests	the duplicate miscount (counting every drawn copy changes the standard error by under 2% at the median), and re-averaging noisy pair means inside each draw, which no counting fixes	the pair-level standard error, checked on disjoint halves: $0.92\times$; nothing survives BH (§5.2)
—	the group partition rebuilt on the pooled records once 9,520 sites were added: -3.07×10^{-5} nats, $p = 0.565$, read as the effect vanishing	a different partition from the one that produced the discovery (adjusted Rand index 0.56)	the partition fixed on the original records: $+6.99 \times 10^{-4}$ nats, $p = 0.027$ (§5.2)
—	the layer-2 group-level discovery test on the original records, 4,000 permutations: $q = 0.053$ over seven layers	Monte Carlo error as large as the distance to the BH boundary	100,000 permutations: $q = 0.046$ over seven layers (App. E)
—	the granite rows at layers 0 and 1 of the CSA family from the original run (680 sites): layer 0 $\rho = +0.116$, $p = 0.015$	a single sample; on 380 unseen documents layer 0 gives $+0.0009$	the extended run (516 documents), chosen after that result had been seen: -0.013 and $+0.001$; the alternative choices give 30-38 equivalences of 44 and no BH departure (App. E, Table A12)
—	P5's BH family of 8, then 16, then 26 tests	not fixed in advance; it grew as layers and models were added	the final 44-test family (App. I); BH q is not monotone in family size, so its q values are not uniformly smaller or larger than the smaller families'
—	the residual-stream PCA of the weighting head fitted before the fold split: the head recovers -3.7% of oracle	transductive: each training fold saw the test folds' features	the PCA fitted inside each fold: -0.9% (§5.3)
8	a positive control that appends the ten true gains to the features: 99.3% recovered	it regresses the target on a design containing the target	a planted signal the head never sees as a label: $+34.7\%$ (§5.3)

M THE 44-TEST CO-ACTIVATION/SYNERGY FAMILY, TEST BY TEST

Model	Layer	Support	Run	Pairs	ρ	95% CI	p	q	TOST p	TOST q
OLMoE	0	full	shallow	1,978	-0.009	$[-0.056, +0.038]$	0.703	0.956	<0.001	**<0.001**
OLMoE	0	random	shallow	1,858	-0.027	$[-0.075, +0.021]$	0.273	0.850	0.001	**0.003**
OLMoE	1	full	shallow	1,973	$+0.003$	$[-0.043, +0.050]$	0.889	0.966	<0.001	**<0.001**
OLMoE	1	random	shallow	1,858	$+0.003$	$[-0.045, +0.051]$	0.894	0.966	<0.001	**<0.001**
OLMoE	2	full	primary	2,016	-0.014	$[-0.060, +0.033]$	0.564	0.926	<0.001	**<0.001**
OLMoE	2	random	primary	2,016	-0.002	$[-0.049, +0.044]$	0.922	0.966	<0.001	**<0.001**
OLMoE	4	full	shallow	1,965	$+0.014$	$[-0.032, +0.061]$	0.545	0.926	<0.001	**<0.001**
OLMoE	4	random	shallow	1,842	$+0.037$	$[-0.011, +0.085]$	0.132	0.729	0.005	**0.009**
OLMoE	6	full	primary	2,016	$+0.022$	$[-0.024, +0.069]$	0.343	0.850	<0.001	**0.001**
OLMoE	6	random	primary	2,016	$+0.020$	$[-0.026, +0.067]$	0.389	0.850	<0.001	**<0.001**
OLMoE	10	full	primary	2,015	$+0.002$	$[-0.044, +0.049]$	0.921	0.966	<0.001	**<0.001**
OLMoE	10	random	primary	2,015	-0.012	$[-0.058, +0.034]$	0.611	0.926	<0.001	**<0.001**
OLMoE	14	full	primary	2,016	-0.003	$[-0.049, +0.044]$	0.907	0.966	<0.001	**<0.001**
OLMoE	14	random	primary	2,016	$+0.024$	$[-0.022, +0.070]$	0.304	0.850	<0.001	**0.001**
granite	0	full	extended	496	-0.013	$[-0.106, +0.080]$	0.785	0.966	0.034	**0.046**

Model	Layer	Support	Run	Pairs	ρ	95% CI	p	q	TOST p	TOST q
granite	0	random	extended	496	$+0.001$	$[-0.092, +0.094]$	0.984	0.998	0.019	**0.027**
granite	1	full	extended	406	-0.018	$[-0.121, +0.086]$	0.739	0.956	0.059	0.072
granite	1	random	extended	406	-0.028	$[-0.131, +0.075]$	0.594	0.926	0.086	0.099
granite	2	full	original	444	$+0.042$	$[-0.057, +0.140]$	0.406	0.850	0.124	0.136
granite	2	random	original	436	$+0.011$	$[-0.089, +0.110]$	0.836	0.966	0.039	0.051
granite	8	full	original	496	$+0.020$	$[-0.073, +0.113]$	0.669	0.956	0.047	0.059
granite	8	random	original	496	-0.043	$[-0.135, +0.051]$	0.372	0.850	0.113	0.128
granite	14	full	original	496	-0.057	$[-0.150, +0.036]$	0.230	0.850	0.183	0.192
granite	14	random	original	496	-0.050	$[-0.143, +0.043]$	0.291	0.850	0.148	0.159
granite	20	full	original	496	-0.074	$[-0.166, +0.020]$	0.122	0.729	0.289	0.296
granite	20	random	original	496	-0.084	$[-0.176, +0.009]$	0.077	0.729	0.369	0.369
Qwen	0	full	Qwen	1,769	$+0.041$	$[-0.009, +0.090]$	0.106	0.729	0.009	**0.014**
Qwen	0	random	Qwen	1,767	$+0.047$	$[-0.003, +0.096]$	0.064	0.729	0.017	**0.025**
Qwen	1	full	Qwen	1,770	$+0.003$	$[-0.047, +0.052]$	0.910	0.966	<0.001	**<0.001**
Qwen	1	random	Qwen	1,770	$+0.021$	$[-0.028, +0.070]$	0.405	0.850	<0.001	**0.002**
Qwen	2	full	Qwen	1,769	-0.030	$[-0.079, +0.019]$	0.235	0.850	0.003	**0.004**
Qwen	2	random	Qwen	1,767	-0.040	$[-0.089, +0.009]$	0.112	0.729	0.009	**0.013**
Qwen	4	full	Qwen	1,770	-0.029	$[-0.078, +0.021]$	0.252	0.850	0.002	**0.004**
Qwen	4	random	Qwen	1,769	-0.065	$[-0.114, -0.015]$	0.010	0.445	0.080	0.095
Qwen	6	full	Qwen	1,770	$+0.015$	$[-0.035, +0.064]$	0.562	0.926	<0.001	**<0.001**
Qwen	6	random	Qwen	1,769	$+0.015$	$[-0.034, +0.065]$	0.545	0.926	<0.001	**<0.001**
Qwen	10	full	Qwen	1,770	$+0.024$	$[-0.025, +0.074]$	0.333	0.850	0.001	**0.002**
Qwen	10	random	Qwen	1,770	$+0.056$	$[+0.007, +0.105]$	0.026	0.581	0.039	0.051
Qwen	14	full	Qwen	1,770	$+0.010$	$[-0.039, +0.060]$	0.686	0.956	<0.001	**<0.001**
Qwen	14	random	Qwen	1,770	-0.014	$[-0.064, +0.035]$	0.572	0.926	<0.001	**<0.001**
Qwen	18	full	Qwen	1,770	-0.009	$[-0.058, +0.041]$	0.726	0.956	<0.001	**<0.001**
Qwen	18	random	Qwen	1,769	$+0.022$	$[-0.027, +0.071]$	0.384	0.850	<0.001	**0.002**
Qwen	22	full	Qwen	1,770	-0.016	$[-0.065, +0.033]$	0.523	0.926	<0.001	**0.001**
Qwen	22	random	Qwen	1,770	-0.000	$[-0.049, +0.049]$	0.998	0.998	<0.001	**<0.001**

Table A12. The 44-test co-activation/synergy family, test by test (`scripts/csa_calibrated.py`, `results/csa_calibrated.json`; `scripts/csa_family_sensitivity.py` writes the same rows, with each row's record file, documents and sites, to `results/csa_family_table.md`). ρ is the Spearman correlation between a pair's co-activation count and its mean Γ over the pairs with at least three observations. Support is either full (the full candidate pool) or random (pairs with at least one uniformly random candidate). The interval, the two-sided p and the TOST p against $|\rho| = 0.1$ use Fisher-z with standard error $1.06/\sqrt{n-3}$, and both q columns are BH-FDR over all 44. Bold marks the 32 equivalences; entries shown as <0.001 are given exactly in the JSON. The run fixes the sample set. Primary is the three-seed OLMoE sweep, with 136 documents and 2,040 sites per layer at 2,013 distinct positions (27 positions per layer were drawn by two seeds, with the random candidates redrawn). Shallow is the one-seed OLMoE extension, with the same 136 documents, 680 sites per layer and an independent draw of positions. Extended is the granite extended run, with 516 documents and 2,580 sites per layer, made up of the 136 original documents at a new draw of positions (174 of those 680 sites coincide with original ones) plus 380 unseen documents. Original is the granite original run (136 documents, 680 sites per layer), and Qwen is the Qwen run (272 documents, 1,360 sites per layer). The extended run at granite layers 0 and 1 was chosen after the original run's layer-0 result had been seen; under the alternative choices equivalence holds at 30-38 of 44 tests and no test departs from zero after BH (minimum q 0.33-0.45; App. E). The granite layer-0 value $+0.074$ quoted with the layer-0 replication is not a family test, since it pools the original run's layer-0 sites with the extended run's unseen ones.