COHERENT OVERLAP: TOKEN-SPECIFIC EXPERT COMMITTEES IN SPARSE MIXTURE-OF-EXPERTS ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Why do sparse mixture-of-experts (MoE) language models route a token to several experts? The standard explanation is geometric complementarity: co-selected experts cover different directions of the token representation. This view has shaped non-overlapping expert designs, diversity objectives, and similarity-based compression, but it has a sharper contextual implication: a selected expert should become more valuable with the experts selected beside it. We test this implication with a prefix-controlled 2×2 factorial that crosses the selected expert with its highest-scoring unselected rival and the actual co-selected context with load-matched alternatives. A positive interaction supports complementarity, whereas a negative interaction supports *relevance-driven selection*, where the router forms a token-specific committee of individually relevant experts from a shared geometric neighborhood. Across six open MoEs, routed-input subspaces share a large layer-wide core and actual routes fit tokens better than matched alternatives. The selected expert beats its rival in 100 of 102 geometric cells, yet the actual context narrows this advantage in all 102. In OLMoE, Mixtral, and DeepSeek-MoE, the negative interaction persists under order-free contexts, sensitivity analyses, and next-token loss, which fits no subspaces. The overlapping experts remain functionally valuable: their outputs separate more than their inputs, later experts beat unselected substitutes in next-token loss, Top-2 beats Top-1 at matched compute, and explicit complementarity regularization does not consistently improve validation loss. We call this pattern **coherent overlap**: MoE routers select useful expert committees from overlapping neighborhoods, so geometric similarity alone is a poor guide for analyzing, pruning, or designing experts. Code is available in the supplementary material.

1 INTRODUCTION

Sparse mixture-of-experts (MoE) language models send each token through a small set of experts chosen by a learned router (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022), and most current MoEs route each token to several experts at once (Jiang et al., 2024; Muennighoff et al., 2025; Dai et al., 2024; Yang et al., 2025). This makes multi-expert routing more than a scaling trick: it raises a mechanistic question with direct consequences for analysis, compression, and architecture design. **When the router has already found one relevant expert for a token, what do the other selected experts add?**

A natural answer is *geometric complementarity*: experts specialize along different representation directions, and the router assembles a route that covers different parts of the token. This idea underlies non-overlapping expert designs (Dai et al., 2024), diversity and orthogonality objectives (Liu et al., 2024; Do et al., 2025; Park & Park, 2026), and similarity-based pruning and merging (Lu et al., 2024; Zhang et al., 2025; Li et al., 2026; Hu et al., 2026). It also suggests a powerful heuristic: geometric overlap should reveal removable computation. The crucial claim, however, is contextual. Complementarity is present only if a selected expert becomes more valuable because of the experts selected with it. Route selectivity and single-expert fit do not test that claim.

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

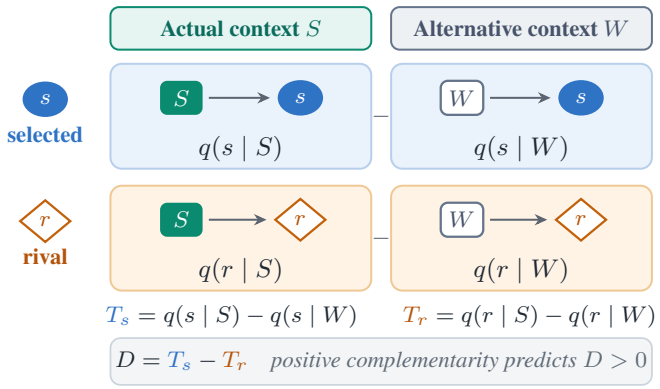

Figure 1: Candidate-by-context complementarity test. The 2×2 factorial crosses the selected expert s and highest-scoring unselected rival r with the actual prefix S and matched alternatives W. The interaction $D = A_{\text{actual}} - A_{\text{alt}} = T_s - T_r$ is positive under complementarity and negative under relevance-driven selection.

We contrast this with *relevance-driven selection*. Here the router builds a token-specific committee from a shared geometric neighborhood: selected experts are individually relevant to the token and therefore overlap in router-input directions, while their nonlinear maps can still differ. The two accounts agree on route coherence and candidate fit: actual routes should beat load-matched alternatives, and a selected expert should beat a strong unselected rival. They diverge only at the interaction. Complementarity predicts that the co-selected context leaves a residual direction the selected candidate is especially suited to fill, increasing its advantage. Relevance-driven selection predicts that the co-selected context has already covered many of the token-relevant directions that make the selected candidate strong, shrinking the same advantage.

We isolate this interaction with the prefix-controlled 2×2 factorial in Figure 1, which evaluates the selected expert and its highest-scoring unselected rival after the actual prefix and after load-matched, route-disjoint alternatives. The prefix is an analytical device for a parallel Top-k layer, so we also test order-free variants. Separating candidate from context is necessary: route comparisons that replace candidate and context together can make later selected experts look worse than controls, while fixed-context comparisons show they are better candidates (Figure 3).

The evidence supports coherent overlap. Across six open MoEs, routed-input subspaces share a large layer-wide core and lie much closer than random subspaces, yet actual routes fit tokens better than load-matched alternatives in all 18 model-layer cells (Figure 2). The factorial separates selection from complementarity: in OLMoE, Mixtral, and DeepSeek-MoE, selected candidates beat their rivals in all 39 cells, but the candidate-by-context interaction is negative in every cell and every 95% interval lies below zero. The sign persists under raw gain, residual matching, leave-one-out contexts, random orders, and permutation-averaged orders. Qwen3, Gemma4, and Qwen3.6 add 63 router-order cells, all with negative point estimates. To avoid bias from fitted geometric bases, we repeat the factorial in next-token loss. The interaction remains negative in OLMoE, Mixtral, and DeepSeek-MoE (37 of 39 cells, none with an interval above zero) and has a negative pooled macro in the three newer models.

The negative interaction changes the interpretation of multi-expert value rather than eliminating it. Experts that overlap in router-input geometry write more separated outputs than inputs, and in OLMoE and DeepSeek-MoE their local Jacobians on the same token are nearly orthogonal. Later selected experts reduce next-token NLL beyond load-near unselected experts, Top-2 beats Top-1 at matched active compute, and a complementarity regularizer lowers output overlap without improving validation loss. For compression, the practical consequence is immediate: skipping the least geometrically novel selected experts costs more NLL than skipping by router weight, and similarity-based skipping can be worse than random. Geometric overlap is how the router selects useful experts, not a certificate of redundant computation.

We call this pattern **coherent overlap**: routing is selective, selected experts are useful, and the selected set lives in an overlapping token-relevant neighborhood. Our contributions are:

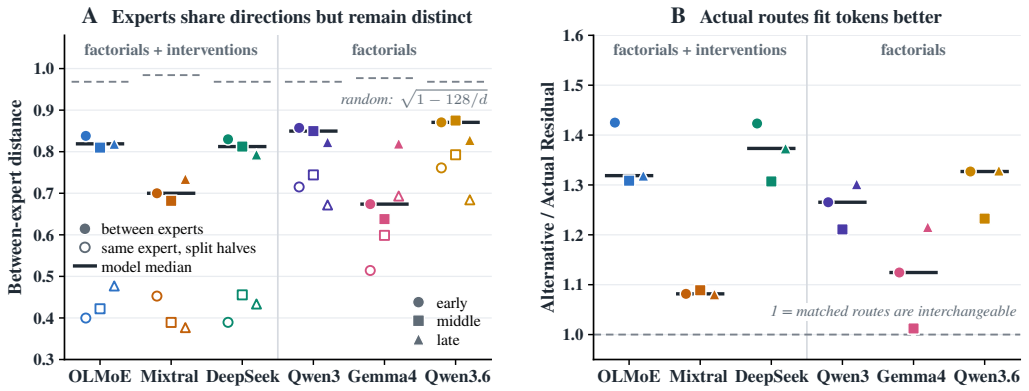

Figure 2: Shared geometry with selective routes. A: Mean between-expert distances of rank-128 routed-input subspaces (filled markers) lie far below random subspaces (dashed, $\sqrt{1-128/d}$ for model width d, 0.968–0.984) but above two fits of the same expert on disjoint data halves (open markers), showing shared directions and expert-specific structure. B: Alternative/actual residual ratios exceed one in all 18 cells (1.012–1.425, median 1.283), so actual routes better fit their tokens. Marker shape gives the layer, bars show model medians, and only support-qualified experts are included.

- **An interaction test for MoE complementarity.** A prefix-controlled 2×2 factorial separates route coherence, candidate fit, and candidate-by-context interaction, revealing why route counterfactuals that change candidate and context together can reverse the conclusion.
- **A cross-architecture account of multi-expert routing.** Across six MoEs, expert input subspaces share a common core while actual routes remain selective, and the geometric interaction is negative in all 102 cells, with order-free controls and the pooled loss-based factorial pointing the same way.
- **A functional separation between overlap and redundancy.** Experts from overlapping input neighborhoods can write distinct outputs and improve prediction, so pruning, merging, skipping, and diversity objectives should be evaluated by retained-route performance rather than geometric similarity alone.

2 Related Work

MoE work often treats a route as a small coalition of specialists, whether in analyses of router behavior, counterfactual tools that score routes or experts, or training and compression methods that operationalize diversity. We test the sharper implication: for a token and route position, should the co-selected context make the selected expert especially valuable?

What routers and experts represent. Analyses of pretrained MoEs have examined knowledge attribution (Wang et al., 2026a), part-of-speech and semantic routing (Antoine et al., 2025; Olson et al., 2025), expert diversity and co-activation (Lo et al., 2025; Tang et al., 2026), domain-invariant expert groups (Wang et al., 2026c), router-expert geometric coupling (Ahrac et al., 2026; Wang et al., 2026b), and routing-induced representation collapse (Chi et al., 2022). These studies establish measurable router and expert structure. We turn one implied mechanism into a sign test: if co-selected experts are complementary, the actual context should amplify a selected expert's advantage over a strong unselected rival.

Counterfactual routes and expert coalitions. Route-level counterfactuals compare selected routes with alternatives (Yoon et al., 2026), while coalition-based scores, including Shapley-style attributions, assign output-level value to experts over subsets and guide pruning (Huang et al., 2025; Zhang, 2026). These tools quantify route quality or expert importance. Our factorial adds the missing contrast by varying candidate and context separately: fixed contexts measure candidate fit, fixed

candidates measure context effects, and their difference-in-differences measures contextual complementarity.

Diversity, specialization, and compression. Training methods encourage diverse or orthogonal experts (Liu et al., 2024; Do et al., 2025; Park & Park, 2026) or explicitly pursue non-overlapping specialization (Dai et al., 2024). Compression methods prune, merge, or skip experts by frequency, similarity, or reconstruction (Lu et al., 2024; Zhang et al., 2025; Li et al., 2026; Hu et al., 2026; Lee et al., 2025; Chen et al., 2025), and adaptive routing varies the number of active experts per token (Zeng et al., 2024). These lines operationalize a common heuristic: geometric overlap suggests capacity that can be separated or removed. Coherent overlap tests that heuristic and shows why route-conditioned evaluation is necessary before using similarity as a proxy for redundancy.

3 Testing Contextual Complementarity

Complementarity is an interaction claim: beyond route coherence and single-expert strength, the actual co-selected context should leave a residual opportunity that the selected expert fills especially well. We define a residual-coverage score, cross candidates with contexts, and describe the layer-wide geometry in which these token-level interactions live.

3.1 Geometric Novelty

At layer ℓ, let $z_x^\ell \in \mathbb{R}^d$ denote the router-input representation for token x. The router selects an ordered set of k experts, $E_x^\ell = (e_1, \dots, e_k)$, in decreasing router-score order. We call e_1 the **leader** and $e_{2:k}$ the **later experts**. The score-ordered prefix before position j is $S_{j-1} = (e_1, \dots, e_{j-1})$. Standard Top-$k$ experts execute in parallel. The score order supplies a canonical family of contexts, and Section 5.3 tests order-free variants. We suppress the layer index below.

Each eligible routed expert e receives an uncentered rank-128 basis $B_e \in \mathbb{R}^{d\times 128}$, fitted on train-split router inputs with native Top-k weights renormalized within each token. For an expert set P, let $B(P) = \mathrm{orth}\big([B_e : e \in P]\big)$. The unexplained energy fraction of token x after the context P is

$$\rho_x(P) = \frac{\|(I - B(P)B(P)^\top)z_x\|_2^2}{\|z_x\|_2^2}, \tag{1}$$

and tokens with zero input energy are excluded. Adding candidate c yields its **fractional novelty**

$$q_x(c \mid P) = \frac{\rho_x(P) - \rho_x(P \cup \{c\})}{\max\big(\rho_x(P), \epsilon\big)}, \qquad \epsilon = 10^{-12}. \tag{2}$$

This is a residual-coverage score: $q = 0.2$ means that the candidate explains 20% of what remained after P. Under coherent overlap, a selected candidate can be well matched to the token yet have low novelty in its actual context, because co-selected experts have already covered shared token-relevant directions. Appendix C details the basis fit, eligibility, and masking.

3.2 A Candidate-by-Context Factorial

For a fixed token, layer, and route position j, the factorial compares two candidates in two kinds of context. The selected candidate is the actual rank-j expert $s = e_j$. The rival r is the highest-scoring eligible expert outside the complete actual Top-k route. The actual context is the preceding prefix $S = S_{j-1}$, and $W_1, \dots, W_M$ are $M = 5$ matched alternative contexts of the same length (Section 3.3). Crossing these choices yields the four cells in Figure 1.

Candidate fit asks whether the selected expert beats the rival under a fixed context:

$$\begin{aligned} A_{\text{actual}} &= q_x(s \mid S) - q_x(r \mid S), \\ A_{\text{alt}} &= \tfrac{1}{M}\textstyle\sum_m \big[q_x(s \mid W_m) - q_x(r \mid W_m)\big]. \end{aligned} \tag{3}$$

Context effects ask how the actual prefix changes the residual opportunity available to each fixed candidate:

$$\begin{aligned} T_s &= q_x(s \mid S) - \tfrac{1}{M}\textstyle\sum_m q_x(s \mid W_m), \\ T_r &= q_x(r \mid S) - \tfrac{1}{M}\textstyle\sum_m q_x(r \mid W_m). \end{aligned} \tag{4}$$

Table 1: Predicted signs. Route coherence uses the normalized residual ratio (Appendix B), and A_{actual} and D are defined in Eqs. 3 and 5.

Quantity	Measured by	Complementarity	Relevance-driven selection
Route coherence	residual ratio, alternative/actual	> 1	> 1
Candidate fit	A_{actual}	> 0	> 0
Candidate-by-context interaction	D	> 0	< 0

The interaction is the difference-in-differences

$$D = A_{\text{actual}} - A_{\text{alt}} = T_s - T_r. \tag{5}$$

This is the estimand for contextual complementarity. Positive D means that the actual context amplifies the selected candidate's advantage over the rival, whereas negative D means that it narrows that advantage. Thus both accounts in Table 1 allow coherent routes and strong selected candidates, but they make opposite predictions for the interaction. Because each basis is fitted on tokens routed to its expert, frequent partners can also contribute directions to that fitted span. The synthetic control in Appendix C.2 shows that this can reverse the sign even for complementary experts. Section 6 therefore repeats the same factorial in next-token loss, where no subspaces are fitted.

3.3 Matched Contexts and Ordering

Alternative contexts set the residual opportunity against which the actual context is judged. For each held-out token, we build five same-length alternative routes that share no expert with the actual route and exclude the rival. Each position uses an eligible expert among the eight nearest in routing load to the selected expert, so prefixes W_m match the load profile while changing expert identity. Geometric coverage depends only on the fitted spans, and in the functional factorial candidate and rival share the actual rank-j router weight (Appendix C).

The score order provides a canonical context sequence for a parallel Top-k layer. At the last position the context already contains every other selected expert, so the estimate is order-free. This includes every Mixtral cell. Section 5.3 extends the test to every rank with leave-one-out, random-order, and permutation-averaged variants.

3.4 Global Expert Separation

The factorial is token-local. To describe the global geometry in which these local interactions occur, each eligible expert e also receives a *centered* rank-p basis G_e ($p = 128$) fitted from its routed router inputs. We compare equal-rank spans with principal angles on the Grassmann manifold (Edelman et al., 1998; Ye & Lim, 2016), which directly measure separation between expert subspaces. For principal angles $\theta_1, \ldots, \theta_p$, the normalized chordal distance is

$$d_{\text{ch}}(U, V) = \left(\frac{1}{p} \sum_{i=1}^{p} \sin^2 \theta_i \right)^{1/2} = \frac{\|UU^\top - VV^\top\|_F}{\sqrt{2p}}. \tag{6}$$

Two random rank-p subspaces of $\mathbb{R}^d$ have $\mathbb{E}\, d_{\text{ch}}^2 = 1 - p/d$, so $d_{\text{ch}} \approx \sqrt{1 - p/d}$ (0.968–0.984 for our models). A hard partition of representation directions would place expert spans at or above this random-subspace scale. Refitting the same expert on two disjoint halves of its sequences gives the distance expected from sampling alone. Appendix B also measures how much of each expert subspace is captured by a layer-wide centered core.

3.5 Models and Protocol

We study six open MoEs (Table 2) spanning Top-k values, expert counts, and shared-expert designs: OLMoE (Top-8 of 64) and Mixtral (Top-2 of 8), both routed-only; DeepSeek-MoE (Top-6 of 64 routed experts plus two shared experts), which targets non-overlapping expert knowledge; and Qwen3, Gemma4, and Qwen3.6 (Top-8 of 128–256), the last two with a shared expert (Muennighoff

et al., 2025; Jiang et al., 2024; Dai et al., 2024; Yang et al., 2025; Gemma Team, 2026; Qwen Team, 2026). We analyze layers at one quarter, one half, and the final block (85% depth for the newer models), giving 18 model-layer cells for the global geometry survey.

The geometric factorial contains 39 position-resolved cells in OLMoE, Mixtral, and DeepSeek-MoE and 63 in the newer models. The functional factorial covers all six models, and the other pretrained-model experiments use the first three. Each geometric cell uses up to 2,048 held-out tokens per layer from a source-record split. Token metrics are averaged within source context, and 95% intervals use 1,000 paired bootstrap resamples (Efron & Tibshirani, 1993). For the newer models, 43 held-out document groups are resampled (Appendix E). Appendices C–J give construction details and cell-level estimates.

4 Shared Geometry, Selective Routes

Experts share a large layer-wide core, yet the router selects subsets that fit each token better than matched alternatives. Complementarity is therefore a contextual question rather than a question of whether experts overlap at all.

4.1 A Shared Core with Expert-Specific Directions

Between-expert distances range from 0.638 to 0.875 across the 18 model-layer cells (Figure 2A), well below the 0.968–0.984 expected from random rank-128 subspaces. Experts therefore occupy a common geometric neighborhood. A layer-wide core captures much of that neighborhood: 64 common directions give median overlap of 0.27–0.44, and 512 directions give 0.77–0.98 (Table 4).

The common core coexists with real expert-specific structure. Split-half refits of the same expert lie at 0.38–0.48 in OLMoE, Mixtral, and DeepSeek-MoE, below their 0.68–0.84 between-expert distances. In Qwen3, Gemma4, and Qwen3.6, whose surveys use 256 sequences, split-half distances are 0.51–0.79 and still below the corresponding between-expert distances (Appendix B.4).

4.2 The Router Selects Within the Shared Geometry

Shared geometry still leaves a clear routing signal. For each token, we compare the normalized residual of the actual route with that of a load-matched alternative route of the same size (Appendix B). The alternative-to-actual ratio exceeds one in all 18 cells, from 1.012 to 1.425 with median 1.283, and every 95% bootstrap interval lies above one (Figure 2B and Table 8).

The router's choice is also recoverable from geometry alone. Ranking experts by geometric residual identifies the router's leader for 47–50% of held-out tokens in OLMoE, 56–61% in DeepSeek-MoE (chance 1.6%), and 66–72% in Mixtral (chance 12.5%). Under a load-preserving route null, experts' routed inputs become less compact: rank-128 residuals rise by 12–15% in OLMoE, 13–23% in DeepSeek-MoE, and 3–10% in Mixtral in all ten permutations (Appendix B.5). Thus selected routes are coherent subsets of a shared neighborhood. Sections 5 and 6 test how those subsets interact.

5 Strong Candidates in Saturating Contexts

Coherent routes leave both accounts in Table 1 viable. The separating question is whether the actual prefix creates a residual opportunity that the selected candidate is especially suited to fill, or has already absorbed the directions that made the candidate attractive (contextual saturation). The factorial supports the second pattern.

5.1 Changing the Context Confounds Candidate Value

Route-level counterfactuals answer a useful practical question (what happens when a selected expert is replaced inside a different route), but they mix two changes that the complementarity claim keeps separate. In Figure 3A, each selected expert's marginal novelty is compared with five load-near, route-disjoint control routes. This joint replacement makes leaders look valuable and later experts look weak: all nine leader cells are positive, while all 39 later-expert cells are negative.

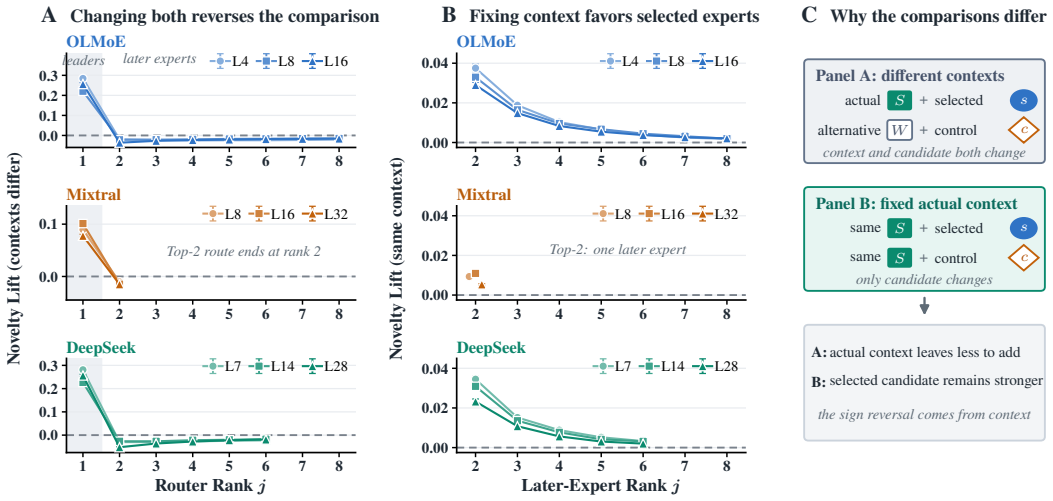

Figure 3: Joint counterfactuals confound candidate value. A: Replacing candidate and context together makes all nine leader cells positive and all 39 later-expert cells negative. B: With the actual prefix fixed, all 39 later-expert candidate cells favor the selected expert (mean lift 0.0018–0.0375). C: The reversal comes from changed residual opportunity. Error bars are 95% source-context bootstrap intervals.

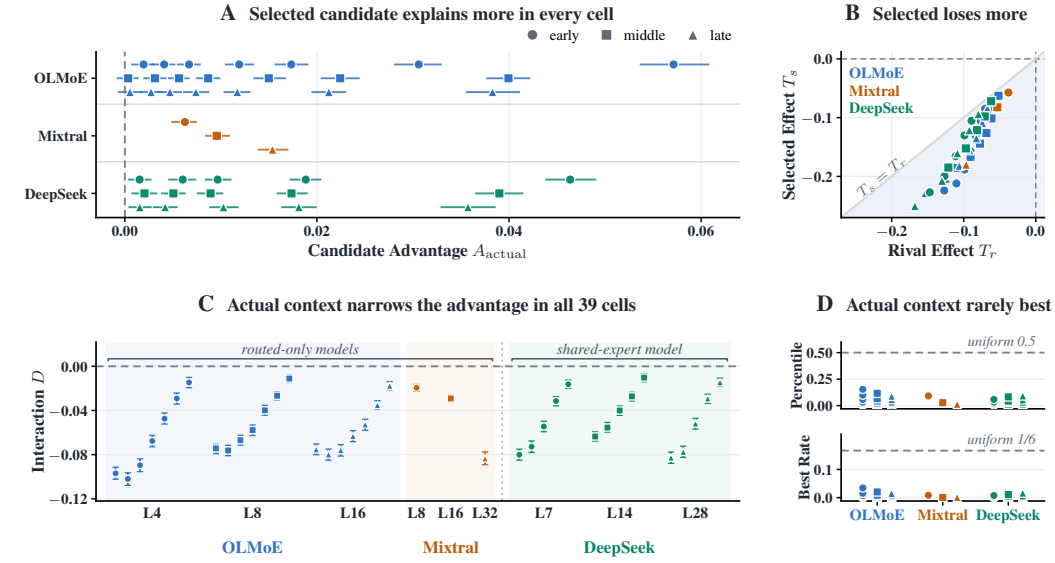

Figure 4: The factorial isolates contextual saturation. A: Candidate fit is positive in all 39 cells (37 intervals exclude zero). B: The actual prefix lowers residual opportunity for both candidates and lowers it more for the selected candidate (all cells below the diagonal). C: The interaction is negative in all 39 cells. D: Selected-candidate novelty ranks near the bottom across prefixes (percentiles 0.006–0.154). Error bars are 95% source-context bootstrap intervals, and dashed lines mark uniform references.

With the actual prefix fixed, the interpretation reverses. Replacing only the candidate favors the selected expert in all 39 later-expert cells, with mean lift from 0.0018 to 0.0375 and intervals above zero (Figure 3B). Later experts can therefore be strong candidates whose apparent marginal novelty collapses when the comparison also changes the residual opportunity.

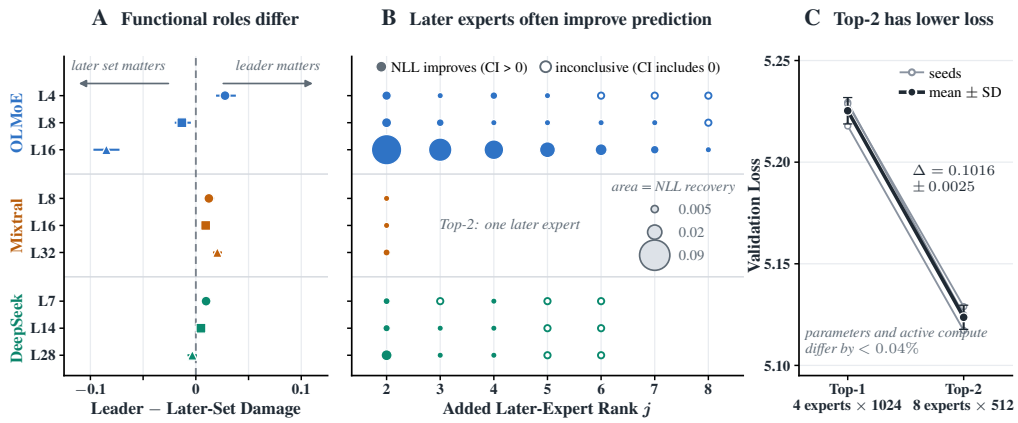

Figure 5: Functional value despite overlap. A: Leaders usually matter more, though aggregate later experts matter more in OLMoE layers 8 and 16. B: Later selected experts reduce next-token NLL in 36 of 39 additions (28 intervals above zero). C: At matched active capacity, Top-2 beats Top-1 in all seeds (bars: mean ± SD).

5.2 The Factorial Reveals Contextual Saturation

Under the actual prefix, the selected candidate beats the highest-scoring eligible rival in all 39 point estimates, and 37 intervals exclude zero (Figure 4A). The context contrast explains why route-level comparisons can reverse: actual prefixes leave less residual opportunity than load-matched alternatives for both fixed candidates, and the reduction is larger for the selected candidate (Figure 4B). The resulting interaction is negative in all 39 cells, with every 95% interval below zero (Figure 4C and Appendix D), and the macro is $-0.0523\,[-0.0538,\ -0.0506]$. The selected expert still wins, but by a smaller margin in the context selected with it, which is the signature of coherent overlap: earlier members of a token-relevant neighborhood absorb part of the same residual signal.

Qwen3, Gemma4, and Qwen3.6 (two with shared experts) add 63 router-order cells. The interaction is negative in every cell (56 intervals below zero, none above), with pooled macro $-0.0235\,[-0.0246,\ -0.0223]$. Candidate fit remains positive in 61 of 63 point estimates, and the leave-one-out version is negative in all 72 rank cells (Appendix E).

5.3 Order Robustness

The interaction survives removing router-score order from the analytic context (Appendix G). Leave-one-out conditioning is negative in all 48 rank cells, every interval below zero, with macro $-0.0560\,[-0.0575,\ -0.0545]$. Random context orders keep all 39 position cells negative, and permutation-averaging keeps all 48 rank cells negative, with every interval below zero in both variants. Magnitudes follow candidate router rank rather than analytic position (Tables 19 and 20).

5.4 Matched-Residual and Sensitivity Controls

A residual-matched control tests generic saturation. We replace the actual prefix with actual prefixes from other tokens that share no expert with the route or rival and match the same residual within 5%. Such prefixes exist for only 13.4% of token-cells. On that subset, unrelated prefixes reproduce much of the interaction, but the true co-selected context narrows the selected candidate's advantage further ($\Delta = -0.0057\,[-0.0065,\ -0.0049]$), with 24 of 39 cell intervals below zero and none above (Appendix H). Raw gain and nearest-residual variants keep all 39 intervals below zero, and a strict train-IQR caliper keeps every point estimate negative and 31 intervals below zero (Appendix F). The negative interaction thus holds across ordering, residual scale, context construction, and scoring.

6 Functional Value of Coherent Overlap

Coherent overlap gives the saturation of Section 5 a functional reading: experts from the same token-relevant neighborhood can still implement different nonlinear maps. We test route-specific predictive value in outputs, next-token loss, matched-compute training, and expert skipping.

First, expert maps transform shared input geometry into separated outputs. On the same held-out tokens (a smaller cache than in Section 4), mean between-expert distance rises from 0.76–0.88 for router inputs to 0.85–0.96 for unweighted outputs, and 99–100% of expert pairs are farther apart in output space (Appendix I). Within selected routes, projected-input cosines are 0.64–0.81, but output cosines fall to 0.02–0.15 in OLMoE and DeepSeek-MoE and 0.28–0.59 in Mixtral, and local Jacobians are nearly orthogonal in OLMoE and DeepSeek-MoE. Thus shared geometry helps the router find relevant experts while the expert MLPs write distinct residual-stream updates.

Second, later selected experts carry selection-specific predictive value. Adding the next selected expert to a frozen route reduces next-token NLL in 36 of 39 additions (28 with intervals above zero), with gains that shrink with rank. Truncating to the leader costs 0.002–0.231 NLL across models and layers, and with the leader fixed the actual second expert beats a load-near unselected expert in all nine model-layer cells by 0.0019–0.0483 (Appendix J).

The same 2×2 interaction appears in prediction, where no basis fit can drive the sign. After the actual prefix, the selected expert lowers NLL more than the rival in 35 of 39 point estimates, but the interaction is negative in 37 of 39 cells, with 25 intervals below zero and none above. Averaged over cells, the selected expert's gain drops from 0.0148 after alternative prefixes to 0.0082 after its actual prefix, while the rival barely changes. The macro interaction is $-0.0062\,[-0.0070,\ -0.0053]$. Qwen3, Gemma4, and Qwen3.6 show the same negative pooled direction, clear in Gemma4 and close to zero in the two Qwen models (Appendix J.4).

Third, training and compression experiments favor the routed committee over a novelty proxy. In a six-layer matched-compute MoE (Clark et al., 2022; Ludziejewski et al., 2024), Top-1 with four width-1,024 experts and Top-2 with eight width-512 experts have identical active intermediate capacity, with total parameters and active FLOPs differing by less than 0.04%. Top-2 wins in all three seeds by 0.1016 ± 0.0025 nats (Figure 5C). A direct output-complementarity regularizer changes the measured geometry but not predictive value: it reduces penalized overlap by 38–92%, while validation loss shows no consistent improvement across seeds. In the matched Top-2 model, the functional interaction is essentially zero at initialization, turns negative after one million training tokens, and remains negative at 20 million. The complementarity objective attenuates this interaction but does not make it positive (Appendices J.6, J.7, and J.9).

The inference-time skipping test reaches the same conclusion. Skipping the least novel selected experts raises NLL more than skipping by router weight in all three models, and skipping the expert most similar to its selected partners is worse than random in OLMoE and DeepSeek-MoE. At the token level, geometric novelty is nearly uncorrelated with the signed NLL change from adding an expert and only weakly related to the magnitude of that change (Appendix J.8). Route-conditioned loss, not geometric overlap, determines whether computation can be removed.

7 Discussion and Conclusion

This paper turns complementarity into a sign test. Multi-expert routing supports geometric complementarity only when the co-selected context increases a selected expert's advantage over a strong unselected rival. Across six MoEs, routes are coherent and selected experts are strong, yet the candidate-by-context interaction is negative in all 102 geometric cells. In OLMoE, Mixtral, and DeepSeek-MoE, the same sign persists under order-free contexts, residual-matched controls, sensitivity analyses, and the next-token loss factorial, where no fitted subspace can bias the result. The evidence points to coherent overlap: the router selects a token-relevant neighborhood, earlier selected experts cover part of the linear residual signal that makes later experts attractive, and later experts remain useful because their nonlinear maps write route-specific updates. Overlap is therefore a coordination structure for choosing useful experts, which makes the token-specific committee the right unit for counterfactual analysis, route-conditioned compression, and diversity objectives judged by prediction rather than geometry alone.

AI USE STATEMENT

Large language models were used for prose polishing, code-drafting assistance, and feedback on experimental design. The authors reviewed all AI-assisted work and remain responsible for the manuscript, references, code, analyses, and reported results.

ETHICS STATEMENT

This work analyzes publicly released MoE language models on public text and code corpora, and trains small MoE models on similar public sources. It involves no human subjects and no new data collection. Model and dataset licenses apply, and raw third-party records are not redistributed.

REPRODUCIBILITY STATEMENT

Section 3 defines the primary estimands, model coverage, and statistical protocol. Appendix A fixes notation, corpora, models, and analyzed layers. Appendices B–H document the subspace fits, route-coherence analysis, factorial construction, cell-level factorial results (including the three newer models), sensitivity analyses, order checks, and residual-matched controls. Appendix I compares input and output geometry, Jacobians, and logit contributions, and Appendix J reports the functional interventions, matched training runs, expert-skipping results, token-level comparisons of novelty and NLL change, and training dynamics. Appendix K summarizes the scope of the evidence, seeds, splits, and compute. The supplementary material provides code, configurations, model revisions, corpus manifests, compact numerical records, figure-generation scripts, and step-by-step instructions in `REPRODUCE.md`. Model checkpoints and raw third-party corpora are obtained from their upstream providers.

REFERENCES

Sagi Ahrac, Noya Hochwald, and Mor Geva. Routers learn the geometry of their experts: Geometric coupling in sparse mixture-of-experts. arXiv preprint arXiv:2605.12476, 2026. URL https://arxiv.org/abs/2605.12476.

Elie Antoine, Frédéric Béchet, and Philippe Langlais. Part-of-speech sensitivity of routers in mixture of experts models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6467–6474, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.431/.

Yuanteng Chen, Yuantian Shao, Peisong Wang, and Jian Cheng. EAC-MoE: Expert-selection aware compressor for mixture-of-experts large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12942–12963, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.633. URL https://aclanthology.org/2025.acl-long.633/.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34600–34613. Curran Associates, Inc., 2022. doi: 10.52202/068431-2507. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/df4f371f1f89ec8ba5014b3310578048-Paper-Conference.pdf.

Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew J. Johnson, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Marc'Aurelio Ranzato, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4057–4086. PMLR, July 2022. URL https://proceedings.mlr.press/v162/clark22a.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2097. URL https://aclanthology.org/N18-2097/.

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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70. URL https://aclanthology.org/2024.acl-long.70/.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

Giang Do, Hung Le, and Truyen Tran. SimSMoE: Toward efficient training mixture of experts via solving representational collapse. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2012–2025, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.107. URL https://aclanthology.org/2025.findings-naacl.107/.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954. URL https://doi.org/10.1137/S0895479895290954.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993. URL https://doi.org/10.1201/9780429246593.

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. URL http://jmlr.org/papers/v23/21-0998.html.

Gemma Team. Gemma 4 technical report. arXiv preprint arXiv:2607.02770, 2026. URL https://arxiv.org/abs/2607.02770.

Wentao Hu, Mingkuan Zhao, Shuangyong Song, Xiaoyan Zhu, Xin Lai, and Jiayin Wang. Mosaic Pruning: A hierarchical framework for generalizable pruning of mixture-of-experts models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21885–21893, 2026. URL https://doi.org/10.1609/aaai.v40i26.39341.

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*, volume 38, pp. 135973–136003, 2025. doi: 10.52202/085713-4538. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c66a9db149261435664284a20b6f1d42-Abstract-Conference.html.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. CodeSearchNet challenge: Evaluating the state of semantic code search. arXiv preprint arXiv:1909.09436, 2019. URL https://arxiv.org/abs/1909.09436.

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

Jaeseong Lee, Seung-won Hwang, Aurick Qiao, Daniel F. Campos, Zhewei Yao, and Yuxiong He. STUN: Structured-then-unstructured pruning for scalable MoE pruning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13660–13676, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.671. URL https://aclanthology.org/2025.acl-long.671/.

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. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Lujun Li, Qiyuan Zhu, Jiacheng Wang, Xiaoyu Qin, Wei Li, Hao Gu, Sirui Han, and Yike Guo. Sub-MoE: Efficient mixture-of-expert LLMs compression via subspace expert merging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22994–23002, 2026. URL https://doi.org/10.1609/aaai.v40i27.39464.

Boan Liu, Liang Ding, Li Shen, Keqin Peng, Yu Cao, Dazhao Cheng, and Dacheng Tao. Diversifying the mixture-of-experts representation for language models with orthogonal optimizer. In *ECAI 2024*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pp. 2966–2973. IOS Press, 2024. URL https://doi.org/10.3233/FAIA240836.

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, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.251. URL https://aclanthology.org/2025.findings-naacl.251/.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.334. URL https://aclanthology.org/2024.acl-long.334/.

Jan Ludziejewski, Jakub Krajewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, Marek Cygan, and Sebastian Jaszczur. Scaling laws for fine-grained mixture of experts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33270–33288. PMLR, July 2024. URL https://proceedings.mlr.press/v235/ludziejewski24a.html.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. Probing semantic routing in large mixture-of-expert models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18263–18278, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.991. URL https://aclanthology.org/2025.findings-emnlp.991/.

Sumin Park and Noseong Park. How many experts are enough? Towards optimal semantic specialization for mixture-of-experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24792–24800, 2026. URL https://doi.org/10.1609/aaai.v40i29.39665.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

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, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.104. URL https://aclanthology.org/2026.eacl-long.104/.

Bo Wang, Junzhuo Li, Hong Chen, Yuanlin Chu, Yuxuan Fan, and Xuming Hu. Deconstructing pre-training: Knowledge attribution analysis in MoE and dense models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33359–33367, 2026a. URL https://doi.org/10.1609/aaai.v40i39.40622.

Xi Wang, Soufiane Hayou, and Eric Nalisnick. The myth of expert specialization in MoEs: Why routing reflects geometry, not necessarily domain expertise. arXiv preprint arXiv:2604.09780, 2026b. URL https://arxiv.org/abs/2604.09780.

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, San Diego, California, United States, July 2026c. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.665. URL https://aclanthology.org/2026.acl-long.665/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016. doi: 10.1137/15M1054201. URL https://doi.org/10.1137/15M1054201.

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

Zihao Zeng, Yibo Miao, Hongcheng Gao, Hao Zhang, and Zhijie Deng. AdaMoE: Token-adaptive routing with null experts for mixture-of-experts language models. In *Findings of the Association*

for Computational Linguistics: EMNLP 2024, pp. 6223–6235, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.361. URL https://aclanthology.org/2024.findings-emnlp.361/.

Yuhao Zhang. SHAPE: Coalition-aware expert pruning for sparse mixture-of-experts LLMs. arXiv preprint arXiv:2606.09886, 2026. URL https://arxiv.org/abs/2606.09886.

Zeliang Zhang, Xiaodong Liu, Hao Cheng, Chenliang Xu, and Jianfeng Gao. Diversifying the expert knowledge for task-agnostic pruning in sparse mixture-of-experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 86–102, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.4. URL https://aclanthology.org/2025.findings-acl.4/.

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

A Appendix Roadmap, Notation, and Experimental Assets

This appendix follows the evidence chain of the main paper. It first fixes the notation, model coverage, data, and implementation details used throughout the experiments. It then expands the three central claims: that selected experts occupy a shared but selective geometry (Appendix B), that the candidate-by-context interaction is negative across geometric controls (Appendices C–H), and that overlapping experts remain functionally useful in outputs, loss, matched training, and expert-skipping tests (Appendices I and J).

We reserve *rival* for the highest-scoring eligible unselected routed expert in the factorial experiment. Figure 3 uses a different object, a load-near *control candidate*, because that diagnostic asks how jointly changing candidate and context can reverse the apparent value of later experts. Unless a table header states otherwise, a scalar written as $\hat{\theta}\,[l, u]$ denotes a point estimate followed by a 95% confidence interval. Mean $\pm$ standard deviation is used for the three-seed matched-training summaries. In all tables, L is the analyzed transformer layer index rather than the model depth.

The pretrained-model geometry, factorial, and intervention analyses use a fixed 8,192-record corpus containing 2,048 records each from C4, CodeSearchNet-Python, the `ccdv/arxiv-summarization` collection, and UltraChat 200k (Raffel et al., 2020; Husain et al., 2019; Cohan et al., 2018; Ding et al., 2023). The matched-compute study uses a separate 8,192-record corpus with the same general, code, and dialogue sources and an ARC science component comprising 1,024 ARC-Easy and 1,024 ARC-Challenge training examples (Clark et al., 2018). The supplementary material includes the construction manifest for the first corpus and fixed source revisions and content hashes for the second. Raw third-party records are not redistributed.

Analyzed layers follow a fixed depth rule. For a model with N transformer blocks, OLMoE, Mixtral, and DeepSeek-MoE use blocks $\mathrm{round}(N/4)$, $\mathrm{round}(N/2)$, and N, while Qwen3, Gemma4, and Qwen3.6 use $\mathrm{round}(N/4)$, $\mathrm{round}(N/2)$, and $\mathrm{round}(0.85N)$. Layer L denotes the router input of block L, counted from one.

One implementation detail is material for reproducing the DeepSeek-MoE runs. Under Transformers 5.5, the remote code of DeepSeek-MoE-16B leaves its rotary `inv_freq` buffers at zero after loading and its default tokenizer drops spaces. We therefore recompute `inv_freq` from the model configuration after loading and tokenize with the model's `tokenizer.json`.

Table 2: Models and analysis coverage. R = routed-only, S = shared expert, and H = shared expert with hybrid attention. All six models enter the geometry, geometric-factorial, and functional-factorial analyses. OLMoE, Mixtral, and DeepSeek-MoE also enter all other pretrained-model analyses. Factorial gives the analyzed route positions j.

Model	Design	Depth	Analyzed layers	Routed	Top-k	Shared	Factorial
OLMoE-1B-7B-0125	R	16	4, 8, 16	64	8	0	$j = 2{:}8$
Mixtral-8x7B-v0.1	R	32	8, 16, 32	8	2	0	$j = 2$
DeepSeek-MoE-16B-Base	S	28	7, 14, 28	64	6	2	$j = 2{:}6$
Qwen3-30B-A3B-Base	R	48	12, 24, 41	128	8	0	$j = 2{:}8$
Gemma4-26B-A4B	S	30	8, 15, 26	128	8	1	$j = 2{:}8$
Qwen3.6-35B-A3B	H	40	10, 20, 34	256	8	1	$j = 2{:}8$

B Global Geometry and Route Coherence

This section supports the first premise of coherent overlap. Expert subspaces are not arranged like independent random blocks but share a large layer-wide neighborhood. At the same time, the router's chosen subset consistently leaves less unexplained router-input energy than load-matched alternatives.

B.1 Fitting global and local routed-input subspaces

For the geometry survey, we use the native Top-k router weights, renormalized to sum to one within each token. An expert is eligible once it has at least 2,048 selected tokens. Its global basis G_e is the

Table 3: Geometry and route coherence across analyzed layers. $|\mathcal{E}|$ is the number of eligible experts and ESSI is defined in Eq. 8. Between is the mean inter-expert chordal distance, within is the load-weighted local tangent dispersion, and Sel. and Alt. are the normalized residuals of actual and matched alternative routes.

Model	L	$\lvert\mathcal{E}\rvert$	Between	Within	ESSI	Sel.	Alt.	Alt./Sel.
OLMoE-1B-7B	4	64	0.838	0.876	0.957	0.260	0.371	1.425
OLMoE-1B-7B	8	64	0.810	0.858	0.944	0.275	0.359	1.309
OLMoE-1B-7B	16	64	0.819	0.834	0.982	0.250	0.329	1.319
Mixtral-8x7B	8	8	0.700	0.894	0.783	0.736	0.796	1.082
Mixtral-8x7B	16	8	0.682	0.878	0.776	0.691	0.753	1.089
Mixtral-8x7B	32	8	0.734	0.853	0.861	0.525	0.567	1.081
DeepSeek-MoE-16B	7	64	0.830	0.844	0.984	0.315	0.448	1.423
DeepSeek-MoE-16B	14	64	0.812	0.853	0.952	0.344	0.449	1.307
DeepSeek-MoE-16B	28	64	0.793	0.817	0.971	0.252	0.346	1.373
Qwen3-30B-A3B	12	79	0.857	0.842	1.018	0.270	0.342	1.265
Qwen3-30B-A3B	24	82	0.850	0.836	1.016	0.289	0.350	1.211
Qwen3-30B-A3B	41	78	0.823	0.797	1.033	0.207	0.270	1.302
Gemma4-26B-A4B	8	90	0.674	0.715	0.942	0.009	0.011	1.124
Gemma4-26B-A4B	15	98	0.638	0.744	0.857	0.130	0.132	1.012
Gemma4-26B-A4B	26	83	0.820	0.803	1.021	0.233	0.283	1.216
Qwen3.6-35B-A3B	10	127	0.871	0.825	1.056	0.241	0.320	1.327
Qwen3.6-35B-A3B	20	131	0.875	0.825	1.060	0.278	0.342	1.232
Qwen3.6-35B-A3B	34	127	0.828	0.789	1.049	0.194	0.257	1.329

leading rank-128 basis of a centered, routing-weighted PCA fit. To compare this global span with local variation inside the same expert, we sample anchors and candidate tokens without replacement in proportion to the expert's routing weights. At anchor a, we center the 256 nearest representations in the candidate pool and fit a local rank-128 tangent basis $L_{e,a}$. Distances use the normalized projection/chordal metric in Eq. 6, a standard Grassmannian metric (Edelman et al., 1998; Ye & Lim, 2016).

OLMoE, Mixtral, and DeepSeek-MoE use all 2,048 corpus sequences, 2,048 anchors, and up to 8,192 candidate tokens per eligible expert. Qwen3, Gemma4, and Qwen3.6 use 256 sequences of 512 tokens from the same corpus, with 512 anchors and up to 4,096 candidate tokens. Global and local bases are fitted on the same routed tokens. The base random seed is zero, with the layer index added for layer-specific sampling.

B.2 Route coherence within the shared geometry

Route coherence asks whether the selected set captures its token better than a matched set of equally many routed experts. The residual uses this section's centered basis family, refit per run with the same construction as G_e: for a routed set E with $U_E = \mathrm{orth}([G_e : e \in E])$ and μ_E the route-weight-averaged mean of its experts' means (alternative sets reuse the actual route's weights), the normalized residual of router input z is $\|(I - U_E U_E^\top)(z - \mu_E)\|_2^2 / \max(\|z - \mu_E\|_2^2, 10^{-12})$. This centered survey quantity is separate from the uncentered factorial residual ρ in Eq. 1.

OLMoE, Mixtral, and DeepSeek-MoE split the 2,048 packed sequences 80/20 at random, fit bases on the training sequences, and average token-level residuals over 50,000 evaluation tokens with one load-matched alternative set per token. Each compared set is restricted to fitted experts. Qwen3, Gemma4, and Qwen3.6 use a 1,024-token support threshold, keep only tokens whose complete routed set is fitted, draw five alternative sets per token, and average within source context before the cell mean. Table 8 gives bootstrap intervals for all 18 cells, and Table 3 records the mean normalized residuals and their ratio (Sel., Alt., and Alt./Sel.), corresponding to Figure 2B. Ratios above one mean that the alternative leaves more unexplained router-input energy. Shared experts, where present, are held fixed outside the routed-set comparison.

Table 4: Median overlap between expert subspaces and a centered layer-wide core.

Model	$c = 64$	$c = 128$	$c = 256$	$c = 512$
OLMoE-1B-7B	0.320	0.501	0.687	0.851
Mixtral-8x7B	0.435	0.695	0.879	0.977
DeepSeek-MoE-16B	0.335	0.521	0.704	0.857
Qwen3-30B-A3B	0.292	0.474	0.661	0.818
Gemma4-26B-A4B	0.403	0.658	0.839	0.937
Qwen3.6-35B-A3B	0.269	0.422	0.594	0.766

B.3 Shared-core overlap

As a complementary overlap diagnostic, let $U_0^{(c)}$ be the leading c directions of a centered PCA fit over all sampled router inputs in a layer. For an eligible rank-p expert basis G_e ($p = 128$), we compute

$$\mathrm{CoreOverlap}_c(e) = \frac{1}{p}\left\|U_0^{(c)\top} G_e\right\|_F^2. \tag{7}$$

This is the fraction of the expert subspace captured by the shared layer-wide core. Table 4 reports the median over all eligible expert-layer bases in each model's three analyzed layers.

B.4 Split-half references and local dispersion

Table 5 gives, for every cell, the between-expert distance and the distance between two global bases of the same expert fitted on disjoint random halves of its sequences (mean over experts). The survey also computes a ratio of the between-expert distance to a local within-expert dispersion, the Expert Subspace Separation Index,

$$\mathrm{ESSI} = \frac{\mathrm{mean}_{e<f}\, d_{\mathrm{ch}}(G_e, G_f)}{\max\{\sum_e \lambda_e \,\mathrm{mean}_{a\in\mathcal{A}_e}\, d_{\mathrm{ch}}(G_e, L_{e,a}),\, \epsilon\}}, \tag{8}$$

where λ_e is the routing-load share of expert e, $\mathcal{A}_e$ its anchors, and $L_{e,a}$ the rank-128 local tangent basis of Appendix B.1. The raw ratio lies between 0.776 and 1.060 (Table 3). Its denominator, however, is dominated by estimation noise at this rank: two local bases fitted at the same anchor from disjoint halves of its candidate pool differ by about as much as one such half-pool local basis differs from its expert's global basis (ratio 0.97–1.10 in every cell), because local spectra are nearly flat around rank 128. Removing this noise with three descriptive corrections, which subtract it in squared distance, correct for attenuation of the overlap, or additionally subtract the noise of the global bases, gives ESSI of 1.01–1.63, with OLMoE between 1.14 and 1.46. We therefore describe global geometry with between-expert distances and their references rather than with the ratio. Table 6 varies the subspace rank and Table 7 removes the leading c directions of a layer-wide core, both with the uncorrected ratio. The rank table uses raw Top-k weights with routing-mass eligibility, and at $p = 128$ its values differ from Table 3 by at most 0.007 in OLMoE and 0.011 in DeepSeek-MoE, and not at all in Mixtral. Every mean between-expert distance lies below the random-subspace value $\sqrt{1 - p/d}$: 0.968 for $d = 2048$ (OLMoE, DeepSeek-MoE, Qwen3, Qwen3.6), 0.977 for $d = 2816$ (Gemma4), and 0.984 for $d = 4096$ (Mixtral).

B.5 Route-coherence intervals and routing nulls

Table 8 reports bootstrap intervals of the route-coherence ratio in all 18 cells, together with the fraction of tokens whose actual route leaves less residual than its alternative (the mean of five alternatives in the three newer models). For OLMoE, Mixtral, and DeepSeek-MoE, the resampled units are connected components of held-out sequences that share a source record, and the ratio is token-weighted as in Table 3. For the three newer models, the units are held-out sequences. Every interval lies above one. Gemma4 layer 15 has the smallest effect: its ratio exceeds one with an interval of $[1.007, 1.018]$, and 58% of its tokens favor the actual route.

Two further analyses in OLMoE, Mixtral, and DeepSeek-MoE show that routing is tied to router-input geometry. Table 9 ranks eligible experts by the rank-128 residual of a held-out token under

Table 5: Reference distances for global and local subspace estimates (rank 128, 256 neighbors). Between is the mean between-expert distance, and split-half refits the same expert on disjoint sequence halves. Within compares local bases fitted from half of each candidate pool with the expert's global basis, noise floor compares two local bases fitted at the same anchor from disjoint halves of that pool, and Ratio is noise floor over within. ESSI is the uncorrected full-pool ratio of Table 3, and Corrected gives the range of three noise-corrected values.

Model	L	Between	Split-half	Within	Noise floor	Ratio	ESSI	Corrected
OLMoE	4	0.838	0.400	0.862	0.838	0.97	0.957	1.14–1.34
OLMoE	8	0.810	0.422	0.847	0.875	1.03	0.944	1.26–1.40
OLMoE	16	0.819	0.477	0.822	0.847	1.03	0.982	1.31–1.46
Mixtral	8	0.700	0.453	0.884	0.909	1.03	0.783	1.01–1.15
Mixtral	16	0.682	0.389	0.866	0.911	1.05	0.776	1.08–1.18
Mixtral	32	0.734	0.377	0.845	0.873	1.03	0.861	1.14–1.27
DeepSeek	7	0.830	0.389	0.831	0.842	1.01	0.984	1.27–1.43
DeepSeek	14	0.812	0.456	0.843	0.865	1.03	0.952	1.26–1.40
DeepSeek	28	0.793	0.434	0.804	0.824	1.02	0.971	1.28–1.43
Qwen3	12	0.857	0.715	0.830	0.867	1.05	1.018	1.23–1.54
Qwen3	24	0.850	0.744	0.827	0.878	1.06	1.016	1.22–1.55
Qwen3	41	0.823	0.672	0.781	0.824	1.06	1.033	1.29–1.58
Gemma4	8	0.674	0.514	0.703	0.742	1.06	0.942	1.20–1.44
Gemma4	15	0.638	0.599	0.735	0.812	1.10	0.857	1.03–1.39
Gemma4	26	0.820	0.693	0.790	0.843	1.07	1.021	1.26–1.58
Qwen3.6	10	0.871	0.761	0.810	0.844	1.04	1.056	1.25–1.59
Qwen3.6	20	0.875	0.793	0.812	0.859	1.06	1.060	1.24–1.63
Qwen3.6	34	0.828	0.684	0.773	0.822	1.06	1.049	1.32–1.63

Table 6: ESSI across subspace ranks. Each cell reports between / within / ESSI.

Model	L	$p = 32$	$p = 64$	$p = 128$
OLMoE	4	0.876 / 0.886 / 0.988	0.862 / 0.884 / 0.976	0.845 / 0.879 / 0.962
OLMoE	8	0.843 / 0.855 / 0.987	0.830 / 0.859 / 0.966	0.815 / 0.859 / 0.949
OLMoE	16	0.849 / 0.818 / 1.039	0.839 / 0.829 / 1.012	0.826 / 0.836 / 0.988
Mixtral	8	0.668 / 0.879 / 0.760	0.686 / 0.890 / 0.770	0.700 / 0.894 / 0.783
Mixtral	16	0.658 / 0.842 / 0.781	0.670 / 0.862 / 0.777	0.682 / 0.878 / 0.776
Mixtral	32	0.709 / 0.802 / 0.884	0.718 / 0.828 / 0.866	0.734 / 0.853 / 0.861
DeepSeek	7	0.887 / 0.855 / 1.037	0.864 / 0.850 / 1.016	0.836 / 0.845 / 0.990
DeepSeek	14	0.861 / 0.853 / 1.010	0.844 / 0.856 / 0.986	0.822 / 0.854 / 0.963
DeepSeek	28	0.859 / 0.825 / 1.041	0.837 / 0.823 / 1.017	0.802 / 0.818 / 0.980

each expert's basis (bases fitted on 80% of the contexts, 100,000 held-out tokens) and compares the ranking with the router's choice. Table 10 compares each expert's rank-128 residual under the actual assignment with a null that permutes complete routes across tokens, which keeps every token's Top-k weights and every expert's load. The column "Matched ratio" permutes only among tokens of the same domain, frequency bin, and position bin. The permutation null uses row-normalized Top-k weights and is computed in-sample over all 1,048,576 corpus tokens with ten permutations.

C Factorial Construction and Geometric Controls

The main text uses a factorial because complementarity is a contextual claim. The construction below keeps the token, router rank, and candidate scale fixed while varying candidate identity and context separately. This section also records controls that clarify what the geometric interaction can and cannot establish on its own.

C.1 Uncentered bases for residual coverage

For the factorial, each routed expert with at least 1,024 train-split tokens receives an uncentered rank-128 router-input basis B_e, fitted with row-normalized native Top-k assignment weights. For

Table 7: ESSI after removing the leading c directions of a layer-wide centered PCA core (rank 128).

Model	L	$c = 64$	$c = 128$	$c = 256$	$c = 512$
OLMoE	4	0.980	0.992	1.001	1.005
OLMoE	8	0.969	0.982	0.992	0.995
OLMoE	16	1.011	1.017	1.018	1.017
Mixtral	8	0.863	0.906	0.947	0.973
Mixtral	16	0.851	0.895	0.937	0.970
Mixtral	32	0.920	0.950	0.976	0.994
DeepSeek	7	1.003	1.014	1.021	1.023
DeepSeek	14	0.985	1.000	1.008	1.011
DeepSeek	28	1.000	1.009	1.016	1.020

Table 8: Route coherence of the selected committee. Values are alternative over actual residuals with 95% bootstrap intervals over resampled units (1,000 resamples). Token share is the fraction of tokens for which the actual route has smaller residual.

Model	L	Units	Ratio [95% CI]	Token share
OLMoE-1B-7B	4	292	$1.425\,[1.416,\ 1.436]$	0.998
OLMoE-1B-7B	8	279	$1.309\,[1.298,\ 1.322]$	0.999
OLMoE-1B-7B	16	290	$1.319\,[1.309,\ 1.329]$	0.996
Mixtral-8x7B	8	277	$1.082\,[1.079,\ 1.085]$	0.942
Mixtral-8x7B	16	278	$1.089\,[1.086,\ 1.092]$	0.967
Mixtral-8x7B	32	281	$1.081\,[1.076,\ 1.087]$	0.674
DeepSeek-MoE-16B	7	291	$1.423\,[1.414,\ 1.434]$	0.997
DeepSeek-MoE-16B	14	284	$1.307\,[1.299,\ 1.316]$	0.995
DeepSeek-MoE-16B	28	277	$1.373\,[1.363,\ 1.385]$	0.983
Qwen3-30B-A3B	12	51	$1.265\,[1.255,\ 1.276]$	0.998
Qwen3-30B-A3B	24	51	$1.211\,[1.200,\ 1.222]$	0.998
Qwen3-30B-A3B	41	51	$1.302\,[1.290,\ 1.314]$	0.998
Gemma4-26B-A4B	8	51	$1.124\,[1.105,\ 1.145]$	0.722
Gemma4-26B-A4B	15	51	$1.012\,[1.007,\ 1.018]$	0.576
Gemma4-26B-A4B	26	51	$1.216\,[1.202,\ 1.231]$	0.973
Qwen3.6-35B-A3B	10	51	$1.327\,[1.310,\ 1.348]$	0.996
Qwen3.6-35B-A3B	20	50	$1.232\,[1.217,\ 1.248]$	0.999
Qwen3.6-35B-A3B	34	51	$1.329\,[1.308,\ 1.353]$	0.999

a context P, we form the orthonormal union $B(P)$ and divide residual energy by raw input energy. Rows with exactly zero input energy are masked as invalid, and fractional novelty uses $\max(\rho_x(P), 10^{-12})$ as its numerical floor.

The train/evaluation split is by connected components of source-record contexts, with an 80/20 fraction and seed zero. Each layer evaluates up to 2,048 held-out tokens (at least 1,024), and up to 2,048 training tokens calibrate the residual matching of Appendix F. The coverage score depends only on the spans of the fitted bases, so the geometric factorial changes expert identity and not mixture scale. Router weights enter through the basis fits.

C.2 Synthetic control for fitted-basis bias

A geometric basis is learned from the tokens routed to its expert. Those tokens also contain directions supplied by frequent co-selected partners, so a fitted basis can make a genuinely complementary route look more saturated than the true expert subspaces would. To measure this effect, we apply the full factorial of the newer models (record split, fitted rank-128 bases, router order, rival, five load-near alternatives, and held-out-unit bootstrap) to synthetic router inputs whose true expert subspaces are known.

The synthetic model has 64 experts in $d = 2048$ with Top-8 routing. Experts form four groups of 16 on a ring. Each owns a 128-dimensional subspace, and one parameter moves the subspaces within a group from mutually orthogonal blocks to a shared neighborhood. The resulting mean

Table 9: Router choice recovered from geometry. Top-1 is the fraction of held-out tokens whose router leader has the smallest residual, and Top-k recall is the fraction of the router's selected set recovered by the k smallest-residual experts. Chance levels are shown for both.

Model	L	Top-1	Chance	Top-k recall	Chance
OLMoE	4	0.499	0.016	0.748	0.125
OLMoE	8	0.465	0.016	0.773	0.125
OLMoE	16	0.495	0.016	0.821	0.125
Mixtral	8	0.658	0.125	0.717	0.250
Mixtral	16	0.720	0.125	0.780	0.250
Mixtral	32	0.714	0.125	0.858	0.250
DeepSeek	7	0.594	0.016	0.756	0.094
DeepSeek	14	0.565	0.016	0.722	0.094
DeepSeek	28	0.615	0.016	0.732	0.094

Table 10: Load-weighted rank-128 residual under actual routes and under route permutations that preserve Top-k weights and expert loads. Ratio is permuted over actual, Matched ratio permutes only among tokens of the same domain, frequency bin, and position bin, and Wins counts the permutations whose residual exceeds that of the actual assignment.

Model	L	Actual	Permuted	Ratio	Matched ratio	Wins
OLMoE	4	0.637	0.732	1.149	1.110	10/10
OLMoE	8	0.634	0.708	1.117	1.090	10/10
OLMoE	16	0.560	0.644	1.149	1.124	10/10
Mixtral	8	0.753	0.777	1.032	1.029	10/10
Mixtral	16	0.698	0.725	1.040	1.035	10/10
Mixtral	32	0.262	0.288	1.098	1.088	10/10
DeepSeek	7	0.602	0.710	1.179	1.139	10/10
DeepSeek	14	0.628	0.713	1.135	1.113	10/10
DeepSeek	28	0.480	0.592	1.232	1.195	10/10

principal cosine β between true co-selected subspaces is 0, 0.23, 0.51, 0.75, or 0.86. Each token mixes the subspaces of nearby experts with graded strengths and isotropic noise, and a linear router with fixed noise selects 92% of the eight strongest experts at every β. Each setting has three independent replicates of 131,072 tokens (21 router-order cells and 24 cells each for leave-one-out and permutation-averaged contexts).

Figure 6A compares D computed from fitted bases with D computed from the true subspaces on the same tokens, rivals, and alternatives. With true subspaces, D is positive at $\beta \leq 0.23$ (router order $+0.063$ and $+0.043$, leave-one-out $+0.180$ and $+0.141$) and negative at $\beta \geq 0.75$. Fitted bases lower D by 0.07–0.17 at $\beta \leq 0.23$: router order gives $-0.003\,[-0.005,\ -0.001]$ and -0.043, leave-one-out gives $+0.027$ and -0.027, and at $\beta = 0.23$ every cell interval lies below zero. The fitted bases of complementary experts remain nearly as far apart as random subspaces (at $\beta = 0$, mean distance 0.963 over all pairs and 0.931 over co-selected pairs, against 0.968 for random subspaces and 0.638–0.875 in the six models), so the fit depresses the interaction far more than it creates the empirical overlap.

This control makes the role of the loss-based factorial explicit. The fitted-basis shift in this control (0.07–0.17) exceeds the magnitude of the router-order geometric macros of the six real models (-0.023 to -0.057), so a negative geometric D alone cannot exclude partly complementary experts. The next-token-loss factorial in Appendix J.4 therefore supplies the sign test that does not rely on fitted subspaces: it keeps the same candidate-by-context contrast while removing fitted subspaces entirely. As an additional check, on the first 256 sequences of OLMoE layer 16, processed like the newer models, replacing native routes by routes that greedily maximize geometric novelty raises the fitted-basis D from -0.064 to -0.032 (Figure 6B), showing that even a router optimized for geometric novelty can retain a negative fitted-basis interaction.

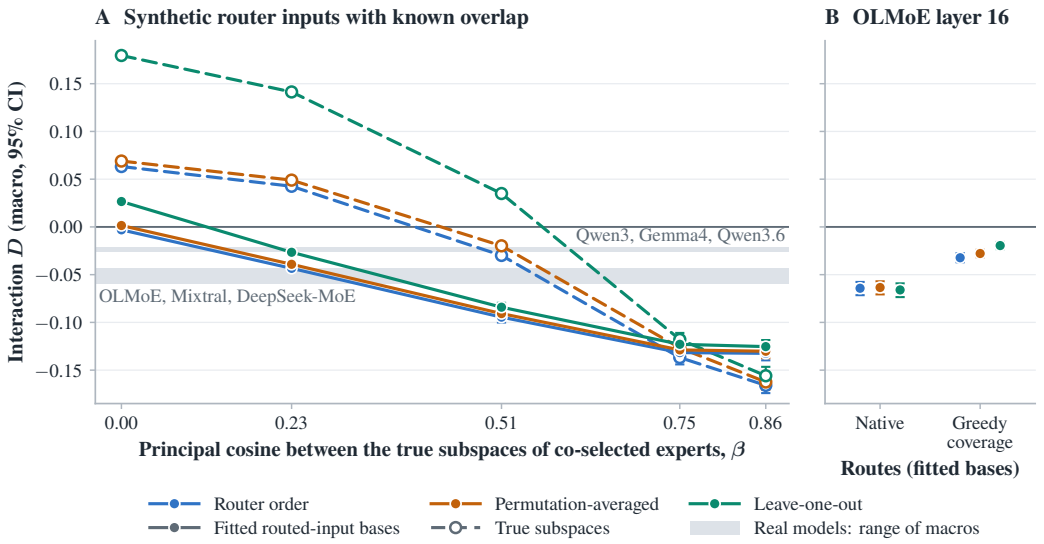

Figure 6: Synthetic control for the geometric interaction. A: Macro D with 95% intervals versus true co-selected-subspace overlap β, using fitted routed-input bases (solid) and true subspaces (dashed) for router-order, permutation-averaged, and leave-one-out contexts. Gray bands span the per-model router-order and leave-one-out macros of OLMoE, Mixtral, and DeepSeek-MoE (lower band) and of Qwen3, Gemma4, and Qwen3.6 (upper band). B: OLMoE layer 16 under native routes and greedy-coverage routes chosen to maximize geometric novelty (fitted bases).

C.3 Load-near control in Figure 3

Figure 3 uses a load-near control rather than the score-defined rival of the factorial. For each held-out token, we construct complete alternative routes disjoint from the actual route. At each position we order eligible experts by absolute routing-load difference from the selected expert, keep the nearest eight, and draw without replacement using a fixed seed.

The changing-context quantity in Figure 3A is

$$\delta_j^{\text{ind}} = q(s_j \mid S_{j-1}) - \frac{1}{5}\sum_{m=1}^{5} q(c_{m,j} \mid C_{m,j-1}), \tag{9}$$

where each C_m is a complete load-near, route-disjoint control route. Thus both context and candidate change. The corresponding same-prefix quantity in Figure 3B is

$$\delta_j^{\text{same}} = q(s_j \mid S_{j-1}) - q(c_j \mid S_{j-1}), \qquad j \geq 2, \tag{10}$$

where c_j is taken from the first complete route in a deterministic 32-route control pool. Both candidates are evaluated with the selected expert's actual rank-j weight. This c_j is a load-near control candidate, whereas the factorial analysis uses the score-defined rival r.

Across the changing-context diagnostic (Table 11), all nine leader intervals are above zero and all 39 later-expert intervals are below zero. Holding the actual context fixed reverses the conclusion: all 39 intervals are above zero, with point estimates from 0.0018 to 0.0375, and every one of the 79,872 later token-cells yields an estimate. This reversal is the practical reason for the factorial: candidate fit and residual opportunity must be separated before interpreting later selected experts.

C.4 Rivals and alternative contexts

The rival is fixed from the complete router-score ordering before any context substitution: it is the highest-scoring eligible routed expert outside the entire actual Top-k route. OLMoE and DeepSeek-MoE use $M = 5$ deterministic, position-wise load-near, unique alternative full routes, whose equal-length prefixes serve as the alternative contexts $W_1, \ldots, W_M$ of Section 3.2. They exclude the

Table 11: Layer-level load-near control for Figure 3. The leader column gives the changing-context difference δ_1^{ind} with its 95% interval. For the later positions $j \geq 2$, each block gives the number of cells with the stated sign and the smallest and largest point estimate of δ_j^{ind} (changing context) or δ_j^{same} (same prefix). Every later token-cell yields an estimate (14,336 per OLMoE layer, 2,048 per Mixtral layer, and 10,240 per DeepSeek-MoE layer).

			Later, changing context			Later, same prefix		
Model	L	Leader δ_1^{ind} [95% CI]	Neg.	Min	Max	Pos.	Min	Max
OLMoE	4	$0.2845\,[0.2760,\ 0.2932]$	7/7	-0.0250	-0.0193	7/7	0.0018	0.0375
	8	$0.2203\,[0.2137,\ 0.2268]$	7/7	-0.0215	-0.0149	7/7	0.0020	0.0329
	16	$0.2564\,[0.2489,\ 0.2639]$	7/7	-0.0362	-0.0159	7/7	0.0021	0.0290
Mixtral	8	$0.0897\,[0.0843,\ 0.0958]$	1/1	-0.0105		1/1	0.0093	
	16	$0.1011\,[0.0977,\ 0.1047]$	1/1	-0.0147		1/1	0.0109	
	32	$0.0777\,[0.0750,\ 0.0807]$	1/1	-0.0150		1/1	0.0054	
DeepSeek-MoE	7	$0.2810\,[0.2729,\ 0.2886]$	5/5	-0.0301	-0.0223	5/5	0.0033	0.0345
	14	$0.2266\,[0.2209,\ 0.2329]$	5/5	-0.0265	-0.0173	5/5	0.0031	0.0309
	28	$0.2563\,[0.2502,\ 0.2618]$	5/5	-0.0520	-0.0189	5/5	0.0020	0.0232

complete actual route and the rival, use a neighbor window of eight, and reuse the same full-route prefixes across $j = 2, \ldots, k$. Mixtral has only one preceding expert at $j = 2$ and uses all five eligible alternative singleton leaders. The two shared experts of DeepSeek-MoE are always active and never enter candidates, rivals, contexts, or alternative routes.

Token-level quantities are averaged within source context. Confidence intervals use 1,000 paired bootstrap resamples over source contexts with seed zero (Efron & Tibshirani, 1993). They quantify held-out context-sampling uncertainty for the analyzed checkpoints.

D Complete Geometric Factorial Results

Table 12 gives every reported cell. A_{actual} is the selected-minus-rival advantage under the actual context, T_s and T_r are the selected and rival context effects, and $D = T_s - T_r$ is the candidate-by-context interaction. Percentile is the fraction of the five alternative-context selected-candidate gains that are no greater than the actual-context gain. Every interval is a paired 95% source-context bootstrap interval.

All 39 D intervals in OLMoE, Mixtral, and DeepSeek-MoE lie below zero. The equal-cell macro, computed with equal cell weights within each source context, is $-0.0523\,[-0.0538,\ -0.0506]$. The per-model macros are $-0.0571\,[-0.0596,\ -0.0548]$ for OLMoE, $-0.0442\,[-0.0465,\ -0.0417]$ for Mixtral, and $-0.0471\,[-0.0491,\ -0.0449]$ for DeepSeek-MoE. The 39 per-cell intervals are nominal 95% intervals without multiplicity correction.

The point estimate of A_{actual} is positive in all 39 cells, and 37 of its intervals exclude zero.

Table 12: Geometric factorial cells for OLMoE, Mixtral, and DeepSeek-MoE (2,048 valid held-out tokens per cell). A_{actual} and D are shown with paired 95% source-context bootstrap intervals, and T_s and T_r as point estimates (their intervals are listed in the supplementary result files). Pctl. is the actual-prefix percentile of the selected candidate.

L	j	A_{actual} [95% CI]	T_s	T_r	D [95% CI]	Pctl.
OLMoE						
4	2	$0.0571\,[0.0536,\ 0.0608]$	-0.2239	-0.1269	$-0.0970\,[-0.1023,\ -0.0914]$	0.006
4	3	$0.0306\,[0.0280,\ 0.0330]$	-0.2120	-0.1101	$-0.1020\,[-0.1078,\ -0.0964]$	0.014
4	4	$0.0173\,[0.0156,\ 0.0191]$	-0.1883	-0.0987	$-0.0896\,[-0.0954,\ -0.0839]$	0.019
4	5	$0.0119\,[0.0104,\ 0.0134]$	-0.1578	-0.0901	$-0.0677\,[-0.0729,\ -0.0625]$	0.033
4	6	$0.0067\,[0.0054,\ 0.0079]$	-0.1307	-0.0834	$-0.0473\,[-0.0523,\ -0.0423]$	0.058
4	7	$0.0041\,[0.0028,\ 0.0053]$	-0.1055	-0.0764	$-0.0292\,[-0.0343,\ -0.0242]$	0.099

Continued on next page

Table 12 (continued)

L	j	A_{actual} [95% CI]	T_s	T_r	D [95% CI]	Pctl.
4	8	0.0020 [0.0007, 0.0031]	−0.0848	−0.0702	−0.0146 [−0.0194, −0.0100]	0.154
8	2	0.0399 [0.0376, 0.0422]	−0.1856	−0.1112	−0.0743 [−0.0790, −0.0706]	0.011
8	3	0.0224 [0.0204, 0.0245]	−0.1667	−0.0905	−0.0762 [−0.0810, −0.0716]	0.011
8	4	0.0150 [0.0134, 0.0168]	−0.1442	−0.0773	−0.0669 [−0.0714, −0.0625]	0.017
8	5	0.0087 [0.0074, 0.0099]	−0.1262	−0.0684	−0.0577 [−0.0627, −0.0531]	0.025
8	6	0.0056 [0.0045, 0.0067]	−0.1013	−0.0614	−0.0399 [−0.0442, −0.0353]	0.043
8	7	0.0031 [0.0019, 0.0043]	−0.0829	−0.0562	−0.0266 [−0.0304, −0.0230]	0.067
8	8	0.0003 [−0.0009, 0.0014]	−0.0628	−0.0517	−0.0111 [−0.0146, −0.0078]	0.116
16	2	0.0383 [0.0355, 0.0412]	−0.2284	−0.1531	−0.0753 [−0.0801, −0.0703]	0.006
16	3	0.0212 [0.0194, 0.0230]	−0.2040	−0.1242	−0.0798 [−0.0851, −0.0749]	0.007
16	4	0.0117 [0.0103, 0.0131]	−0.1820	−0.1060	−0.0760 [−0.0813, −0.0709]	0.009
16	5	0.0074 [0.0061, 0.0088]	−0.1552	−0.0920	−0.0632 [−0.0683, −0.0582]	0.022
16	6	0.0047 [0.0034, 0.0059]	−0.1350	−0.0823	−0.0527 [−0.0579, −0.0479]	0.033
16	7	0.0027 [0.0015, 0.0040]	−0.1092	−0.0741	−0.0351 [−0.0388, −0.0311]	0.058
16	8	0.0005 [−0.0007, 0.0018]	−0.0847	−0.0671	−0.0176 [−0.0213, −0.0138]	0.089
Mixtral						
8	2	0.0063 [0.0049, 0.0075]	−0.0573	−0.0379	−0.0194 [−0.0226, −0.0163]	0.092
16	2	0.0096 [0.0084, 0.0109]	−0.0823	−0.0532	−0.0291 [−0.0319, −0.0263]	0.028
32	2	0.0154 [0.0138, 0.0170]	−0.1801	−0.0965	−0.0835 [−0.0892, −0.0776]	0.013
DeepSeek-MoE						
7	2	0.0464 [0.0438, 0.0491]	−0.2271	−0.1470	−0.0801 [−0.0850, −0.0751]	0.010
7	3	0.0188 [0.0171, 0.0205]	−0.1995	−0.1267	−0.0728 [−0.0781, −0.0676]	0.008
7	4	0.0097 [0.0083, 0.0111]	−0.1656	−0.1111	−0.0545 [−0.0593, −0.0498]	0.019
7	5	0.0060 [0.0045, 0.0074]	−0.1303	−0.0989	−0.0314 [−0.0358, −0.0275]	0.033
7	6	0.0015 [0.0003, 0.0028]	−0.1053	−0.0891	−0.0162 [−0.0201, −0.0124]	0.059
14	2	0.0390 [0.0365, 0.0415]	−0.1849	−0.1213	−0.0636 [−0.0677, −0.0594]	0.011
14	3	0.0174 [0.0158, 0.0191]	−0.1524	−0.0969	−0.0555 [−0.0596, −0.0509]	0.014
14	4	0.0089 [0.0075, 0.0102]	−0.1214	−0.0814	−0.0400 [−0.0444, −0.0359]	0.031
14	5	0.0051 [0.0038, 0.0063]	−0.0979	−0.0707	−0.0273 [−0.0314, −0.0235]	0.042
14	6	0.0020 [0.0007, 0.0034]	−0.0721	−0.0619	−0.0102 [−0.0140, −0.0065]	0.083
28	2	0.0357 [0.0329, 0.0386]	−0.2504	−0.1677	−0.0827 [−0.0880, −0.0776]	0.009
28	3	0.0181 [0.0163, 0.0200]	−0.2075	−0.1300	−0.0775 [−0.0825, −0.0723]	0.014
28	4	0.0103 [0.0088, 0.0118]	−0.1606	−0.1087	−0.0518 [−0.0569, −0.0471]	0.032
28	5	0.0042 [0.0029, 0.0055]	−0.1214	−0.0922	−0.0292 [−0.0335, −0.0252]	0.055
28	6	0.0016 [0.0004, 0.0028]	−0.0939	−0.0795	−0.0145 [−0.0180, −0.0107]	0.092

E GEOMETRIC FACTORIAL IN QWEN3, GEMMA4, AND QWEN3.6

For Qwen3, Gemma4, and Qwen3.6, the geometric factorial uses the cached router inputs and router outputs from the geometry survey. The cached features are the inputs of each routed MoE block's router: the sparse-MoE input for Qwen3 and Qwen3.6, and the residual-stream input of the Gemma4 router module, whose router applies its own normalization and per-channel scale. Shared experts lie outside the routed space and never enter candidates, rivals, contexts, or alternative routes.

The cache holds 256 sequences of 512 tokens with documents packed back to back, so connected components of source contexts would form a single training group. We therefore split documents into training and held-out sets (80/20, seed zero) and resample the 43 connected groups of held-out documents per model. For OLMoE, these groups coincide with the source contexts of Appendix D. With source contexts as units instead, 61 of 63 intervals lie below zero (pooled macro -0.0243). Restricting the primary analysis to tokens whose whole route is eligible gives -0.0237. The 101,324–105,155 training tokens per model leave 55–81% of the routed experts eligible (Table 13). The 34 cells in which at least 90% of the 2,048 evaluation tokens enter the estimate all have intervals below zero. In every cell both context effects are negative with $T_s < T_r$, and the actual-prefix percentile ranges from 0.012 to 0.240. Table 14 summarizes the router-order and leave-one-out interactions by model, and Table 15 gives every router-order cell.

Table 13: Factorial coverage in the three newer models: eligible routed experts, training and evaluation tokens, evaluation tokens whose full route is eligible, and bootstrap units (held-out document groups).

Model	L	Eligible	Train tokens	Eval tokens	Complete routes	Units
Qwen3	12	91/128	105,081	2,048	1,804	43
Qwen3	24	88/128	105,081	2,048	1,828	43
Qwen3	41	87/128	105,081	2,048	1,830	43
Gemma4	8	95/128	105,155	2,048	1,756	43
Gemma4	15	104/128	105,155	2,048	1,792	43
Gemma4	26	93/128	105,155	2,048	1,784	43
Qwen3.6	10	141/256	101,324	2,048	1,084	43
Qwen3.6	20	160/256	101,324	2,048	1,299	43
Qwen3.6	34	140/256	101,324	2,048	1,274	43

Table 14: Geometric-factorial summary for the three newer models. Counts for D are negative / interval below zero / interval above zero, and counts for A_{actual} are positive / interval above zero / interval below zero. Macros are equal-cell averages, and macro D has a 95% interval over held-out document groups.

Ordering	Model	Cells	D counts	Macro D [95% CI]	A_{actual} counts	Macro A_{actual}
Router order	Qwen3	21	21/19/0	$-0.0234\,[-0.0256, -0.0213]$	21/18/0	0.0108
Router order	Gemma4	21	21/20/0	$-0.0236\,[-0.0253, -0.0219]$	20/14/0	0.0074
Router order	Qwen3.6	21	21/17/0	$-0.0234\,[-0.0254, -0.0210]$	20/18/0	0.0106
Router order	Pooled	63	63/56/0	$-0.0235\,[-0.0246, -0.0223]$	61/50/0	0.0096
Leave-one-out	Qwen3	24	24/22/0	$-0.0245\,[-0.0271, -0.0222]$	24/21/0	0.0079
Leave-one-out	Gemma4	24	24/23/0	$-0.0214\,[-0.0229, -0.0199]$	23/14/0	0.0052
Leave-one-out	Qwen3.6	24	24/21/0	$-0.0259\,[-0.0282, -0.0234]$	23/19/0	0.0084
Leave-one-out	Pooled	72	72/66/0	$-0.0239\,[-0.0252, -0.0227]$	70/54/0	0.0071

Table 15: Router-order geometric-factorial cells in the three newer models. Coverage is the share of 2,048 evaluation tokens entering the cell, and A_{actual} and D have paired 95% intervals over held-out document groups.

Model	L	j	Coverage	A_{actual} [95% CI]	D [95% CI]
Qwen3	12	2	0.984	$0.0320\,[\ 0.0280, 0.0353]$	$-0.0343\,[-0.0385, -0.0300]$
Qwen3	12	3	0.977	$0.0179\,[\ 0.0154, 0.0201]$	$-0.0344\,[-0.0379, -0.0311]$
Qwen3	12	4	0.960	$0.0097\,[\ 0.0070, 0.0121]$	$-0.0264\,[-0.0298, -0.0224]$
Qwen3	12	5	0.949	$0.0060\,[\ 0.0040, 0.0083]$	$-0.0180\,[-0.0219, -0.0138]$
Qwen3	12	6	0.932	$0.0041\,[\ 0.0020, 0.0063]$	$-0.0103\,[-0.0138, -0.0067]$
Qwen3	12	7	0.910	$0.0021\,[\ 0.0006, 0.0038]$	$-0.0054\,[-0.0090, -0.0012]$
Qwen3	12	8	0.881	$0.0023\,[\ 0.0002, 0.0045]$	$-0.0011\,[-0.0062, \ 0.0036]$
Qwen3	24	2	0.982	$0.0276\,[\ 0.0245, 0.0307]$	$-0.0292\,[-0.0332, -0.0254]$
Qwen3	24	3	0.971	$0.0160\,[\ 0.0133, 0.0189]$	$-0.0310\,[-0.0380, -0.0254]$
Qwen3	24	4	0.959	$0.0094\,[\ 0.0071, 0.0119]$	$-0.0284\,[-0.0339, -0.0235]$
Qwen3	24	5	0.947	$0.0068\,[\ 0.0043, 0.0094]$	$-0.0239\,[-0.0312, -0.0176]$
Qwen3	24	6	0.929	$0.0033\,[\ 0.0019, 0.0047]$	$-0.0174\,[-0.0244, -0.0114]$
Qwen3	24	7	0.912	$0.0010\,[-0.0003, 0.0022]$	$-0.0115\,[-0.0170, -0.0072]$
Qwen3	24	8	0.893	$0.0016\,[-0.0009, 0.0051]$	$-0.0061\,[-0.0114, -0.0008]$
Qwen3	41	2	0.977	$0.0399\,[\ 0.0346, 0.0454]$	$-0.0460\,[-0.0508, -0.0412]$
Qwen3	41	3	0.964	$0.0196\,[\ 0.0163, 0.0235]$	$-0.0504\,[-0.0562, -0.0448]$
Qwen3	41	4	0.956	$0.0117\,[\ 0.0097, 0.0137]$	$-0.0422\,[-0.0472, -0.0371]$
Qwen3	41	5	0.941	$0.0062\,[\ 0.0049, 0.0077]$	$-0.0332\,[-0.0377, -0.0284]$
Qwen3	41	6	0.925	$0.0048\,[\ 0.0030, 0.0064]$	$-0.0197\,[-0.0246, -0.0152]$
Qwen3	41	7	0.911	$0.0030\,[\ 0.0013, 0.0047]$	$-0.0165\,[-0.0212, -0.0113]$
Qwen3	41	8	0.894	$0.0016\,[-0.0004, 0.0038]$	$-0.0034\,[-0.0076, \ 0.0008]$
Gemma4	8	2	0.959	$0.0362\,[\ 0.0315, 0.0407]$	$-0.0538\,[-0.0596, -0.0483]$
Gemma4	8	3	0.941	$0.0142\,[\ 0.0110, 0.0175]$	$-0.0580\,[-0.0643, -0.0523]$

Continued on next page

Table 15 (continued)

Model	L	j	Coverage	A_{actual} [95% CI]	D [95% CI]
Gemma4	8	4	0.927	0.0044 [0.0019, 0.0068]	−0.0463 [−0.0520, −0.0410]
Gemma4	8	5	0.917	0.0018 [−0.0014, 0.0049]	−0.0317 [−0.0375, −0.0261]
Gemma4	8	6	0.896	−0.0010 [−0.0039, 0.0021]	−0.0238 [−0.0292, −0.0185]
Gemma4	8	7	0.877	0.0009 [−0.0033, 0.0057]	−0.0108 [−0.0160, −0.0049]
Gemma4	8	8	0.857	0.0017 [−0.0018, 0.0049]	−0.0035 [−0.0069, −0.0000]
Gemma4	15	2	0.959	0.0157 [0.0134, 0.0181]	−0.0197 [−0.0228, −0.0170]
Gemma4	15	3	0.949	0.0096 [0.0073, 0.0117]	−0.0191 [−0.0220, −0.0163]
Gemma4	15	4	0.936	0.0062 [0.0046, 0.0080]	−0.0160 [−0.0192, −0.0128]
Gemma4	15	5	0.924	0.0035 [0.0015, 0.0053]	−0.0133 [−0.0167, −0.0099]
Gemma4	15	6	0.909	0.0026 [0.0010, 0.0045]	−0.0075 [−0.0104, −0.0044]
Gemma4	15	7	0.889	0.0009 [−0.0009, 0.0027]	−0.0040 [−0.0061, −0.0015]
Gemma4	15	8	0.875	0.0010 [−0.0006, 0.0026]	−0.0031 [−0.0050, −0.0011]
Gemma4	26	2	0.957	0.0276 [0.0250, 0.0301]	−0.0441 [−0.0493, −0.0393]
Gemma4	26	3	0.945	0.0135 [0.0118, 0.0152]	−0.0452 [−0.0501, −0.0410]
Gemma4	26	4	0.932	0.0084 [0.0071, 0.0097]	−0.0370 [−0.0417, −0.0328]
Gemma4	26	5	0.917	0.0047 [0.0030, 0.0061]	−0.0281 [−0.0322, −0.0238]
Gemma4	26	6	0.905	0.0028 [0.0014, 0.0045]	−0.0184 [−0.0230, −0.0140]
Gemma4	26	7	0.887	0.0014 [0.0002, 0.0026]	−0.0101 [−0.0136, −0.0068]
Gemma4	26	8	0.871	0.0001 [−0.0014, 0.0014]	−0.0021 [−0.0065, 0.0036]
Qwen3.6	10	2	0.865	0.0406 [0.0348, 0.0463]	−0.0361 [−0.0401, −0.0316]
Qwen3.6	10	3	0.812	0.0205 [0.0162, 0.0249]	−0.0315 [−0.0371, −0.0243]
Qwen3.6	10	4	0.760	0.0090 [0.0062, 0.0115]	−0.0305 [−0.0359, −0.0256]
Qwen3.6	10	5	0.702	0.0063 [0.0039, 0.0088]	−0.0228 [−0.0282, −0.0173]
Qwen3.6	10	6	0.644	0.0046 [0.0008, 0.0087]	−0.0158 [−0.0209, −0.0105]
Qwen3.6	10	7	0.596	0.0022 [−0.0003, 0.0047]	−0.0049 [−0.0145, 0.0074]
Qwen3.6	10	8	0.529	−0.0006 [−0.0029, 0.0018]	−0.0086 [−0.0135, −0.0037]
Qwen3.6	20	2	0.934	0.0242 [0.0218, 0.0266]	−0.0262 [−0.0309, −0.0220]
Qwen3.6	20	3	0.904	0.0143 [0.0117, 0.0173]	−0.0243 [−0.0280, −0.0204]
Qwen3.6	20	4	0.864	0.0088 [0.0067, 0.0110]	−0.0177 [−0.0229, −0.0127]
Qwen3.6	20	5	0.814	0.0055 [0.0034, 0.0079]	−0.0149 [−0.0214, −0.0094]
Qwen3.6	20	6	0.762	0.0038 [0.0013, 0.0069]	−0.0116 [−0.0166, −0.0070]
Qwen3.6	20	7	0.702	0.0025 [0.0003, 0.0049]	−0.0040 [−0.0087, 0.0005]
Qwen3.6	20	8	0.634	0.0024 [−0.0003, 0.0058]	−0.0030 [−0.0077, 0.0017]
Qwen3.6	34	2	0.898	0.0351 [0.0295, 0.0404]	−0.0541 [−0.0612, −0.0472]
Qwen3.6	34	3	0.858	0.0171 [0.0137, 0.0210]	−0.0573 [−0.0634, −0.0514]
Qwen3.6	34	4	0.811	0.0105 [0.0077, 0.0134]	−0.0481 [−0.0548, −0.0415]
Qwen3.6	34	5	0.771	0.0057 [0.0033, 0.0079]	−0.0354 [−0.0425, −0.0286]
Qwen3.6	34	6	0.721	0.0027 [0.0007, 0.0047]	−0.0232 [−0.0289, −0.0176]
Qwen3.6	34	7	0.667	0.0037 [0.0014, 0.0061]	−0.0137 [−0.0190, −0.0092]
Qwen3.6	34	8	0.622	0.0028 [0.0003, 0.0057]	−0.0063 [−0.0125, 0.0006]

F Robustness and Numerical Validation

The robustness checks keep the same paired candidate and context construction while changing the scale of the score or the way residual opportunity is matched. Raw gain removes the fractional normalization. Nearest-residual matching pairs actual and alternative contexts by train-IQR-scaled residual opportunity. The 5th-percentile caliper uses the fifth percentile of training-set nearest-match distances after IQR scaling. Table 16 reports the cell-level estimates.

Table 16: Cell-level robustness of the geometric interaction. Each column gives one variant with paired 95% source-context bootstrap intervals. The primary fractional-novelty interaction of each cell is in Table 12.

L	j	Raw gain [95% CI]	Nearest residual [95% CI]	5th-percentile caliper [95% CI]
OLMoE				
4	2	$-0.0936\,[-0.0985,\ -0.0889]$	$-0.0942\,[-0.0995,\ -0.0891]$	$-0.0119\,[-0.0191,\ -0.0057]$
4	3	$-0.0784\,[-0.0826,\ -0.0742]$	$-0.0942\,[-0.0996,\ -0.0894]$	$-0.0172\,[-0.0238,\ -0.0113]$
4	4	$-0.0589\,[-0.0626,\ -0.0553]$	$-0.0796\,[-0.0853,\ -0.0742]$	$-0.0132\,[-0.0206,\ -0.0057]$
4	5	$-0.0399\,[-0.0430,\ -0.0366]$	$-0.0580\,[-0.0631,\ -0.0532]$	$-0.0121\,[-0.0185,\ -0.0065]$
4	6	$-0.0243\,[-0.0269,\ -0.0217]$	$-0.0408\,[-0.0456,\ -0.0366]$	$-0.0138\,[-0.0220,\ -0.0061]$
4	7	$-0.0132\,[-0.0154,\ -0.0109]$	$-0.0240\,[-0.0284,\ -0.0196]$	$-0.0066\,[-0.0150,\ 0.0019]$
4	8	$-0.0061\,[-0.0079,\ -0.0041]$	$-0.0137\,[-0.0184,\ -0.0098]$	$-0.0055\,[-0.0129,\ 0.0018]$
8	2	$-0.0669\,[-0.0703,\ -0.0634]$	$-0.0713\,[-0.0760,\ -0.0668]$	$-0.0133\,[-0.0177,\ -0.0089]$
8	3	$-0.0557\,[-0.0589,\ -0.0525]$	$-0.0697\,[-0.0744,\ -0.0650]$	$-0.0129\,[-0.0185,\ -0.0074]$
8	4	$-0.0423\,[-0.0450,\ -0.0394]$	$-0.0603\,[-0.0652,\ -0.0559]$	$-0.0087\,[-0.0146,\ -0.0029]$
8	5	$-0.0316\,[-0.0340,\ -0.0292]$	$-0.0508\,[-0.0556,\ -0.0462]$	$-0.0251\,[-0.0450,\ -0.0111]$
8	6	$-0.0193\,[-0.0216,\ -0.0174]$	$-0.0348\,[-0.0398,\ -0.0300]$	$-0.0230\,[-0.0439,\ -0.0083]$
8	7	$-0.0112\,[-0.0129,\ -0.0097]$	$-0.0234\,[-0.0274,\ -0.0194]$	$-0.0168\,[-0.0314,\ -0.0048]$
8	8	$-0.0040\,[-0.0053,\ -0.0028]$	$-0.0104\,[-0.0140,\ -0.0070]$	$-0.0091\,[-0.0190,\ -0.0007]$
16	2	$-0.0662\,[-0.0699,\ -0.0624]$	$-0.0716\,[-0.0765,\ -0.0672]$	$-0.0181\,[-0.0247,\ -0.0114]$
16	3	$-0.0554\,[-0.0587,\ -0.0524]$	$-0.0742\,[-0.0787,\ -0.0694]$	$-0.0221\,[-0.0281,\ -0.0164]$
16	4	$-0.0450\,[-0.0482,\ -0.0423]$	$-0.0694\,[-0.0745,\ -0.0644]$	$-0.0211\,[-0.0270,\ -0.0161]$
16	5	$-0.0326\,[-0.0353,\ -0.0300]$	$-0.0562\,[-0.0608,\ -0.0519]$	$-0.0156\,[-0.0220,\ -0.0099]$
16	6	$-0.0240\,[-0.0261,\ -0.0218]$	$-0.0464\,[-0.0516,\ -0.0420]$	$-0.0099\,[-0.0158,\ -0.0042]$
16	7	$-0.0145\,[-0.0161,\ -0.0129]$	$-0.0304\,[-0.0344,\ -0.0269]$	$-0.0060\,[-0.0116,\ -0.0001]$
16	8	$-0.0063\,[-0.0076,\ -0.0050]$	$-0.0145\,[-0.0178,\ -0.0113]$	$-0.0035\,[-0.0091,\ 0.0021]$
Mixtral				
8	2	$-0.0152\,[-0.0176,\ -0.0131]$	$-0.0152\,[-0.0178,\ -0.0128]$	$-0.0010\,[-0.0032,\ 0.0011]$
16	2	$-0.0221\,[-0.0242,\ -0.0200]$	$-0.0256\,[-0.0285,\ -0.0226]$	$-0.0040\,[-0.0070,\ -0.0009]$
32	2	$-0.0297\,[-0.0322,\ -0.0273]$	$-0.0791\,[-0.0853,\ -0.0733]$	$-0.0066\,[-0.0117,\ -0.0010]$
DeepSeek-MoE				
7	2	$-0.0760\,[-0.0798,\ -0.0719]$	$-0.0796\,[-0.0847,\ -0.0748]$	$-0.0137\,[-0.0263,\ -0.0045]$
7	3	$-0.0541\,[-0.0577,\ -0.0505]$	$-0.0690\,[-0.0743,\ -0.0640]$	$-0.0193\,[-0.0275,\ -0.0118]$
7	4	$-0.0344\,[-0.0374,\ -0.0315]$	$-0.0505\,[-0.0551,\ -0.0462]$	$-0.0202\,[-0.0322,\ -0.0091]$
7	5	$-0.0177\,[-0.0198,\ -0.0154]$	$-0.0285\,[-0.0323,\ -0.0244]$	$-0.0094\,[-0.0165,\ -0.0021]$
7	6	$-0.0076\,[-0.0093,\ -0.0060]$	$-0.0146\,[-0.0182,\ -0.0112]$	$-0.0106\,[-0.0237,\ 0.0003]$
14	2	$-0.0574\,[-0.0613,\ -0.0538]$	$-0.0601\,[-0.0646,\ -0.0559]$	$-0.0104\,[-0.0179,\ -0.0040]$
14	3	$-0.0397\,[-0.0431,\ -0.0365]$	$-0.0492\,[-0.0535,\ -0.0449]$	$-0.0120\,[-0.0181,\ -0.0062]$
14	4	$-0.0240\,[-0.0266,\ -0.0217]$	$-0.0341\,[-0.0384,\ -0.0303]$	$-0.0117\,[-0.0177,\ -0.0057]$
14	5	$-0.0144\,[-0.0165,\ -0.0125]$	$-0.0222\,[-0.0260,\ -0.0188]$	$-0.0115\,[-0.0172,\ -0.0058]$
14	6	$-0.0048\,[-0.0066,\ -0.0031]$	$-0.0079\,[-0.0115,\ -0.0050]$	$-0.0062\,[-0.0115,\ -0.0010]$
28	2	$-0.0653\,[-0.0691,\ -0.0616]$	$-0.0814\,[-0.0866,\ -0.0761]$	$-0.0087\,[-0.0167,\ -0.0007]$
28	3	$-0.0465\,[-0.0492,\ -0.0438]$	$-0.0707\,[-0.0754,\ -0.0663]$	$-0.0065\,[-0.0133,\ -0.0002]$
28	4	$-0.0258\,[-0.0282,\ -0.0236]$	$-0.0454\,[-0.0498,\ -0.0410]$	$-0.0060\,[-0.0127,\ 0.0005]$
28	5	$-0.0123\,[-0.0139,\ -0.0105]$	$-0.0242\,[-0.0279,\ -0.0202]$	$-0.0048\,[-0.0102,\ 0.0011]$
28	6	$-0.0050\,[-0.0061,\ -0.0038]$	$-0.0115\,[-0.0150,\ -0.0084]$	$-0.0021\,[-0.0071,\ 0.0030]$

Raw-gain and nearest-residual intervals lie below zero in all 39 cells. The strict-caliper point estimate is also negative in every cell, with 31 of 39 intervals below zero. The caliper retains 3,875 of 49,152 token-cells (7.88%) in OLMoE and Mixtral and 2,322 of 30,720 (7.56%) in DeepSeek-MoE, so this last variant is best read as a deliberately stringent, low-coverage check.

Stratifying the router-order interaction within each cell by the router-logit margin between s and r, actual-prefix residual $\rho_x(S)$, and candidate training load (quintiles), and by token domain and sequence position (thirds), leaves D negative in every stratum of OLMoE, Mixtral, DeepSeek-MoE, and their pool, with every interval below zero. The same holds for the leave-one-out interaction and, without the domain split, for Qwen3, Gemma4, and Qwen3.6. The interaction is weakest for the most heavily loaded candidates (-0.040 in the top load quintile against -0.055 to -0.056 in the others, three models pooled), so it is not driven by heavily used experts.

Table 17: Independent CPU tall-SVD recomputation agrees with the residual computation of the factorial on 32 sampled actual and 32 alternative prefixes per layer. Columns report maximum absolute residual and projector discrepancies, rank mismatches, and non-finite values.

Model	L	Max residual	Max projector	Rank mism.	Non-finite
OLMoE	4	1.25×10^{-7}	1.77×10^{-4}	0	0
OLMoE	8	1.55×10^{-7}	1.12×10^{-4}	0	0
OLMoE	16	1.83×10^{-7}	9.78×10^{-5}	0	0
Mixtral	8	2.10×10^{-7}	1.18×10^{-5}	0	0
Mixtral	16	1.53×10^{-7}	2.36×10^{-5}	0	0
Mixtral	32	7.79×10^{-8}	1.02×10^{-4}	0	0
DeepSeek	7	1.18×10^{-7}	4.98×10^{-5}	0	0
DeepSeek	14	8.98×10^{-8}	6.38×10^{-5}	0	0
DeepSeek	28	1.56×10^{-7}	6.16×10^{-5}	0	0

Across the nine model-layer runs (Table 17), the largest residual and projector discrepancies are 2.10×10^{-7} and 1.77×10^{-4}, with no rank mismatch and no non-finite value.

G Order-Robustness of the Interaction

The factorial defines each candidate's context by router-score order. We replace this order in three ways while keeping the rival, the five alternative routes, the uncentered bases, and the paired source-context bootstrap of Appendix C. For token x with selected set E, a candidate $c \in E$ and a context $P \subseteq E \setminus \{c\}$ of size ℓ, the interaction is

$$D(c; P) = \big[q(c \mid P) - q(r \mid P)\big] - \frac{1}{M} \sum_{m} \big[q(c \mid W_{m,\ell}) - q(r \mid W_{m,\ell})\big],$$

where $W_{m,\ell}$ is the length-ℓ prefix of alternative route m. *Leave-one-out* sets $P = E \setminus \{c\}$ for every router rank of c. *Random order* draws 16 uniformly random orders of E per token (both orders for Mixtral) and reports D by position, the candidate at position j being conditioned on its predecessors. *Permutation-averaged* reports, for each router rank, the average of D over the orders and positions at which that expert is the candidate with a non-empty context.

Table 18 summarizes the four orderings, and Tables 19 and 20 give every cell. All 135 cells of the three new orderings are negative with intervals below zero.

Table 18: Equal-cell macro interaction under each ordering for pooled OLMoE and Mixtral cells, DeepSeek-MoE cells, and all three models, with paired 95% source-context intervals.

Ordering	Models	Cells	Macro D [95% CI]
Router order (main text)	OLMoE, Mixtral	24	$-0.0555\,[-0.0578,\ -0.0534]$
	DeepSeek	15	$-0.0471\,[-0.0491,\ -0.0449]$
	All three	39	$-0.0523\,[-0.0538,\ -0.0506]$
Leave-one-out	OLMoE, Mixtral	30	$-0.0567\,[-0.0588,\ -0.0547]$
	DeepSeek	18	$-0.0548\,[-0.0569,\ -0.0524]$
	All three	48	$-0.0560\,[-0.0575,\ -0.0545]$
Random order	OLMoE, Mixtral	24	$-0.0543\,[-0.0565,\ -0.0522]$
	DeepSeek	15	$-0.0461\,[-0.0480,\ -0.0441]$
	All three	39	$-0.0511\,[-0.0526,\ -0.0496]$
Permutation-averaged	OLMoE, Mixtral	30	$-0.0533\,[-0.0554,\ -0.0515]$
	DeepSeek	18	$-0.0462\,[-0.0480,\ -0.0442]$
	All three	48	$-0.0507\,[-0.0521,\ -0.0492]$

Table 19: Interaction by candidate router rank for leave-one-out contexts and permutation-averaged random orders. Every interval lies below zero.

Model	L	Rank	Leave-one-out D [95% CI]	Permutation-averaged D [95% CI]
OLMoE	4	1	$-0.1098\,[-0.1154,\,-0.1044]$	$-0.0923\,[-0.0974,\,-0.0871]$
OLMoE	4	2	$-0.1086\,[-0.1141,\,-0.1030]$	$-0.0937\,[-0.0987,\,-0.0886]$
OLMoE	4	3	$-0.0961\,[-0.1019,\,-0.0908]$	$-0.0874\,[-0.0927,\,-0.0826]$
OLMoE	4	4	$-0.0807\,[-0.0858,\,-0.0755]$	$-0.0759\,[-0.0808,\,-0.0709]$
OLMoE	4	5	$-0.0624\,[-0.0674,\,-0.0574]$	$-0.0591\,[-0.0636,\,-0.0544]$
OLMoE	4	6	$-0.0444\,[-0.0490,\,-0.0396]$	$-0.0435\,[-0.0484,\,-0.0390]$
OLMoE	4	7	$-0.0281\,[-0.0332,\,-0.0236]$	$-0.0278\,[-0.0326,\,-0.0233]$
OLMoE	4	8	$-0.0146\,[-0.0194,\,-0.0100]$	$-0.0141\,[-0.0189,\,-0.0095]$
OLMoE	8	1	$-0.0834\,[-0.0881,\,-0.0787]$	$-0.0740\,[-0.0783,\,-0.0700]$
OLMoE	8	2	$-0.0776\,[-0.0822,\,-0.0737]$	$-0.0704\,[-0.0748,\,-0.0666]$
OLMoE	8	3	$-0.0707\,[-0.0755,\,-0.0663]$	$-0.0662\,[-0.0706,\,-0.0622]$
OLMoE	8	4	$-0.0608\,[-0.0650,\,-0.0567]$	$-0.0580\,[-0.0620,\,-0.0541]$
OLMoE	8	5	$-0.0527\,[-0.0574,\,-0.0484]$	$-0.0512\,[-0.0556,\,-0.0473]$
OLMoE	8	6	$-0.0369\,[-0.0413,\,-0.0328]$	$-0.0375\,[-0.0417,\,-0.0333]$
OLMoE	8	7	$-0.0260\,[-0.0295,\,-0.0225]$	$-0.0255\,[-0.0292,\,-0.0219]$
OLMoE	8	8	$-0.0111\,[-0.0146,\,-0.0078]$	$-0.0113\,[-0.0148,\,-0.0081]$
OLMoE	16	1	$-0.0892\,[-0.0942,\,-0.0844]$	$-0.0793\,[-0.0839,\,-0.0749]$
OLMoE	16	2	$-0.0832\,[-0.0880,\,-0.0782]$	$-0.0758\,[-0.0803,\,-0.0713]$
OLMoE	16	3	$-0.0762\,[-0.0814,\,-0.0715]$	$-0.0714\,[-0.0760,\,-0.0671]$
OLMoE	16	4	$-0.0702\,[-0.0751,\,-0.0655]$	$-0.0666\,[-0.0711,\,-0.0623]$
OLMoE	16	5	$-0.0583\,[-0.0630,\,-0.0534]$	$-0.0566\,[-0.0611,\,-0.0522]$
OLMoE	16	6	$-0.0495\,[-0.0546,\,-0.0449]$	$-0.0490\,[-0.0540,\,-0.0445]$
OLMoE	16	7	$-0.0338\,[-0.0375,\,-0.0298]$	$-0.0345\,[-0.0382,\,-0.0309]$
OLMoE	16	8	$-0.0176\,[-0.0213,\,-0.0138]$	$-0.0189\,[-0.0226,\,-0.0150]$
Mixtral	8	1	$-0.0197\,[-0.0222,\,-0.0174]$	$-0.0197\,[-0.0222,\,-0.0174]$
Mixtral	8	2	$-0.0194\,[-0.0226,\,-0.0163]$	$-0.0194\,[-0.0226,\,-0.0163]$
Mixtral	16	1	$-0.0310\,[-0.0334,\,-0.0285]$	$-0.0310\,[-0.0334,\,-0.0285]$
Mixtral	16	2	$-0.0291\,[-0.0319,\,-0.0263]$	$-0.0291\,[-0.0319,\,-0.0263]$
Mixtral	32	1	$-0.0771\,[-0.0820,\,-0.0726]$	$-0.0771\,[-0.0820,\,-0.0726]$
Mixtral	32	2	$-0.0835\,[-0.0892,\,-0.0776]$	$-0.0835\,[-0.0892,\,-0.0776]$
DeepSeek	7	1	$-0.0973\,[-0.1020,\,-0.0926]$	$-0.0744\,[-0.0785,\,-0.0704]$
DeepSeek	7	2	$-0.0932\,[-0.0987,\,-0.0880]$	$-0.0748\,[-0.0793,\,-0.0705]$
DeepSeek	7	3	$-0.0719\,[-0.0769,\,-0.0668]$	$-0.0617\,[-0.0662,\,-0.0573]$
DeepSeek	7	4	$-0.0533\,[-0.0580,\,-0.0488]$	$-0.0464\,[-0.0505,\,-0.0422]$
DeepSeek	7	5	$-0.0313\,[-0.0355,\,-0.0275]$	$-0.0281\,[-0.0320,\,-0.0246]$
DeepSeek	7	6	$-0.0162\,[-0.0201,\,-0.0124]$	$-0.0149\,[-0.0186,\,-0.0113]$
DeepSeek	14	1	$-0.0762\,[-0.0801,\,-0.0724]$	$-0.0611\,[-0.0648,\,-0.0579]$
DeepSeek	14	2	$-0.0688\,[-0.0731,\,-0.0647]$	$-0.0576\,[-0.0611,\,-0.0541]$
DeepSeek	14	3	$-0.0523\,[-0.0564,\,-0.0478]$	$-0.0460\,[-0.0497,\,-0.0423]$
DeepSeek	14	4	$-0.0378\,[-0.0418,\,-0.0340]$	$-0.0343\,[-0.0379,\,-0.0308]$
DeepSeek	14	5	$-0.0255\,[-0.0294,\,-0.0220]$	$-0.0244\,[-0.0282,\,-0.0210]$
DeepSeek	14	6	$-0.0102\,[-0.0140,\,-0.0065]$	$-0.0105\,[-0.0141,\,-0.0069]$
DeepSeek	28	1	$-0.0975\,[-0.1022,\,-0.0930]$	$-0.0748\,[-0.0787,\,-0.0710]$
DeepSeek	28	2	$-0.0905\,[-0.0959,\,-0.0854]$	$-0.0732\,[-0.0776,\,-0.0692]$
DeepSeek	28	3	$-0.0743\,[-0.0790,\,-0.0695]$	$-0.0639\,[-0.0682,\,-0.0598]$
DeepSeek	28	4	$-0.0491\,[-0.0538,\,-0.0446]$	$-0.0453\,[-0.0497,\,-0.0412]$
DeepSeek	28	5	$-0.0270\,[-0.0312,\,-0.0230]$	$-0.0258\,[-0.0301,\,-0.0218]$
DeepSeek	28	6	$-0.0145\,[-0.0180,\,-0.0107]$	$-0.0143\,[-0.0177,\,-0.0107]$

Table 20: Interaction by context position under router order and random order. With random order, magnitudes no longer shrink at later positions.

Model	L	j	Router order D [95% CI]	Random order D [95% CI]
OLMoE	4	2	$-0.0970\,[-0.1023,\,-0.0914]$	$-0.0380\,[-0.0418,\,-0.0345]$
OLMoE	4	3	$-0.1020\,[-0.1078,\,-0.0964]$	$-0.0573\,[-0.0618,\,-0.0532]$
OLMoE	4	4	$-0.0896\,[-0.0954,\,-0.0839]$	$-0.0637\,[-0.0681,\,-0.0595]$

Continued on next page

Table 20 (continued)

Model	L	j	Router order D [95% CI]	Random order D [95% CI]
OLMoE	4	5	$-0.0677\,[-0.0729, -0.0625]$	$-0.0671\,[-0.0717, -0.0630]$
OLMoE	4	6	$-0.0473\,[-0.0523, -0.0423]$	$-0.0692\,[-0.0736, -0.0649]$
OLMoE	4	7	$-0.0292\,[-0.0343, -0.0242]$	$-0.0694\,[-0.0738, -0.0649]$
OLMoE	4	8	$-0.0146\,[-0.0194, -0.0100]$	$-0.0671\,[-0.0716, -0.0625]$
OLMoE	8	2	$-0.0743\,[-0.0790, -0.0706]$	$-0.0344\,[-0.0376, -0.0318]$
OLMoE	8	3	$-0.0762\,[-0.0810, -0.0716]$	$-0.0472\,[-0.0507, -0.0437]$
OLMoE	8	4	$-0.0669\,[-0.0714, -0.0625]$	$-0.0517\,[-0.0553, -0.0482]$
OLMoE	8	5	$-0.0577\,[-0.0627, -0.0531]$	$-0.0525\,[-0.0567, -0.0488]$
OLMoE	8	6	$-0.0399\,[-0.0442, -0.0353]$	$-0.0536\,[-0.0576, -0.0498]$
OLMoE	8	7	$-0.0266\,[-0.0304, -0.0230]$	$-0.0530\,[-0.0569, -0.0493]$
OLMoE	8	8	$-0.0111\,[-0.0146, -0.0078]$	$-0.0525\,[-0.0562, -0.0486]$
OLMoE	16	2	$-0.0753\,[-0.0801, -0.0703]$	$-0.0402\,[-0.0434, -0.0366]$
OLMoE	16	3	$-0.0798\,[-0.0851, -0.0749]$	$-0.0547\,[-0.0589, -0.0509]$
OLMoE	16	4	$-0.0760\,[-0.0813, -0.0709]$	$-0.0589\,[-0.0631, -0.0548]$
OLMoE	16	5	$-0.0632\,[-0.0683, -0.0582]$	$-0.0608\,[-0.0651, -0.0565]$
OLMoE	16	6	$-0.0527\,[-0.0579, -0.0479]$	$-0.0616\,[-0.0659, -0.0577]$
OLMoE	16	7	$-0.0351\,[-0.0388, -0.0311]$	$-0.0604\,[-0.0644, -0.0565]$
OLMoE	16	8	$-0.0176\,[-0.0213, -0.0138]$	$-0.0593\,[-0.0636, -0.0553]$
Mixtral	8	2	$-0.0194\,[-0.0226, -0.0163]$	$-0.0195\,[-0.0223, -0.0169]$
Mixtral	16	2	$-0.0291\,[-0.0319, -0.0263]$	$-0.0301\,[-0.0328, -0.0275]$
Mixtral	32	2	$-0.0835\,[-0.0892, -0.0776]$	$-0.0803\,[-0.0857, -0.0751]$
DeepSeek	7	2	$-0.0801\,[-0.0850, -0.0751]$	$-0.0297\,[-0.0328, -0.0269]$
DeepSeek	7	3	$-0.0728\,[-0.0781, -0.0676]$	$-0.0471\,[-0.0507, -0.0433]$
DeepSeek	7	4	$-0.0545\,[-0.0593, -0.0498]$	$-0.0539\,[-0.0581, -0.0500]$
DeepSeek	7	5	$-0.0314\,[-0.0358, -0.0275]$	$-0.0586\,[-0.0626, -0.0545]$
DeepSeek	7	6	$-0.0162\,[-0.0201, -0.0124]$	$-0.0603\,[-0.0642, -0.0564]$
DeepSeek	14	2	$-0.0636\,[-0.0677, -0.0594]$	$-0.0244\,[-0.0268, -0.0220]$
DeepSeek	14	3	$-0.0555\,[-0.0596, -0.0509]$	$-0.0383\,[-0.0415, -0.0350]$
DeepSeek	14	4	$-0.0400\,[-0.0444, -0.0359]$	$-0.0428\,[-0.0465, -0.0395]$
DeepSeek	14	5	$-0.0273\,[-0.0314, -0.0235]$	$-0.0446\,[-0.0483, -0.0410]$
DeepSeek	14	6	$-0.0102\,[-0.0140, -0.0065]$	$-0.0451\,[-0.0487, -0.0419]$
DeepSeek	28	2	$-0.0827\,[-0.0880, -0.0776]$	$-0.0301\,[-0.0330, -0.0272]$
DeepSeek	28	3	$-0.0775\,[-0.0825, -0.0723]$	$-0.0472\,[-0.0510, -0.0437]$
DeepSeek	28	4	$-0.0518\,[-0.0569, -0.0471]$	$-0.0544\,[-0.0581, -0.0507]$
DeepSeek	28	5	$-0.0292\,[-0.0335, -0.0252]$	$-0.0572\,[-0.0613, -0.0534]$
DeepSeek	28	6	$-0.0145\,[-0.0180, -0.0107]$	$-0.0587\,[-0.0625, -0.0546]$

H Residual-Matched Cross-Token Control

This control asks how much of the negative interaction is explained by residual opportunity alone. In most token-cells, the alternative contexts leave more residual energy than the actual context. We therefore compare the actual prefix with unrelated prefixes that leave nearly the same residual on the same token. The control keeps the token, position, candidate s, and rival r fixed, and replaces the actual context S_x with actual length-$(j-1)$ prefixes $S_{x'}$ from other held-out tokens in different source contexts. Donor prefixes must share no expert with the route or rival, matching the exclusion used for alternative contexts. Each donor is evaluated on token x. We keep donors with $|\ln \rho_x(S_{x'}) - \ln \rho_x(S_x)| \le 0.05$ and average over up to five nearest distinct donor sets.

The primary contrast is $\Delta = A_{\text{actual}} - A_{\text{null}}$, where $A_{\text{null}} = \text{mean}_i[q_x(s \mid S_{x'_i}) - q_x(r \mid S_{x'_i})]$. The control interaction $D_{\text{null}} = A_{\text{null}} - A_{\text{alt}}$ satisfies $D = D_{\text{null}} + \Delta$ on matched token-cells. Negative Δ means that the co-selected context removes more of the selected candidate's advantage than an unrelated context of matched residual opportunity. Statistics follow the main analysis. As for D, fitted bases push Δ downward because a candidate's basis carries the directions of frequent partners (Appendix C.2). The design, caliper, and variants were fixed before any control quantity was computed.

Matched donors exist for 13.4% of token-cells (10.7% in OLMoE, 27.9% in Mixtral, 14.3% in DeepSeek-MoE). In 85–98.5% of token-cells, the actual context leaves less residual than every admissible unrelated prefix, so the matched subset has higher residual opportunity and smaller D

than the full set. Table 21 reports the primary analysis and variants, and Table 22 gives every cell. The selected candidate's own novelty drives Δ: the component of Δ that comes from $q_x(s \mid \cdot)$ is negative in all 39 cells. Of the 15 cells whose interval covers zero, seven are late OLMoE positions where the matched D is itself close to zero, and eight are DeepSeek-MoE cells, five of them at positions 5 and 6. Donors that may include other experts selected for the same token behave almost like the actual context (D_{null}/D of 0.90–0.96, three models pooled). Matched donors leave slightly more residual than the actual context (mean log gap 0.025 in OLMoE and Mixtral, 0.022 in DeepSeek-MoE), and a correction to zero gap gives $\Delta = -0.0041\,[-0.0053,\ -0.0029]$ for OLMoE and Mixtral pooled and $-0.0019\,[-0.0035,\ -0.0002]$ for DeepSeek-MoE. The no-caliper variant covers every token-cell but pairs tokens with donors that leave far more residual (mean log gap 0.23), so its Δ overstates the co-selection-specific component.

Table 21: Residual-matched cross-token control: equal-cell macros for pooled OLMoE and Mixtral cells, DeepSeek-MoE cells, and all three models. Coverage is the share of token-cells with matched donors. Counts give cells with negative Δ / interval below zero / interval above zero, and D is the observed interaction on the same matched token-cells. Macros are formed within each source context before averaging over contexts, so with sparse coverage they differ from the mean of the rows of Table 22.

Variant	Coverage	Cells	Δ [95% CI]	D_{null}	D
OLMoE and Mixtral pooled (24 cells)					
Primary (5 donors, caliper 0.05)	12.9%	23/17/0	$-0.0067\,[-0.0078,\ -0.0056]$	-0.0152	-0.0220
Nearest donor only	12.9%	23/16/0	$-0.0062\,[-0.0074,\ -0.0051]$	-0.0157	-0.0220
Caliper 0.02	6.3%	23/14/0	$-0.0043\,[-0.0056,\ -0.0029]$	-0.0143	-0.0186
Caliper 0.10	28.4%	23/20/0	$-0.0104\,[-0.0112,\ -0.0096]$	-0.0160	-0.0264
Nearest donor, no caliper	100.0%	24/24/0	$-0.0362\,[-0.0378,\ -0.0346]$	-0.0193	-0.0555
DeepSeek-MoE (15 cells)					
Primary (5 donors, caliper 0.05)	14.3%	14/7/0	$-0.0045\,[-0.0059,\ -0.0032]$	-0.0140	-0.0185
Nearest donor only	14.3%	14/7/0	$-0.0038\,[-0.0052,\ -0.0023]$	-0.0147	-0.0185
Caliper 0.02	8.3%	12/5/0	$-0.0014\,[-0.0033,\ \ 0.0004]$	-0.0142	-0.0156
Caliper 0.10	28.9%	15/12/0	$-0.0070\,[-0.0081,\ -0.0060]$	-0.0139	-0.0210
Nearest donor, no caliper	100.0%	15/15/0	$-0.0315\,[-0.0332,\ -0.0298]$	-0.0156	-0.0471
All three models pooled (39 cells)					
Primary (5 donors, caliper 0.05)	13.4%	37/24/0	$-0.0057\,[-0.0065,\ -0.0049]$	-0.0139	-0.0196
Nearest donor only	13.4%	37/23/0	$-0.0053\,[-0.0061,\ -0.0045]$	-0.0143	-0.0196
Caliper 0.02	7.0%	35/19/0	$-0.0039\,[-0.0051,\ -0.0027]$	-0.0129	-0.0169
Caliper 0.10	28.5%	38/32/0	$-0.0091\,[-0.0097,\ -0.0084]$	-0.0146	-0.0237
Nearest donor, no caliper	100.0%	39/39/0	$-0.0344\,[-0.0356,\ -0.0331]$	-0.0179	-0.0523

Table 22: Residual-matched control by cell (primary variant). The table reports matched token-cells, Δ with paired 95% interval, the residual-matched control interaction, and the observed interaction on the same token-cells.

Model	L	j	Tokens	Δ [95% CI]	D_{null}	D
OLMoE	4	2	414	$-0.0087\,[-0.0116,\ -0.0057]$	-0.0267	-0.0354
OLMoE	4	3	281	$-0.0041\,[-0.0075,\ -0.0008]$	-0.0405	-0.0447
OLMoE	4	4	186	$-0.0065\,[-0.0102,\ -0.0031]$	-0.0265	-0.0330
OLMoE	4	5	151	$-0.0045\,[-0.0080,\ -0.0012]$	-0.0147	-0.0193
OLMoE	4	6	132	$-0.0020\,[-0.0056,\ \ 0.0017]$	-0.0078	-0.0098
OLMoE	4	7	127	$-0.0013\,[-0.0062,\ \ 0.0036]$	-0.0028	-0.0041
OLMoE	4	8	137	$\ \ 0.0022\,[-0.0018,\ \ 0.0074]$	-0.0030	-0.0008
OLMoE	8	2	467	$-0.0116\,[-0.0148,\ -0.0089]$	-0.0185	-0.0301
OLMoE	8	3	346	$-0.0108\,[-0.0139,\ -0.0076]$	-0.0198	-0.0307
OLMoE	8	4	258	$-0.0086\,[-0.0120,\ -0.0048]$	-0.0165	-0.0251
OLMoE	8	5	190	$-0.0096\,[-0.0136,\ -0.0054]$	-0.0119	-0.0215
OLMoE	8	6	156	$-0.0037\,[-0.0080,\ \ 0.0009]$	-0.0091	-0.0129
OLMoE	8	7	133	$-0.0072\,[-0.0116,\ -0.0024]$	-0.0054	-0.0126
OLMoE	8	8	141	$-0.0002\,[-0.0055,\ \ 0.0059]$	-0.0035	-0.0037

Continued on next page

Table 22 (continued)

Model	L	j	Tokens	Δ [95% CI]	D_{null}	D
OLMoE	16	2	406	-0.0117 $[-0.0167, -0.0074]$	-0.0233	-0.0350
OLMoE	16	3	311	-0.0100 $[-0.0144, -0.0055]$	-0.0266	-0.0366
OLMoE	16	4	213	-0.0106 $[-0.0146, -0.0068]$	-0.0222	-0.0327
OLMoE	16	5	164	-0.0048 $[-0.0091, -0.0006]$	-0.0124	-0.0172
OLMoE	16	6	145	-0.0058 $[-0.0096, -0.0021]$	-0.0067	-0.0125
OLMoE	16	7	132	-0.0023 $[-0.0078, 0.0031]$	-0.0057	-0.0080
OLMoE	16	8	131	-0.0034 $[-0.0089, 0.0017]$	0.0004	-0.0030
Mixtral	8	2	960	-0.0033 $[-0.0044, -0.0023]$	-0.0011	-0.0044
Mixtral	16	2	512	-0.0061 $[-0.0078, -0.0045]$	-0.0019	-0.0080
Mixtral	32	2	241	-0.0117 $[-0.0155, -0.0079]$	-0.0032	-0.0149
DeepSeek	7	2	299	-0.0024 $[-0.0058, 0.0007]$	-0.0180	-0.0204
DeepSeek	7	3	218	-0.0068 $[-0.0101, -0.0032]$	-0.0163	-0.0231
DeepSeek	7	4	184	-0.0040 $[-0.0084, 0.0003]$	-0.0140	-0.0180
DeepSeek	7	5	153	-0.0010 $[-0.0052, 0.0029]$	-0.0061	-0.0071
DeepSeek	7	6	143	-0.0025 $[-0.0066, 0.0018]$	-0.0043	-0.0068
DeepSeek	14	2	462	-0.0067 $[-0.0096, -0.0038]$	-0.0177	-0.0244
DeepSeek	14	3	371	-0.0073 $[-0.0104, -0.0041]$	-0.0172	-0.0245
DeepSeek	14	4	284	-0.0055 $[-0.0088, -0.0019]$	-0.0119	-0.0174
DeepSeek	14	5	263	-0.0018 $[-0.0049, 0.0012]$	-0.0063	-0.0081
DeepSeek	14	6	277	-0.0017 $[-0.0047, 0.0016]$	-0.0014	-0.0031
DeepSeek	28	2	391	-0.0060 $[-0.0102, -0.0017]$	-0.0185	-0.0246
DeepSeek	28	3	380	-0.0039 $[-0.0081, 0.0008]$	-0.0168	-0.0206
DeepSeek	28	4	340	-0.0064 $[-0.0100, -0.0030]$	-0.0091	-0.0156
DeepSeek	28	5	327	-0.0040 $[-0.0073, -0.0007]$	-0.0107	-0.0147
DeepSeek	28	6	316	0.0003 $[-0.0037, 0.0041]$	-0.0062	-0.0059

I Input and Output Geometry of Experts

Shared input geometry would imply redundancy only if experts also wrote the same function. This section compares routed inputs with expert outputs, local Jacobians, and last-layer logit contributions on the same tokens. The pattern is consistent across views: the router chooses experts from a shared input neighborhood, while the selected expert maps remain differentiated enough to contribute distinct residual-stream updates.

The analysis uses cached per-token expert outputs from 128 held-out OLMoE contexts and 64 held-out Mixtral and DeepSeek-MoE contexts (65,536, 32,768, and 32,768 tokens). For each routed slot, the unweighted expert output is the recorded weighted contribution divided by its router weight, which is never zero for a routed slot. The DeepSeek-MoE capture records the native Top-6 route of a forward pass and excludes the shared-expert output. For each expert, we fit centered bases of router inputs and outputs on the same routed tokens, using the ESSI construction of Appendix B in float64: routing weights are renormalized per token, local tangents use 256 neighbors, and within-expert dispersion is load-weighted.

Every OLMoE and Mixtral expert has at least 2,048 routed tokens in the capture. In DeepSeek-MoE, distance, split-half, pair, dispersion, and capture statistics use the 62, 60, and 59 experts with at least 2,048 routed tokens at layers 7, 14, and 28, while the cosine analyses use every routed slot. Table 23 reports rank 128, and ranks 32 and 64 give the same pattern. Because local estimation noise affects both inputs and outputs, the table reports a split-half separation index: the between-expert distance of bases fitted on disjoint halves divided by the same-expert split-half distance.

This output cache is smaller than the geometry-survey pool, so input-side between-expert distances are 0.03–0.09 higher than in Table 3 and input-side ESSI is 0.015–0.10 higher. Subsampling the survey features to the same number of sequences reproduces the distance shift in all three models (between-expert distance 0.75–0.88) and, where computed (OLMoE and Mixtral), the ESSI shift. Inputs and outputs are therefore compared on the same tokens throughout. In every layer, outputs have higher ESSI at ranks 32 and 128 with expert-bootstrap intervals above zero, and the split-half separation index is higher for outputs than for inputs.

Table 23: Input versus output geometry on the same tokens (rank 128). Between is the mean between-expert chordal distance, and the split-half index divides the between-expert split-half distance by the same-expert split-half distance. Δ index is the output-minus-input difference of this index with an expert-bootstrap 95% interval. Pairs is the fraction of expert pairs farther apart in output than in input. The cosines compare the leader and a later co-selected expert in output space and in projected input space.

Model	L	Between in/out	Index in/out	Δ index [95% CI]	Pairs	Cos out	Cos in
OLMoE	4	0.876 / 0.954	1.241 / 1.468	$+0.227$ [0.203, 0.252]	1.000	0.032	0.642
OLMoE	8	0.851 / 0.957	1.192 / 1.478	$+0.286$ [0.260, 0.315]	1.000	0.026	0.682
OLMoE	16	0.851 / 0.953	1.178 / 1.496	$+0.318$ [0.269, 0.368]	0.992	0.120	0.737
Mixtral	8	0.793 / 0.922	1.065 / 1.082	$+0.017$ [0.012, 0.023]	1.000	0.336	0.685
Mixtral	16	0.759 / 0.946	1.093 / 1.124	$+0.031$ [0.026, 0.037]	1.000	0.284	0.757
Mixtral	32	0.783 / 0.852	1.132 / 1.300	$+0.169$ [0.146, 0.190]	1.000	0.588	0.811
DeepSeek	7	0.877 / 0.953	1.149 / 1.317	$+0.168$ [0.152, 0.186]	1.000	0.035	0.668
DeepSeek	14	0.859 / 0.962	1.104 / 1.284	$+0.180$ [0.164, 0.197]	1.000	0.024	0.683
DeepSeek	28	0.828 / 0.957	1.117 / 1.534	$+0.417$ [0.373, 0.467]	1.000	0.149	0.745

For co-selected experts on the same token, the output cosine is computed between the mean-subtracted outputs of the leader and each later expert, and the input analogue projects the token's input onto each expert's rank-128 input subspace. The same pairs on different tokens give cosines near zero. In OLMoE and DeepSeek-MoE the output alignment falls with route position (OLMoE layer 16: 0.148 at $j = 2$, 0.086 at $j = 8$; DeepSeek-MoE layer 28: 0.217 at $j = 2$, 0.093 at $j = 6$). Weighted by variance, the share of one expert's variance captured by another expert's basis is 0.51–0.56 for inputs but 0.18–0.32 for outputs in OLMoE and 0.49–0.58 versus 0.15–0.20 in DeepSeek-MoE. In Mixtral it falls from 0.70 for inputs to 0.57 for outputs at layer 16, is unchanged at layer 8, and rises from 0.74 to 0.88 at layer 32, where all experts share a few dominant output directions carrying 58–71% of output variance. Beyond these directions, Mixtral's outputs are also more separated than its inputs. Each OLMoE expert's outputs lie in the 1,024-dimensional column space of its own down-projection.

Jacobians. Two experts could write different outputs yet respond alike to changes of their shared input. We therefore compare the experts' Jacobians $J_e(x) \in \mathbb{R}^{d\times d}$, the derivatives of each unweighted expert output with respect to the expert input, which the router also reads. For the leader and each later co-selected expert on the same token we compute the Frobenius cosine $\langle J_a, J_b\rangle_F/(\|J_a\|_F\|J_b\|_F)$ on 8,192 tokens per layer, the same number from each of the contexts above, and average it over later route positions (Table 24). The cosine is estimated with 1,024 Gaussian probes shared by the experts of the token. On 32 tokens per layer, these estimates lie within 0.015 of the exact value, and because the probes are drawn anew for every token, their noise enters the bootstrap intervals. In OLMoE and DeepSeek-MoE the cosine is 0.009–0.069, against a projected-input alignment of 0.642–0.745 for the same pairs, and a load-matched expert not selected for the token gives 0.003–0.038, slightly less in every layer: experts whose input subspaces overlap apply nearly orthogonal local maps to the same token. In Mixtral the cosine is 0.20–0.43, and at layers 8 and 16 this alignment is shared by all eight experts: after subtracting their mean Jacobian on the token, the co-selected cosine is -0.145 and -0.143, close to the $-1/7$ expected if the Jacobians were a common part plus exchangeable independent parts. At layer 32 a pair-specific part remains: the co-selected cosine is -0.079, against -0.188 for the unselected expert.

Logit contributions. At the last analyzed layer of each model (OLMoE 16, Mixtral 32, DeepSeek-MoE 28), only the final RMSNorm and the unembedding follow the MoE block, so an expert output c contributes $W_U(g \odot c)$ to the next-token logits, scaled by a positive factor shared by all experts on the token (g: final-norm weight), and the comparison needs no approximation. On all 65,408, 32,704, and 32,704 tokens with a next token, we compare the vocabulary-centered logit contributions of the raw (not mean-subtracted) outputs of the leader and each later co-selected expert (Table 25). In OLMoE and DeepSeek-MoE the two contributions differ: their median cosine is 0.214 and 0.343, against a projected-input alignment of 0.737 and 0.745, and the 20 tokens each expert raises most overlap with a Jaccard index of 0.006 and 0.011. In Mixtral they are closely aligned (median cosine 0.887, Jaccard index 0.256), mostly in a way that depends neither on the pair nor on the token: a load-matched unselected expert gives 0.840, and the same expert pair, with the later expert's contribution

Table 24: Jacobian alignment between the leader and a later co-selected expert on the same token, measured by the Frobenius cosine and averaged over later route positions (8,192 tokens per layer, 95% source-context bootstrap interval). Unselected is a load-matched expert outside the route, and Other token uses the later expert's Jacobian on a token from another context. Mean removed subtracts the mean Jacobian of all routed experts on the same token, and Top-3 removed removes each expert's three leading output directions. Both are given as co-selected / unselected. Cos in is the projected-input cosine from Table 23.

Model	L	Co-selected [95% CI]	Unselected	Other token	Mean removed	Top-3 removed	Cos in
OLMoE	4	0.027 [0.026, 0.027]	0.018	0.020	−0.016 / −0.021	0.026 / 0.018	0.642
OLMoE	8	0.022 [0.022, 0.022]	0.016	0.016	−0.017 / −0.020	0.021 / 0.016	0.682
OLMoE	16	0.042 [0.041, 0.043]	0.038	0.034	−0.006 / −0.022	0.023 / 0.017	0.737
Mixtral	8	0.222 [0.218, 0.225]	0.212	0.137	−0.145 / −0.148	0.208 / 0.198	0.685
Mixtral	16	0.201 [0.198, 0.205]	0.189	0.126	−0.143 / −0.149	0.183 / 0.171	0.757
Mixtral	32	0.430 [0.421, 0.439]	0.371	0.274	−0.079 / −0.188	0.315 / 0.246	0.811
DeepSeek	7	0.013 [0.013, 0.014]	0.004	0.005	−0.012 / −0.020	0.013 / 0.004	0.668
DeepSeek	14	0.009 [0.009, 0.010]	0.003	0.004	−0.016 / −0.021	0.009 / 0.003	0.683
DeepSeek	28	0.069 [0.067, 0.071]	0.020	0.050	0.037 / −0.016	0.055 / 0.015	0.745

Table 25: Last-layer logit contributions from the leader and a later expert on the same token (later route positions averaged, 95% source-context intervals). Co-selected is the later selected expert, and Unselected is a load-matched expert that is not selected. Cos is the median cosine of the vocabulary-centered contributions, and Top-20 is the Jaccard index of the 20 tokens each contribution raises most. Same sign is the share of tokens on which both contributions move the correct next token in the same direction, and Both + is the share on which both raise it. Cos in follows Table 23.

Model	L	Pair	Cos [95% CI]	Top-20 [95% CI]	Same sign	Both +	Cos in
OLMoE	16	Co-selected	0.214 [0.206, 0.223]	0.006 [0.006, 0.006]	0.919	0.916	0.737
OLMoE	16	Unselected	0.146 [0.138, 0.155]	0.003 [0.002, 0.003]	0.808	0.801	
Mixtral	32	Co-selected	0.887 [0.880, 0.893]	0.256 [0.246, 0.266]	0.955	0.914	0.811
Mixtral	32	Unselected	0.840 [0.831, 0.848]	0.185 [0.177, 0.195]	0.925	0.888	
DeepSeek	28	Co-selected	0.343 [0.330, 0.356]	0.011 [0.010, 0.012]	0.819	0.797	0.745
DeepSeek	28	Unselected	0.175 [0.164, 0.185]	0.002 [0.002, 0.003]	0.645	0.611	

taken on a token from another context, still gives 0.537. In OLMoE, Mixtral, and DeepSeek-MoE, the co-selected expert's contribution is closer to the leader's than that of a load-matched expert not selected for the token (median cosine higher by 0.07, 0.05, and 0.17), and it moves the correct next token in the same direction as the leader on more tokens (91.9%, 95.5%, and 81.9%, against 80.8%, 92.5%, and 64.5%). In OLMoE and DeepSeek-MoE, co-selected experts thus apply different maps to the same token. In all three models, they agree on the prediction more than an unselected expert does, consistent with relevance-driven selection.

J Functional Evidence and Matched Training

The geometric analyses ask whether the selected context increases a candidate's residual-space advantage. The functional analyses ask the complementary question: whether the same overlapping experts still improve prediction, and whether making selected experts more geometrically diverse is enough to improve the model. The answer supports coherent overlap. Later experts contribute measurable loss reductions, actual selected experts beat load-near substitutes, matched Top-2 training outperforms Top-1, and novelty-based skipping removes more useful computation than simpler route-conditioned baselines.

J.1 Leader and later-expert value

We first locate where loss value sits inside the selected committee. Table 26 compares the damage from replacing the leader with the aggregate damage from replacing all later experts. Positive values

place more loss value on the leader, and negative values indicate that the later selected set carries the larger aggregate contribution. The pattern is model- and depth-dependent: later experts carry predictive value beyond the leader, and in deeper OLMoE layers their aggregate contribution is larger.

Table 26: Where prediction loss is carried in the selected committee. Difference is the NLL increase from replacing the leader minus that from replacing all later experts (nats, 95% interval), so positive values place more loss value on the leader.

Model	L	Difference [95% CI]	Interpretation
OLMoE	4	$0.02769\,[\ 0.01917,\ 0.03796]$	Leader larger
OLMoE	8	$-0.01323\,[-0.02000, -0.00472]$	Later set larger
OLMoE	16	$-0.08483\,[-0.09712, -0.07241]$	Later set larger
Mixtral	8	$0.01239\,[\ 0.00804,\ 0.01741]$	Leader larger
Mixtral	16	$0.00934\,[\ 0.00626,\ 0.01244]$	Leader larger
Mixtral	32	$0.02041\,[\ 0.01626,\ 0.02481]$	Leader larger
DeepSeek	7	$0.00971\,[\ 0.00535,\ 0.01412]$	Leader larger
DeepSeek	14	$0.00485\,[\ 0.00184,\ 0.00807]$	Leader larger
DeepSeek	28	$-0.00331\,[-0.00791,\ 0.00096]$	Inconclusive

J.2 NLL RECOVERY FROM ADDING LATER EXPERTS

Adding selected experts one at a time gives a finer view of that committee value. Table 27 freezes a route prefix and measures the NLL recovered by the next router-selected expert. Across the 39 additions, 28 intervals lie above zero and the remaining 11 include zero. No interval lies below zero.

Table 27: NLL recovery from adding the next selected expert j to a frozen route ($\times 10^{-3}$ nats, 95% intervals, 64 held-out contexts). Intervals lie above zero for 28 of the 39 additions and include zero for the other 11.

L	j	Recovery [95% CI]	L	j	Recovery [95% CI]
OLMoE			*DeepSeek-MoE*		
4	2	$8.80\,[\ 4.68, 12.57]$	7	2	$4.87\,[\ 3.04,\ 6.67]$
4	3	$3.50\,[\ 0.94,\ 5.93]$	7	3	$0.80\,[-0.43,\ 2.21]$
4	4	$5.81\,[\ 2.79,\ 9.97]$	7	4	$1.69\,[\ 0.64,\ 2.78]$
4	5	$2.34\,[\ 0.84,\ 4.10]$	7	5	$-0.44\,[-1.20,\ 0.29]$
4	6	$0.79\,[-0.17,\ 1.79]$	7	6	$0.34\,[-0.45,\ 1.16]$
4	7	$1.21\,[-0.02,\ 2.63]$	14	2	$5.32\,[\ 3.87,\ 6.91]$
4	8	$-0.83\,[-1.81,\ 0.16]$	14	3	$1.95\,[\ 0.93,\ 2.87]$
8	2	$9.87\,[\ 5.76, 13.23]$	14	4	$0.75\,[\ 0.05,\ 1.47]$
8	3	$6.03\,[\ 4.01,\ 8.22]$	14	5	$-0.17\,[-0.83,\ 0.46]$
8	4	$3.79\,[\ 2.56,\ 5.01]$	14	6	$0.11\,[-0.54,\ 0.75]$
8	5	$1.86\,[\ 0.85,\ 2.86]$	28	2	$11.87\,[\ 9.99, 13.83]$
8	6	$2.16\,[\ 1.15,\ 3.19]$	28	3	$3.81\,[\ 2.33,\ 5.24]$
8	7	$1.28\,[\ 0.11,\ 2.89]$	28	4	$1.22\,[\ 0.56,\ 1.85]$
8	8	$0.13\,[-0.69,\ 1.01]$	28	5	$0.30\,[-0.20,\ 0.82]$
16	2	$90.61\,[\ 83.64, 97.48]$	28	6	$0.05\,[-0.43,\ 0.60]$
16	3	$53.77\,[\ 49.47, 58.21]$	*Mixtral*		
16	4	$37.90\,[\ 34.10, 41.83]$	8	2	$3.15\,[\ 1.64,\ 4.74]$
16	5	$24.72\,[\ 22.25, 27.49]$	16	2	$2.26\,[\ 1.24,\ 3.30]$
16	6	$14.03\,[\ 11.54, 16.56]$	32	2	$4.97\,[\ 3.63,\ 6.34]$
16	7	$7.35\,[\ 6.04,\ 8.78]$			
16	8	$3.03\,[\ 2.14,\ 3.92]$			

By model, the intervals lie above zero for 17 of the 21 OLMoE additions, all three Mixtral additions, and 8 of the 15 DeepSeek-MoE additions, and include zero for the others.

Table 28: NLL advantage of the router's actual second expert over a load-near unselected expert after the same leader (64 held-out contexts).

Model	L	Unselected minus actual [95% CI]
OLMoE	4	$0.0123\,[0.0080,\ 0.0167]$
OLMoE	8	$0.0096\,[0.0042,\ 0.0140]$
OLMoE	16	$0.0483\,[0.0425,\ 0.0548]$
Mixtral	8	$0.0035\,[0.0019,\ 0.0050]$
Mixtral	16	$0.0019\,[0.0008,\ 0.0031]$
Mixtral	32	$0.0041\,[0.0030,\ 0.0051]$
DeepSeek	7	$0.0057\,[0.0036,\ 0.0078]$
DeepSeek	14	$0.0067\,[0.0051,\ 0.0086]$
DeepSeek	28	$0.0123\,[0.0105,\ 0.0142]$

J.3 Selection-specific NLL value

The next intervention asks whether the router's specific second choice matters beyond expert popularity. Table 28 holds the leader and rank-2 weight fixed, then replaces the actual second expert by a route-disjoint unselected expert with the nearest routing load. Every interval favors the actual selected expert. Replacing the whole later prefix with load-near unselected experts gives intervals above zero in all 39 OLMoE, Mixtral, and DeepSeek-MoE cells.

J.4 Functional factorial in next-token loss

The loss-level factorial is the direct functional analogue of the geometric sign test: it keeps the selected expert, rival, actual prefix, and alternative prefixes fixed, but replaces residual coverage by next-token NLL. It therefore bypasses fitted subspaces while preserving the same candidate-by-context contrast. In OLMoE, Mixtral, and DeepSeek-MoE, the functional factorial uses the 64 held-out contexts (512 tokens each) of the other NLL analyses. At layer L, the routed MoE output of every token is replaced by $\sum_{e\in P} w_e f_e(z)$ for a specified expert set P, without renormalization: context experts carry the actual weights $w_1, \ldots, w_{j-1}$ of the ranks they occupy, and the added candidate, the selected expert s or the rival r, carries the actual rank-j weight w_j. In DeepSeek-MoE, the shared-expert output is added unchanged. Layers before L are unchanged, and later layers route natively. The rival and the five alternative routes follow Appendix C, applied to each token's route in this forward pass (in Mixtral, the five experts outside the route other than the rival). Every token has a rival and five alternatives. With $N(P)$ the mean next-token NLL of a context under P and the gain $G(c \mid P) = N(P) - N(P \cup \{c\})$, the factorial quantities are

$$A_{\text{actual}} = N(S \cup \{r\}) - N(S \cup \{s\}), \qquad A_{\text{alt}} = \tfrac{1}{5}\textstyle\sum_m \big[N(W_m \cup \{r\}) - N(W_m \cup \{s\})\big],$$
$$T_s = G(s \mid S) - \tfrac{1}{5}\textstyle\sum_m G(s \mid W_m), \qquad T_r \text{ likewise}, \qquad D = A_{\text{actual}} - A_{\text{alt}} = T_s - T_r.$$

Complementarity predicts $D > 0$. Intervals use 1,000 paired bootstrap resamples over contexts, and macros average the cells with equal weights within each context. Table 29 reports every cell. Averaged over the 39 cells, the selected expert's gain is 0.0082 after its actual prefix and 0.0148 after alternative prefixes, and the rival's gain is 0.0033 and 0.0037.

In Qwen3, Gemma4, and Qwen3.6 we run the same functional factorial at the layers of Appendix E on the cached 512-token sequences of the geometric factorial that contain held-out next-token targets (91, 92, and 100 sequences). The intervention replaces the routed output of every token at layer L as above (shared-expert and dense branches unchanged, with Gemma4's per-expert scale treated as part of the expert), the rival and the five alternative routes use the eligible experts and training loads of that factorial, and the NLL is taken on next-token targets inside held-out documents. Intervals resample the 43 held-out document groups of each model with 1,000 draws. The pooled macro averages each model's 21 cells within each group. Replaying the actual route reproduces the native NLL exactly in all three models. Table 30 summarizes the 63 cells: the interaction is negative in 44, with 15 intervals below zero and one above (Qwen3.6 layer 10), and the pooled macro is $-0.0026\,[-0.0039,\ -0.0011]$. Per-expert NLL effects are about an order of magnitude smaller than in OLMoE for Qwen3 and Qwen3.6 and comparable to OLMoE for Gemma4.

Table 29: Functional factorial in next-token NLL (OLMoE, Mixtral, and DeepSeek-MoE; 64 held-out contexts). A_{actual} is the selected expert's NLL advantage over the rival after the actual prefix, T_s and T_r are each candidate's gain change when the actual prefix replaces alternatives, and $D = T_s - T_r$. Positive values favor the selected expert or actual prefix.

Model	L	j	A_{actual} [95% CI]	T_s	T_r	D [95% CI]
OLMoE	4	2	0.0087 [0.0032, 0.0136]	−0.0155	0.0028	−0.0182 [−0.0219, −0.0147]
OLMoE	4	3	0.0016 [−0.0018, 0.0048]	−0.0078	0.0028	−0.0106 [−0.0135, −0.0078]
OLMoE	4	4	0.0049 [0.0018, 0.0091]	−0.0026	0.0015	−0.0041 [−0.0075, 0.0004]
OLMoE	4	5	0.0028 [0.0011, 0.0050]	−0.0045	−0.0001	−0.0044 [−0.0063, −0.0022]
OLMoE	4	6	0.0013 [−0.0004, 0.0036]	−0.0027	−0.0004	−0.0023 [−0.0043, −0.0001]
OLMoE	4	7	0.0020 [−0.0000, 0.0048]	0.0000	−0.0002	0.0003 [−0.0016, 0.0027]
OLMoE	4	8	0.0002 [−0.0015, 0.0020]	−0.0015	−0.0004	−0.0011 [−0.0025, 0.0003]
OLMoE	8	2	0.0084 [0.0033, 0.0123]	−0.0039	−0.0003	−0.0036 [−0.0052, −0.0017]
OLMoE	8	3	0.0049 [0.0020, 0.0074]	−0.0025	−0.0003	−0.0022 [−0.0041, −0.0004]
OLMoE	8	4	0.0034 [0.0012, 0.0055]	−0.0028	−0.0009	−0.0019 [−0.0031, −0.0004]
OLMoE	8	5	0.0011 [−0.0010, 0.0030]	−0.0021	−0.0003	−0.0018 [−0.0029, −0.0006]
OLMoE	8	6	0.0012 [−0.0003, 0.0028]	−0.0011	0.0005	−0.0015 [−0.0025, −0.0006]
OLMoE	8	7	0.0007 [−0.0005, 0.0021]	−0.0007	−0.0004	−0.0003 [−0.0014, 0.0008]
OLMoE	8	8	−0.0001 [−0.0010, 0.0009]	−0.0012	−0.0008	−0.0005 [−0.0014, 0.0004]
OLMoE	16	2	0.0489 [0.0426, 0.0544]	−0.0524	−0.0089	−0.0436 [−0.0499, −0.0368]
OLMoE	16	3	0.0267 [0.0221, 0.0311]	−0.0442	−0.0059	−0.0383 [−0.0451, −0.0319]
OLMoE	16	4	0.0198 [0.0163, 0.0236]	−0.0328	−0.0020	−0.0308 [−0.0383, −0.0242]
OLMoE	16	5	0.0132 [0.0106, 0.0154]	−0.0210	0.0012	−0.0222 [−0.0265, −0.0179]
OLMoE	16	6	0.0070 [0.0048, 0.0092]	−0.0106	0.0036	−0.0142 [−0.0164, −0.0121]
OLMoE	16	7	0.0032 [0.0019, 0.0045]	−0.0054	0.0047	−0.0101 [−0.0121, −0.0083]
OLMoE	16	8	0.0008 [−0.0000, 0.0015]	0.0009	0.0053	−0.0045 [−0.0055, −0.0035]
Mixtral	8	2	0.0020 [0.0004, 0.0037]	−0.0022	−0.0018	−0.0004 [−0.0018, 0.0010]
Mixtral	16	2	0.0003 [−0.0008, 0.0014]	−0.0018	−0.0011	−0.0007 [−0.0016, 0.0003]
Mixtral	32	2	0.0019 [0.0010, 0.0027]	−0.0059	−0.0040	−0.0019 [−0.0026, −0.0013]
DeepSeek	7	2	0.0044 [0.0026, 0.0065]	−0.0023	−0.0004	−0.0020 [−0.0032, −0.0006]
DeepSeek	7	3	0.0001 [−0.0012, 0.0016]	−0.0016	−0.0006	−0.0010 [−0.0022, 0.0002]
DeepSeek	7	4	0.0007 [−0.0006, 0.0019]	−0.0011	0.0002	−0.0012 [−0.0025, 0.0002]
DeepSeek	7	5	−0.0007 [−0.0016, 0.0002]	−0.0012	−0.0005	−0.0007 [−0.0017, 0.0002]
DeepSeek	7	6	0.0000 [−0.0009, 0.0010]	−0.0000	−0.0003	0.0003 [−0.0006, 0.0011]
DeepSeek	14	2	0.0052 [0.0036, 0.0066]	−0.0017	0.0006	−0.0022 [−0.0033, −0.0012]
DeepSeek	14	3	0.0019 [0.0010, 0.0029]	−0.0014	0.0000	−0.0014 [−0.0021, −0.0007]
DeepSeek	14	4	0.0007 [−0.0002, 0.0015]	−0.0007	−0.0002	−0.0004 [−0.0011, 0.0003]
DeepSeek	14	5	−0.0003 [−0.0011, 0.0005]	−0.0010	−0.0002	−0.0007 [−0.0013, −0.0001]
DeepSeek	14	6	−0.0003 [−0.0010, 0.0003]	−0.0002	0.0002	−0.0004 [−0.0011, 0.0003]
DeepSeek	28	2	0.0105 [0.0086, 0.0123]	−0.0075	−0.0026	−0.0050 [−0.0058, −0.0040]
DeepSeek	28	3	0.0033 [0.0020, 0.0046]	−0.0058	−0.0018	−0.0040 [−0.0046, −0.0033]
DeepSeek	28	4	0.0011 [0.0005, 0.0017]	−0.0031	−0.0014	−0.0017 [−0.0022, −0.0013]
DeepSeek	28	5	0.0003 [−0.0002, 0.0007]	−0.0015	−0.0010	−0.0005 [−0.0010, −0.0000]
DeepSeek	28	6	0.0000 [−0.0005, 0.0006]	−0.0009	−0.0008	−0.0001 [−0.0006, 0.0004]
All 39 cells			0.0049 [0.0043, 0.0056]	−0.0065	−0.0004	−0.0062 [−0.0070, −0.0053]
OLMoE (21 cells)			0.0077 [0.0064, 0.0087]	−0.0102	0.0001	−0.0103 [−0.0119, −0.0086]
Mixtral (3 cells)			0.0014 [0.0006, 0.0023]	−0.0033	−0.0023	−0.0010 [−0.0017, −0.0002]
DeepSeek (15 cells)			0.0018 [0.0014, 0.0022]	−0.0020	−0.0006	−0.0014 [−0.0016, −0.0012]

The interaction is negative in every source domain. The 64 contexts of OLMoE, Mixtral, and DeepSeek-MoE come from four domains (code, dialogue, general text, and science). Short records are packed together, so 36–40 contexts contain two or three records, and the three tokenizers pack different records into a context. Each context is therefore assigned, separately for each model, to the domain that holds most of its next-token targets (8–24 contexts per domain and model). Averaged over the 39 cells with equal weights, D is $-0.0124\,[-0.0137, -0.0109]$ in code, $-0.0065\,[-0.0074, -0.0055]$ in dialogue, $-0.0046\,[-0.0061, -0.0028]$ in general text, and $-0.0046\,[-0.0054, -0.0038]$ in science, with intervals resampling contexts within each model. All twelve model-by-domain macros are negative, although the Mixtral intervals for code and science include zero. Keeping only contexts whose majority domain holds at least 90% of the targets gives the same result. The evaluation text of the newer models is general-domain only.

Table 30: Functional factorial in Qwen3, Gemma4, and Qwen3.6. Equal-cell macros cover positions $j = 2, \ldots, 8$ by layer, all 21 cells per model, and the pooled 63 cells (n: cells). Sign counts are $D < 0$ / interval below zero / interval above zero.

Model	L	n	A_{actual} [95% CI]	T_s	T_r	D [95% CI]	Sign
Qwen3	all	21	0.0011 [0.0006, 0.0015]	−0.0002	0.0000	−0.0003 [−0.0007, 0.0001]	11/2/0
Qwen3	12	7	0.0006 [−0.0001, 0.0013]	0.0001	0.0002	−0.0001 [−0.0008, 0.0006]	2/1/0
Qwen3	24	7	0.0005 [−0.0004, 0.0015]	0.0000	0.0001	−0.0001 [−0.0008, 0.0008]	4/0/0
Qwen3	41	7	0.0020 [0.0014, 0.0028]	−0.0008	−0.0002	−0.0006 [−0.0014, 0.0001]	5/1/0
Gemma4	all	21	0.0014 [−0.0015, 0.0040]	−0.0202	−0.0134	−0.0069 [−0.0103, −0.0030]	17/9/0
Gemma4	8	7	−0.0000 [−0.0087, 0.0079]	−0.0298	−0.0213	−0.0085 [−0.0193, 0.0038]	5/1/0
Gemma4	15	7	0.0012 [−0.0028, 0.0044]	−0.0221	−0.0142	−0.0079 [−0.0124, −0.0039]	6/4/0
Gemma4	26	7	0.0031 [0.0014, 0.0044]	−0.0088	−0.0046	−0.0042 [−0.0061, −0.0023]	6/4/0
Qwen3.6	all	21	0.0010 [−0.0001, 0.0021]	−0.0006	−0.0000	−0.0005 [−0.0011, 0.0001]	16/4/1
Qwen3.6	10	7	0.0017 [−0.0003, 0.0037]	−0.0003	−0.0008	0.0005 [−0.0007, 0.0021]	4/0/1
Qwen3.6	20	7	0.0001 [−0.0012, 0.0016]	−0.0004	0.0003	−0.0008 [−0.0016, 0.0001]	5/1/0
Qwen3.6	34	7	0.0011 [−0.0005, 0.0026]	−0.0009	0.0004	−0.0013 [−0.0024, −0.0003]	7/3/0
Pooled	all	63	0.0012 [0.0002, 0.0021]	−0.0070	−0.0045	−0.0026 [−0.0039, −0.0011]	44/15/1

Table 31: NLL increase (nats) from truncating OLMoE and DeepSeek-MoE routes to their top-j experts, with 95% paired bootstrap intervals over 64 held-out contexts.

	OLMoE		
j	$L = 4$	$L = 8$	$L = 16$
1	0.0216 [0.0163, 0.0270]	0.0251 [0.0208, 0.0292]	0.2314 [0.2144, 0.2481]
2	0.0128 [0.0085, 0.0187]	0.0153 [0.0122, 0.0189]	0.1408 [0.1288, 0.1517]
3	0.0093 [0.0045, 0.0161]	0.0092 [0.0068, 0.0123]	0.0870 [0.0785, 0.0961]
4	0.0035 [0.0010, 0.0063]	0.0054 [0.0033, 0.0077]	0.0491 [0.0434, 0.0555]
5	0.0012 [−0.0006, 0.0030]	0.0036 [0.0017, 0.0061]	0.0244 [0.0200, 0.0288]
6	0.0004 [−0.0009, 0.0018]	0.0014 [−0.0004, 0.0035]	0.0104 [0.0086, 0.0126]
7	−0.0008 [−0.0018, 0.0001]	0.0001 [−0.0007, 0.0009]	0.0030 [0.0022, 0.0040]
	DeepSeek-MoE		
j	$L = 7$	$L = 14$	$L = 28$
1	0.0073 [0.0048, 0.0098]	0.0080 [0.0060, 0.0099]	0.0173 [0.0145, 0.0206]
2	0.0024 [0.0006, 0.0041]	0.0026 [0.0014, 0.0038]	0.0054 [0.0036, 0.0071]
3	0.0016 [0.0004, 0.0029]	0.0007 [−0.0001, 0.0016]	0.0016 [0.0006, 0.0025]
4	−0.0001 [−0.0011, 0.0010]	−0.0001 [−0.0007, 0.0006]	0.0004 [−0.0002, 0.0010]
5	0.0003 [−0.0004, 0.0012]	0.0001 [−0.0006, 0.0007]	0.0001 [−0.0004, 0.0006]

J.5 Route truncation

Route truncation provides a simple absolute-value check on the later selected experts. Table 31 reports the NLL increase when OLMoE or DeepSeek-MoE keeps only its top-j selected experts at one layer, with the retained router weights unchanged. Keeping only the leader in Mixtral costs $0.0032\,[0.0016,\ 0.0047]$, $0.0023\,[0.0013,\ 0.0033]$, and $0.0050\,[0.0037,\ 0.0063]$ NLL at layers 8, 16, and 32.

J.6 Matched-compute Top-1 versus Top-2

The matched-compute training comparison tests whether two selected experts help because they add active capacity or because routing a committee is useful. The six-layer Top-1 and Top-2 models match active intermediate capacity exactly (1,024 units), while total parameters (27,108,608 vs. 27,114,752) and active FLOPs per token (36,777,984 vs. 36,790,272) differ by less than 0.04%. Top-1 uses four experts of width 1,024, and Top-2 uses eight experts of width 512. Each seed trains for 20 million tokens (2,442 optimizer updates).

Table 32: Matched-compute Top-1 versus Top-2 validation loss. Difference is Top-1 minus Top-2, so positive values favor Top-2.

Seed	Top-1	Top-2	Difference
0	5.229414	5.124888	0.104527
1	5.228661	5.128909	0.099751
2	5.217812	5.117210	0.100602
Mean $\pm$ SD	5.225296 ± 0.006492	5.123669 ± 0.005944	0.101627 ± 0.002547

Table 33: Validation loss over training (mean over three seeds) and paired Top-1 minus Top-2 difference (mean $\pm$ SD).

Tokens	Top-1	Top-2	Difference
1M	7.9348	7.9242	0.0106 ± 0.0077
2M	6.6587	6.6290	0.0298 ± 0.0146
5M	5.8934	5.8417	0.0517 ± 0.0010
10M	5.4535	5.3546	0.0988 ± 0.0050
20M	5.2253	5.1237	0.1016 ± 0.0025

Across the three seeds (Table 32), Top-2 has lower validation loss in every seed, with a paired difference of 0.1016 ± 0.0025 nats (mean $\pm$ sample SD). Validation loss is the mean next-token cross-entropy in nats on the held-out split of the training corpus, tokenized with the Pythia tokenizer. Table 33 follows the gap over training.

J.7 Training with an output-complementarity objective

This training intervention tests whether forcing output-level diversity can turn a matched Top-2 route into a better predictor. In the Top-2 setting, the objective adds the gate-weighted fraction of each later expert's output that lies in the span of higher-ranked selected experts' outputs, with weight 0.05, 0.2, or 1.0. Every other setting matches the baseline. Table 35 reports validation loss, the NLL increase when the later expert is removed at the last MoE layer (tail gain), and the relative change of the final logits under the same removal (tail lift). Table 34 measures the objective's target on the 448,256 validation tokens per layer: the squared cosine between the two selected experts' outputs, which for Top-2 equals the penalized span fraction, averaged over the six MoE layers. At weight 0.05 the regularizer lowers it in every layer and seed, reducing the six-layer mean by 38% (24–57% in individual layers), while the later expert's router weight moves by only 1.7%. Weights 0.2 and 1.0 reduce the six-layer mean by 70% and 92%, again lowering it in every layer and seed. The geometry moves as designed, but validation loss does not improve: the paired change is within seed variation at weights 0.05 and 0.2 and positive in all three seeds at weight 1.0, and the tail gain falls further. At weight 0.05, the reduction of the penalized fraction shrinks during training (0.19 at one million tokens, 0.045 at ten million, 0.036 at twenty million) because the baseline becomes more complementary on its own. The small early validation-loss cost at weight 0.05 (+0.0023 at one million and +0.0038 at five million tokens, all three seeds) is no longer consistent across seeds at ten million (+0.0025) and vanishes by twenty million (-0.0001).

J.8 Per-token expert skipping

Expert skipping turns coherent overlap into an inference-time compression question. A novelty proxy would predict that dropping the least novel selected expert is safer than following the router weight or output magnitude. Tables 36 and 37 show the opposite pattern in the models where there is room to choose: novelty-based criteria consistently discard useful route members.

The skipping protocol is as follows. At one layer, every token of the 64 held-out contexts of the other NLL analyses keeps $k - m$ of its k selected experts (OLMoE $m = 1, 2, 4$, DeepSeek-MoE $m = 1, 2, 3$, Mixtral $m = 1$). The kept experts carry their router weights, shared experts are added unchanged, and later layers route natively. The skipped experts are chosen per token by six criteria:

Table 34: Effects of the complementarity objective at three weights. The upper block gives final-checkpoint values for each arm, and the lower block gives paired differences (regularized minus baseline for the same seed), both as mean ± SD over three seeds. Penalized fraction, signed cosine, and rank-2 router weight are six-layer means. Tail gain is the NLL cost of removing the later expert at the last MoE layer. Every paired difference has the same sign in all three seeds except validation loss at weights 0.05 and 0.2.

Arm	Penalized fraction	Signed cosine	Rank-2 router weight	Validation loss	Tail gain
Final checkpoint, mean ± SD over three seeds					
Baseline	0.0937 ± 0.0034	0.2580 ± 0.0069	0.3178 ± 0.0026	5.1237 ± 0.0059	0.0114 ± 0.0012
Weight 0.05	0.0580 ± 0.0012	0.1890 ± 0.0029	0.3124 ± 0.0050	5.1236 ± 0.0026	0.0102 ± 0.0007
Weight 0.2	0.0278 ± 0.0022	0.1136 ± 0.0038	0.3112 ± 0.0047	5.1268 ± 0.0025	0.0098 ± 0.0008
Weight 1.0	0.0079 ± 0.0003	0.0392 ± 0.0004	0.3111 ± 0.0009	5.1336 ± 0.0038	0.0087 ± 0.0006
Paired difference from the baseline (same seed), mean ± SD					
Weight 0.05	-0.0357 ± 0.0045	-0.0690 ± 0.0074	-0.0053 ± 0.0038	-0.0001 ± 0.0065	-0.0011 ± 0.0009
Weight 0.2	-0.0659 ± 0.0056	-0.1444 ± 0.0107	-0.0066 ± 0.0031	0.0031 ± 0.0083	-0.0016 ± 0.0008
Weight 1.0	-0.0858 ± 0.0033	-0.2188 ± 0.0067	-0.0067 ± 0.0018	0.0099 ± 0.0097	-0.0027 ± 0.0009

Table 35: Matched Top-2 training with and without the complementarity objective at three weights (final checkpoint, 20 million tokens).

Arm	Seed	Validation loss	Tail gain	Tail lift
Baseline	0	5.1249	0.0100	0.0561
Baseline	1	5.1289	0.0122	0.0590
Baseline	2	5.1172	0.0118	0.0561
Weight 0.05	0	5.1265	0.0096	0.0528
Weight 0.05	1	5.1216	0.0100	0.0530
Weight 0.05	2	5.1225	0.0110	0.0553
Weight 0.2	0	5.1273	0.0090	0.0505
Weight 0.2	1	5.1240	0.0097	0.0510
Weight 0.2	2	5.1289	0.0106	0.0511
Weight 1.0	0	5.1325	0.0082	0.0478
Weight 1.0	1	5.1304	0.0085	0.0476
Weight 1.0	2	5.1378	0.0094	0.0485

R, lowest router weight (Top-$(k-m)$ truncation); *O*, smallest weighted output norm $\|w_e f_e(z)\|$; *G-fit*, lowest single-expert fit $1-\rho_x(\{e\})$; *G-LOO*, lowest leave-one-out novelty $q_x(e \mid E_x \setminus \{e\})$, removed greedily one expert at a time; *S*, smallest mean chordal distance between the expert's basis and the bases of the other selected experts, also greedy; and *Rand*, uniform (mean of three seeds). G-fit, G-LOO, and S use the uncentered rank-128 factorial bases of Section 3.1. The outcome is the NLL increase over the full route per context, and intervals use 1,000 paired bootstrap resamples over contexts. We call one criterion better on a model when the paired-difference interval favors it in at least two thirds of that model's layer-budget cells. Replaying the full route reproduces the native NLL exactly, and R reproduces the corresponding truncation costs in Table 31 and the Mixtral values of Appendix J.5.

Table 36 reports every cell. G-LOO is worse than R in 9 of 9 OLMoE cells, 8 of 9 DeepSeek-MoE cells, and 2 of 3 Mixtral cells, and worse than O in 8, 6, and 2. It is better than random skipping everywhere. R and O are comparable in all three models (O is better at OLMoE layer 16). S is worse than R in every OLMoE and DeepSeek-MoE cell and worse than random skipping in 8 of 9 cells of each model, by $0.0126\,[0.0104,\ 0.0148]$ and $0.0054\,[0.0040,\ 0.0070]$ NLL on average. In these two models, S skips the leader in 8–66% of tokens and G-LOO in 3–23% (22–25% in Mixtral), whereas R never does. With the leader never skipped (Table 37, which omits Mixtral because with $k=2$ every criterion then skips the same expert as R), G-LOO* remains worse than R in 7 of 9 OLMoE and 6 of 9 DeepSeek-MoE cells ($0.0093\,[0.0082,\ 0.0104]$ and $0.0012\,[0.0007,\ 0.0016]$), and S* remains worse than random skipping among ranks $2, \ldots, k$. In OLMoE, G-LOO* beats random skipping on average (by $0.0030\,[0.0016,\ 0.0042]$) but only in the three layer-16 cells. Rescaling the kept weights

Table 36: Skipping selected experts by router weight, output size, and geometric novelty. Values are NLL increases over the full route ($\times 10^{-3}$ nats, 64 held-out contexts, kept experts at router weights) when skipping m of k selected experts at layer L. The final column is the paired G-LOO−R difference with 95% interval. In Mixtral ($k = 2$), S equals R and G-fit equals G-LOO.

Model	L	m	R	O	G-fit	G-LOO	S	Rand	G-LOO−R [95% CI]
OLMoE	4	1	−0.8	0.1	1.5	1.2	17.3	7.3	2.0 [0.1, 4.1]
OLMoE	4	2	0.4	−0.3	1.8	4.2	29.7	14.9	3.9 [1.3, 6.3]
OLMoE	4	4	3.5	2.8	7.0	13.7	51.3	32.8	10.2 [6.0, 14.6]
OLMoE	8	1	0.1	0.1	0.4	2.2	5.8	3.6	2.1 [0.9, 3.3]
OLMoE	8	2	1.4	2.2	3.0	4.4	11.7	6.9	3.0 [1.3, 4.6]
OLMoE	8	4	5.4	5.1	8.3	12.7	23.8	17.2	7.3 [4.5, 10.2]
OLMoE	16	1	3.0	2.3	5.0	16.1	28.6	21.9	13.1 [10.0, 16.1]
OLMoE	16	2	10.4	7.7	16.7	37.0	67.3	49.5	26.6 [22.1, 31.9]
OLMoE	16	4	49.1	35.7	73.4	97.8	162.0	130.2	48.7 [41.3, 57.2]
OLMoE mean			8.1	6.2	13.0	21.0	44.2	31.6	13.0 [11.2, 14.9]
DeepSeek	7	1	0.3	−0.1	0.5	0.9	7.5	3.1	0.6 [−0.4, 1.6]
DeepSeek	7	2	−0.1	0.7	1.0	1.6	14.4	7.1	1.7 [0.4, 3.0]
DeepSeek	7	3	1.6	0.3	2.9	3.6	18.5	11.6	2.0 [0.5, 3.5]
DeepSeek	14	1	0.1	0.2	0.9	1.0	3.8	2.5	0.9 [0.2, 1.6]
DeepSeek	14	2	−0.1	0.7	1.0	1.4	7.9	5.4	1.5 [0.4, 2.7]
DeepSeek	14	3	0.7	1.0	2.4	2.5	11.7	9.2	1.8 [0.5, 3.2]
DeepSeek	28	1	0.1	0.6	0.6	1.3	10.6	3.7	1.2 [0.5, 1.9]
DeepSeek	28	2	0.4	0.2	1.7	3.2	14.1	7.5	2.9 [1.7, 4.0]
DeepSeek	28	3	1.6	0.9	3.7	5.9	23.9	13.9	4.3 [2.7, 5.9]
DeepSeek mean			0.5	0.5	1.7	2.4	12.5	7.1	1.9 [1.4, 2.4]
Mixtral	8	1	3.2	3.5	4.0	4.0	3.2	8.8	0.8 [−0.9, 2.7]
Mixtral	16	1	2.3	2.0	3.4	3.4	2.3	7.0	1.1 [0.2, 2.1]
Mixtral	32	1	5.0	5.1	7.4	7.4	5.0	22.0	2.5 [1.4, 3.4]
Mixtral mean			3.5	3.5	4.9	4.9	3.5	12.6	1.5 [0.7, 2.2]

to the full route's total weight gives the same comparisons in OLMoE and DeepSeek-MoE, except that G-LOO* and R become comparable in DeepSeek-MoE. In Mixtral, G-LOO, O, and R become comparable. Across tokens, leave-one-out novelty and weighted output norm correlate at Spearman 0.21–0.56.

Geometric novelty and the effect of one expert. We also ask whether geometric novelty predicts, token by token, how much a selected expert changes the prediction. In the 39 cells of the functional factorial (Appendix J.4, with 64 held-out contexts of 511 loss positions), we take for every token the novelty $q_x(s \mid S)$ of the selected rank-j expert under its actual prefix and the reduction G_x of the token's next-token NLL when that expert is added to the frozen route with its router weight (the per-token form of $G(s \mid S)$ in Appendix J.4, positive when the expert lowers the loss). We compute Spearman correlations across the tokens of each context, averaged over contexts and then over cells (Table 38, with 95% intervals from 1,000 context bootstrap resamples). Because the intervention covers every token of the context, G_x at earlier layers also carries effects that reach the token through attention in later layers. Its sign varies from token to token: G_x is positive on 37–56% of the tokens of a cell, and on 37–53% at the last analyzed layers. Novelty does not predict the signed change ($-0.001\,[-0.003,\ 0.001]$), and the router weight barely does ($0.007\,[0.005,\ 0.009]$). The size of the change, $|G_x|$, correlates with the router weight (0.129) and the weighted output norm (0.151), and hardly with novelty (0.024). At the last analyzed layers, where only the final norm and the unembedding follow, the correlations are 0.271 for the router weight, 0.334 for the output norm, and 0.052 for novelty. Novelty thus carries little information about how much a selected expert changes a given prediction, consistent with the skipping results above. Across the 64 contexts of a cell, the geometric and the NLL interactions are weakly aligned (Spearman $0.083\,[0.007,\ 0.156]$), mainly in OLMoE (0.131). The per-model intervals of DeepSeek-MoE and Mixtral include zero.

Table 37: Expert skipping with the leader never skipped. G-fit*, G-LOO*, S*, and Rand* choose among ranks $2, \ldots, k$, and R and O are as in Table 36 ($\times 10^{-3}$ nats). Mixtral is omitted because with $k = 2$ every criterion skips the rank-2 expert, as R does.

Model	L	m	R	O	G-fit*	G-LOO*	S*	Rand*	G-LOO*−R [95% CI]
OLMoE	4	1	−0.8	0.1	1.4	0.3	4.3	0.6	1.1 [−0.6, 2.9]
OLMoE	4	2	0.4	−0.3	1.7	4.0	8.6	4.8	3.6 [1.6, 5.8]
OLMoE	4	4	3.5	2.8	4.8	9.5	13.2	10.4	6.0 [3.1, 9.5]
OLMoE	8	1	0.1	0.1	0.3	1.8	4.8	1.7	1.6 [0.5, 2.8]
OLMoE	8	2	1.4	2.2	2.5	3.3	8.6	4.7	1.9 [−0.1, 3.8]
OLMoE	8	4	5.4	5.1	6.3	11.3	15.1	11.7	5.8 [2.7, 9.1]
OLMoE	16	1	3.0	2.3	4.9	12.7	27.0	15.9	9.7 [8.1, 11.2]
OLMoE	16	2	10.4	7.7	15.7	30.4	58.3	36.7	20.0 [17.5, 22.7]
OLMoE	16	4	49.1	35.7	65.2	82.9	125.5	96.5	33.8 [29.9, 37.7]
OLMoE mean			8.1	6.2	11.4	17.4	29.5	20.3	9.3 [8.2, 10.4]
DeepSeek	7	1	0.3	−0.1	0.5	0.1	2.9	1.1	−0.2 [−1.3, 0.9]
DeepSeek	7	2	−0.1	0.7	1.3	1.3	3.8	2.3	1.4 [0.3, 2.6]
DeepSeek	7	3	1.6	0.3	2.4	1.8	4.8	3.2	0.2 [−1.2, 1.5]
DeepSeek	14	1	0.1	0.2	0.4	0.5	1.5	1.2	0.4 [−0.3, 1.2]
DeepSeek	14	2	−0.1	0.7	0.8	1.2	3.6	2.8	1.3 [0.3, 2.3]
DeepSeek	14	3	0.7	1.0	1.7	2.0	5.5	4.0	1.3 [0.5, 2.3]
DeepSeek	28	1	0.1	0.6	0.5	1.0	5.3	2.3	1.0 [0.3, 1.7]
DeepSeek	28	2	0.4	0.2	1.4	2.4	7.7	4.9	2.1 [1.1, 3.0]
DeepSeek	28	3	1.6	0.9	2.8	4.6	10.7	8.8	3.0 [1.6, 4.4]
DeepSeek mean			0.5	0.5	1.3	1.7	5.1	3.4	1.2 [0.7, 1.6]

Table 38: Geometric novelty and one-expert NLL change. Within-context rows report Spearman correlations between selected rank-j expert quantities and the NLL reduction G_x from adding that expert to a frozen route. Values are averaged over contexts and cells (OLMoE 21, Mixtral 3, DeepSeek-MoE 15, all 39 cells) with 95% context-bootstrap interval for the pooled estimate. Signed change: difference is the novelty correlation minus the router-weight correlation. The context-level row correlates geometric and NLL interactions across the 64 contexts of each cell.

Measure	OLMoE	Mixtral	DeepSeek	All [95% CI]
Signed change: novelty q	−0.002	0.005	−0.000	−0.001 [−0.003, 0.001]
Signed change: router weight	0.010	0.017	0.000	0.007 [0.005, 0.009]
Signed change: difference	−0.012	−0.012	−0.000	−0.007 [−0.010, −0.005]
Size $\lvert G\rvert$: novelty q	0.013	0.003	0.044	0.024 [0.019, 0.030]
Size $\lvert G\rvert$: router weight	0.117	0.076	0.158	0.129 [0.121, 0.138]
Size $\lvert G\rvert$: output norm	0.130	0.088	0.194	0.151 [0.145, 0.158]
Context level: geometric vs. NLL D	0.131	−0.042	0.043	0.083 [0.007, 0.156]

J.9 THE FUNCTIONAL INTERACTION OVER TRAINING

We repeat the functional factorial of Appendix J.4 in the matched Top-2 model (six MoE layers, eight experts, $d = 256$) at initialization and after 1, 5, 10, and 20 million training tokens, for three seeds. Initial weights are regenerated from each seed, and training them for 123 updates reproduces the 1M-token checkpoint exactly. At one MoE layer at a time, the routed output of every token of the 1,751 held-out validation sequences (256 tokens each) is replaced as in Appendix J.4, with the actual router weights, and later layers route natively. The selected candidate is the rank-2 expert and the rival the highest-scoring unselected expert. As in Mixtral, the five alternative contexts are the experts outside the route other than the rival. Table 39 reports the six-layer macro with the three seeds pooled. At initialization, when the loss barely depends on the experts (the selected expert lowers NLL by 0.0003 after its leader, against 0.0294 after 20 million tokens), $D = 0.00009\,[-0.000003,\ 0.00019]$, below 0.0002 in every seed. After one million tokens, $D = -0.0031\,[-0.0032,\ -0.0030]$, and after 20 million it is $-0.0208\,[-0.0214,\ -0.0202]$, below zero in each seed. The complementarity objective moves the interaction in the intended direction as its weight grows. At 20 million tokens, $D =$

Table 39: Functional interaction over training in the matched Top-2 model (six-layer macro over 1,751 held-out validation sequences with three seeds pooled, and 95% paired bootstrap intervals over sequences). Compl. is the complementarity objective with the given weight, trained from the same initialization as Top-2. ΔD is the difference from Top-2 at the same checkpoint, so Top-2 is the reference arm for this column.

Arm	Tokens	A_{actual}	T_s	T_r	D [95% CI]	ΔD vs. Top-2 [95% CI]
Top-2	0	−0.0002	0.0006	0.0005	0.0001 [−0.0000, 0.0002]	
	1M	0.0019	−0.0062	−0.0031	−0.0031 [−0.0032, −0.0030]	
	5M	0.0075	−0.0662	−0.0456	−0.0206 [−0.0212, −0.0199]	*reference arm*
	10M	0.0109	−0.0751	−0.0524	−0.0227 [−0.0234, −0.0220]	
	20M	0.0150	−0.0795	−0.0587	−0.0208 [−0.0214, −0.0202]	
Compl. 0.05	1M	0.0016	−0.0057	−0.0031	−0.0026 [−0.0027, −0.0025]	0.0005 [0.0005, 0.0005]
	5M	0.0069	−0.0575	−0.0401	−0.0174 [−0.0180, −0.0168]	0.0032 [0.0030, 0.0033]
	10M	0.0108	−0.0675	−0.0476	−0.0199 [−0.0205, −0.0192]	0.0029 [0.0027, 0.0031]
	20M	0.0144	−0.0719	−0.0535	−0.0184 [−0.0190, −0.0178]	0.0024 [0.0022, 0.0026]
Compl. 0.2	1M	0.0016	−0.0049	−0.0029	−0.0020 [−0.0021, −0.0020]	0.0011 [0.0010, 0.0011]
	5M	0.0073	−0.0529	−0.0373	−0.0156 [−0.0160, −0.0152]	0.0050 [0.0046, 0.0054]
	10M	0.0105	−0.0596	−0.0429	−0.0167 [−0.0171, −0.0163]	0.0060 [0.0056, 0.0064]
	20M	0.0141	−0.0634	−0.0479	−0.0154 [−0.0158, −0.0150]	0.0054 [0.0051, 0.0057]
Compl. 1.0	1M	0.0015	−0.0035	−0.0021	−0.0015 [−0.0015, −0.0014]	0.0017 [0.0016, 0.0017]
	5M	0.0078	−0.0548	−0.0401	−0.0147 [−0.0150, −0.0144]	0.0059 [0.0054, 0.0063]
	10M	0.0118	−0.0614	−0.0450	−0.0164 [−0.0167, −0.0160]	0.0064 [0.0058, 0.0069]
	20M	0.0160	−0.0622	−0.0485	−0.0137 [−0.0141, −0.0133]	0.0071 [0.0067, 0.0075]

-0.0184, -0.0154, and -0.0137 at weights 0.05, 0.2, and 1.0, differences from Top-2 of 0.0024, 0.0054, and 0.0071 with every interval above zero (Table 39). The sign therefore emerges with training as the experts become predictive and remains negative even when output complementarity is explicitly encouraged.

K Interpretation and Reproducibility

K.1 What the evidence supports

The appendix supports a coherent-overlap interpretation of sparse MoE routing. The router selects experts from a shared token-relevant region of representation space. Within that region, earlier selected experts cover part of the residual signal that makes later selected experts attractive, and the selected experts remain functionally valuable because their expert maps differ and their loss contributions are selection-specific. The same candidate-by-context sign appears in residual geometry and in next-token NLL. Output-complementarity training and novelty-based skipping show that geometric diversity alone is an insufficient optimization proxy.

The rank-128 geometry is a linear diagnostic on router-input representations. It describes the inputs routed to each expert, not the expert's function, so input overlap alone does not rule out functional specialization. It is deliberately paired with loss-level interventions, output/Jacobian/logit comparisons, matched training, and expert skipping. The fitted-basis control in Appendix C.2 explains why the NLL factorial is the sign test that does not rely on fitted subspaces: it keeps the candidate, rival, actual prefix, and alternative prefixes fixed while removing subspace fitting from the estimand. The geometry survey, geometric factorial, and functional factorial cover all six models, while the richer functional analyses and expert-skipping experiments cover OLMoE, Mixtral, and DeepSeek-MoE. The matched-training comparison also controls active intermediate capacity, parameters, and active FLOPs to within 0.04% in the tested six-layer model.

The score-ordered prefix is an analytical device for a parallel Top-k MoE. It creates a well-defined context in which to ask whether a candidate becomes more valuable after its actual co-selected predecessors than after matched alternatives. The leave-one-out, random-order, permutation-averaged,

and residual-matched variants test the same interaction under different admissible contexts and preserve the conclusion.

K.2 REPRODUCIBILITY

The factorial analyses of OLMoE, Mixtral, and DeepSeek-MoE use seed zero, source-record connected-component splits, train-only expert fits, and context-level paired bootstrap intervals. The supplementary material contains configurations, model revisions, corpus manifests, compact result tables, figure scripts, and analysis code. The reported GPU experiments were conducted on a server equipped with eight NVIDIA GeForce RTX 4090 GPUs (24 GB each), and each matched-compute Top-1/Top-2 training run used a single GPU.

Centered global bases and uncentered factorial bases play different roles in the evidence chain. Centered bases describe how expert input subspaces are arranged within a layer. Uncentered factorial bases define the residual signal used in the candidate-by-context interaction.