BEYOND THE PREVIOUS LAYER: INCREMENTAL PREDICTIVE VALUE OF ROUTING HISTORY IN FROZEN SPARSE MIXTURE-OF-EXPERTS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Routing in a sparse mixture-of-experts (MoE) transformer is usually examined one layer at a time, yet a single token passes through many routers in sequence. We ask how much *older* expert selections still predict once the full score vector of the immediately previous router is known. Answering that takes a conditional measurement: we hold that score vector fixed, add older selections, and compare against a distribution-matched control in which history rows are permuted across tokens, which separates alignment from added width. On a frozen grid of four pretrained MoE language models at four depths each, older aligned selections raise held-out R^2 for the next router's scores by 0.0455 on average (95% CI $[0.0435, 0.0474]$, positive in all 16 cells), while the misaligned control recovers none of this gain. The same history raises next-layer expert-set recall from 0.6350 to 0.6681 ($+3.31$ pp), and both effects reproduce on an unseen corpus at four token positions ($+3.29$ pp). The gains vary by nearly an order of magnitude across architectures, and a headroom-normalized diagnostic computed on separate text ranks the cells by realized gain (Spearman $\rho = 0.744$); supplying history to the top half of the cells under that ranking retains about 71% of the full average recall gain. Two conditions bound the scope: widening the conditioning state to three recent score vectors leaves only a small residual, carried by the narrowest router in the grid, and against a strong gate-reading reranker the history-specific complementarity is small and the runtime selector we tested does not convert it. Cross-layer routing history is a real conditional signal whose magnitude is an architectural property rather than a uniform feature of sparse MoE routing.

1 INTRODUCTION

A token in a sparse mixture-of-experts transformer is routed many times. At every MoE layer a small router scores the experts and a top-k rule selects a few of them, so a single forward pass leaves behind a trajectory of selections, one per routed layer. Analyses of these trajectories almost always proceed one layer at a time: which experts a layer prefers, how balanced its load is, how its scores respond to language or domain. What the trajectory contributes as a trajectory is harder to establish, because the immediately previous router state is itself a rich continuous object. Once the full score vector produced by the previous router is known, it is not obvious that anything earlier remains informative.

That conditioning requirement separates our question from existing cross-layer work (§2), which establishes that *some* structure spans MoE depth without holding the previous score vector fixed. A signal that exists unconditionally can still be absorbed entirely by recent state; whether it is, and by how much, has to be measured.

This paper measures precisely that question. Let $\mathbf{g}_m \in \mathbb{R}^E$ denote the router score vector at routed layer m over its E experts, and $S_m \in \{0,1\}^E$ the expert set the model actually selected there. We hold $G = \mathbf{g}_{m-1}$ fixed as the conditioning state, take $H = [S_{m-4}; S_{m-3}; S_{m-2}]$ as older selection history, and compare three predictors of the routing decision at layer m: G alone, $[G; H]$, and $[G; \pi(H)]$, where π permutes history rows across samples. The third arm is what makes the

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.

comparison work: it matches $[G; H]$ in width, active coordinate count and feature marginals, differing only in whether the history of each token belongs to that token, which separates alignment from width where a zero-padded $[G; \mathbf{0}]$ cannot (§3).

A probe increment has a precise reading: a probe that improves when given H shows that older selections carry information decodable by that probe for that target, a statement about a prediction task rather than about conditional mutual information, causal memory or a Markov order (§3 fixes this scope for the whole paper). Predicting a model's native expert set is likewise a different objective from finding a better routing policy, lowering language-modelling loss, or making inference faster.

Contributions. A conditional measurement protocol with a matched control. We define a measurement of routing history that holds the complete score vector of the previous layer fixed, and we separate the direct increment ($[G; H]$ versus G) from the alignment-specific increment ($[G; H]$ versus $[G; \pi(H)]$).

Incremental predictive value, and its reproduction. On 4 architectures $\times$ 4 depths the direct increment is $+0.0455$ in held-out R^2 (95% CI $[0.0435, 0.0474]$), positive in 16/16 cells, ranging from $+0.011$ (OLMoE) to $+0.154$ (JetMoE). It transfers to the discrete outcome a routing system acts on, raising next-layer expert-set Recall@native-k from 0.6350 to 0.6681 ($+3.31$ pp, 95% CI $[3.01, 3.61]$ pp), and both effects reproduce on C4 at four token positions ($+3.29$ pp, 95% CI $[2.97, 3.61]$ pp).

Heterogeneity, a diagnostic use, and the conditions that bound it. The gains vary by nearly an order of magnitude across architectures, and a headroom-normalized diagnostic $\mathrm{HNV} = (R - B)/(1 - B)$ computed on separate anchors ranks the cells by realized gain at $\rho = 0.744$ (95% CI $[0.565, 0.859]$). Ranking by it and supplying history to 8 of 16 cells retains about 71% of the average recall gain of supplying it everywhere, above the same-budget random mean, though simpler rules score similarly. Two measured conditions bound the scope: widening the conditioning state to three recent score vectors leaves a pooled residual of only $+0.0031$, positive in 5 of 16 cells and carried by JetMoE, and against a strong gate-reading reranker the history-specific oracle complementarity is 0.56 pp, which the runtime selector we tested does not convert.

Code, configurations and frozen result tables are available in an anonymous repository, accompanied by an anonymous interactive view of the routing captures and the study graph.[1]

2 Related work

Sparse MoE routing. Sparsely gated MoE layers replace a dense feed-forward block with many experts and a router activating a few of them (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022), and are now standard in open language models including OLMoE (Muennighoff et al., 2024), JetMoE (Shen et al., 2024), the Qwen MoE line (Bai et al., 2023), DeepSeek-MoE (Dai et al., 2024) and Mixtral (Jiang et al., 2024). Work on routing design concerns how to choose experts more effectively, through routing objectives, soft assignment (Puigcerver et al., 2024), or scaling relationships between routed capacity and loss (Clark et al., 2022). Our models are frozen and we propose no routing rule; the question is what the past selections of a trained router predict about its subsequent ones.

What has been measured across layers. Several studies already examine MoE depth, which is why our gap must be stated narrowly. Li et al. (2026) decompose the input of a router into contributions from earlier attention and MoE components, reporting that MoE outputs usually dominate subsequent routing and that some components influence routing many layers later, which attributes the current decision to upstream computations. Bandarkar et al. (2026) measure layer-wise routing divergence between languages, find cross-lingual alignment concentrated in middle layers, then steer middle-layer experts to improve multilingual accuracy; their conditioning variable is language. Li & Zhou (2025) concatenate routing weights of all layers into a sentence embedding, establishing that router

[1] Anonymous code and reproducibility artifact: https://anonymous.4open.science/r/moe-routing-anonymous-artifact-FD5A. Anonymous interactive expert-node visualization: https://anonymous-artifact.net/. All quantitative evidence for the claims here is contained in this PDF; the site is a navigational aid.

scores carry usable information, but pooling depth rather than separating it. Liang et al. (2026) measure *local routing consistency* across consecutive tokens on 20 MoE models and relate it to expert-cache hit rates; their axis is the token axis. Lo et al. (2024) and Su et al. (2026) characterize experts and their gate statistics within layers, and Qiu et al. (2024) connect routers across layers by construction, feeding a recurrent router during training. Systems work exploits cross-layer routing regularity for prefetching and offloading (Eliseev & Mazur, 2023; Yao et al., 2024), optimizing latency, which we do not measure.

None of these studies holds the complete score vector of the previous router fixed and asks what older discrete selections still contribute on top of it. That conditional question is the one we measure, distinct from attribution of the current decision (Li et al., 2026), from consistency along the token axis (Liang et al., 2026), and from architectural coupling of routers (Qiu et al., 2024).

Probing as measurement. Our instrument is a linear probe on frozen activations, a standard tool whose interpretation requires care: the success of a probe shows that information is decodable, not that the model uses it (Li et al., 2023). Probing studies across model families report per-model probe quality rather than architecture-level laws (Gurnee & Tegmark, 2024), and cross-layer analyses treat depth-varying quantities descriptively (Lawson et al., 2025). That the control condition can determine the conclusion is itself an established finding (Zhang & Nanda, 2024), which is why our shuffled-history control forms part of the estimand rather than a robustness check, and we follow work reporting positive and saturating effects together (Jelassi et al., 2025) and separating oracle bounds from achievable performance (Liang et al., 2026).

3 Measurement framework

3.1 Routing states

Consider a decoder-only transformer whose MoE layers are indexed $1, \ldots, L$ over routed layers only; layers without a router, such as a leading dense feed-forward block, are excluded from the index, as are shared or always-on experts from every state. At routed layer ℓ with E experts and native top-k, the router produces a score vector $\mathbf{g}_\ell \in \mathbb{R}^E$ and the model selects the set $S_\ell \in \{0,1\}^E$, $\sum_i S_{\ell,i} = k$. We take $\mathbf{g}_\ell$ to be the pre-softmax router logits recomputed from the input of the gate itself, the quantity on which the native top-k operates.

Fix a target layer m. All quantities are measured for one token position and one anchor at a time, so every state belongs to the same token: the target is $Y = \mathbf{g}_m$, the conditioning state is $G = \mathbf{g}_{m-1} \in \mathbb{R}^E$, and older selection history is $H = [S_{m-4}; S_{m-3}; S_{m-2}] \in \{0,1\}^{3E}$. The complete score vector G is not a sufficient statistic for the computation of that layer, but it is the strongest recent-state conditioner available from routing alone.

3.2 Three predictors and two estimands

Figure 1 shows the resulting design. The probe does not consume H at its raw $3E$ width: the $3E$ binary coordinates are standardized, reduced by PCA to E components and standardized again, every transform fitted on the fitting half only, giving an E-dimensional history block. We fit one probe family, multi-output ridge regression with $\alpha = 1$ and no α search, on three designs: G alone, the aligned design $[G; H]$, and a misaligned control $[G; \pi(H)]$ in which π permutes history rows, rather than coordinates, across samples, so each token receives the history of another token. The aligned and misaligned designs are therefore both exactly $2E$ wide, while the recent-only design is E wide. Writing A, R and B for held-out uniform-average multi-output R^2 under the three designs, we report two increments: the *direct increment* $\Delta_G = R - A$, value beyond the scores of the previous layer, and the *alignment-specific value* $\Psi = R - B$, value beyond same-capacity misaligned history. $\Delta_G > 0$ says older history helps relative to G alone; $\Psi > 0$ says the help depends on the history belonging to the token being predicted rather than on the design being wider. Neither implies the other, so both numbers appear wherever either is claimed.

The control must match capacity, not just width. A zero-padded block equalizes nominal dimension but not fitted capacity: against $[G; \mathbf{0}]$, only the history arm pays the ridge estimation cost of E additional fitted columns, so part of any difference is attributable to width alone. An earlier

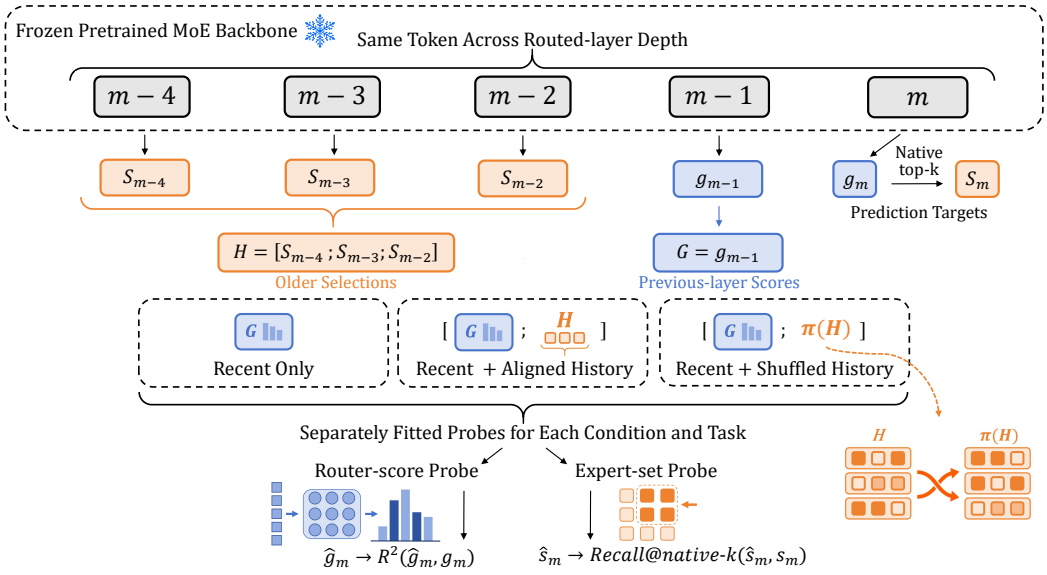

Figure 1: Primary conditional-probing design. One token is followed across the routed layers of a frozen pretrained MoE backbone, where the native top-k rule produces the stored selection S_m. Older selections form $H = [S_{m-4}; S_{m-3}; S_{m-2}]$ and the previous layer provides $G = \mathbf{g}_{m-1}$; the three conditions combine these states as recent only, recent plus aligned history, and recent plus shuffled history. The inset shows what π does: after preprocessing, it exchanges whole history rows between samples, so each token receives another token's history with the order of layers within a row preserved; coordinates within a row are never shuffled and G is left unchanged. A separate ridge probe is fitted independently for each condition and each task, scored by held-out R^2 on $\mathbf{g}_m$ and by Recall@native-k on S_m, and the two tasks are evaluated on non-overlapping anchor manifolds. Both targets are fitting and evaluation quantities, never evaluation-time inputs. Preprocessing, including the reduction of H from $3E$ to E dimensions, is described in the text. The figure covers this main design, and omits the headroom-normalized diagnostic, the oracle analyses, the runtime selector and richer conditioning.

study on a wider router used that pairing and measured the asymmetry on synthetic data, where the null case with no planted history yielded $\Delta = -0.0209$ instead of zero at $E = 60$ with 512 fitting rows, a bias comparable in size to the effect under study. The row permutation instead preserves width, active coordinate count and per-coordinate marginals exactly, so both arms carry the same estimation cost and differ only in sample alignment. We use 20 fixed permutations per cell, with independent permutations for the fitting and evaluation halves (Appendix B).

3.3 Headroom-normalized history value

Ψ is an absolute R^2 difference, so cells whose baseline B is already near 1 retain little room in which any increment can appear. To compare cells with very different baselines we use

$$\mathrm{HNV}_i = \frac{R_i - B_i}{1 - B_i} = \frac{\Psi_i}{1 - B_i}, \tag{1}$$

the fraction of the remaining R^2 headroom of the misaligned baseline that correctly aligned history closes. Here B_i is the mean over the 20 same-capacity shuffled controls of cell i, and R_i and B_i are evaluated on the same held-out split, so the ratio compares like with like. All 16 denominators are comfortably positive (minimum $1 - B = 0.163$), so neither clamping nor special-casing arises.

HNV is a normalized empirical ratio, not an information-theoretic quantity and not a formal partial R^2, since the aligned and misaligned designs are same-capacity alternatives rather than a nested pair. Normalization does not by itself remove baseline geometry: on a synthetic case holding fractional closure constant while sweeping the baseline, $\mathrm{Spearman}(A, \Psi) = -1.000$ whereas $\mathrm{Spearman}(A, \mathrm{HNV}) = -0.381$ (Appendix D).

3.4 EXPERT-SET OUTCOME

R^2 on scores is a convenient continuous target, but not the object a routing system acts on. We therefore also predict the *selected set*: from the scores predicted by a probe we take the native top-k and measure Recall@native-k against S_m, the set selected by the model itself, read from the gate rather than reconstructed by a top-k that exact ties could perturb. Both sets contain exactly k experts, so recall equals precision here, and utility is $U_i = \text{Recall}_i(\text{recent}+\text{history}) - \text{Recall}_i(\text{recent})$, with an alignment-specific counterpart U_i^{align} formed against the shuffled-history arm. Throughout, "information" means decodable by this probe family on these targets, and the oracle quantities in §5 pick among predictors using ground-truth utility, so they bound complementarity from above and are reported separately from what a runtime rule achieves.

4 EXPERIMENTAL SETUP

Models and cells. We study four pretrained MoE language models with pinned revisions: OLMoE-1B-7B-0125 ($L = 16$, $E = 64$, $k = 8$) (Muennighoff et al., 2024), JetMoE-8B ($L = 24$, $E = 8$, $k = 2$) (Shen et al., 2024), Qwen1.5-MoE-A2.7B ($L = 24$, $E = 60$, $k = 4$) (Bai et al., 2023) and DeepSeek-MoE-16B-base ($L = 28$ routed, $E = 64$, $k = 6$) (Dai et al., 2024). All backbone weights remain frozen in evaluation mode: no backbone parameter is trained, quantized or edited, and no activation is altered, while the probes of §3 are fitted on the fitting half only. Four target layers per model are chosen mechanically as $\arg\min_m |m/L - f|$ for $f \in \{0.3, 0.5, 0.7, 0.9\}$, ties to the smaller m and never by performance, giving a fixed grid of 16 model–depth cells. The routed component is the routed-MLP gate of each model, identified by module identity: the attention-side router of JetMoE is excluded, as are the shared experts of DeepSeek and its leading dense feed-forward block, and scores are recomputed from the input of each gate in the model dtype rather than read from a convenience router-logits API. The recomputation reproduces the native top-k selection of each model exactly (Appendix B).

Data and splits. Anchors are non-overlapping 4096-character spans of WikiText-103-raw (Merity et al., 2016) at a pinned dataset revision, taken in increasing offset order; a span is eligible only if all four tokenizers yield at least 128 tokens. Each anchor contributes one sample, at token position 127 (zero-based), and all four models receive byte-identical raw text. A study uses 3072 anchors: 2048 for fitting and the next 1024, disjoint, held out. Every scaler and the history PCA are fitted on the fitting half only, and no held-out quantity enters any fitted transform or any model-selection decision. There is no hyperparameter search on the evaluation split, because there is no hyperparameter search at all.

The expert-set studies use a disjoint anchor manifest whose candidate spans begin strictly after the maximum character offset of every anchor used for the score-level grid. This matters for one claim in particular: the HNV diagnostic is computed on the first manifest and the expert-set utility it is asked to predict on the second, so the correlation relates a diagnostic and an outcome measured on non-overlapping text. External validation uses C4 (Raffel et al., 2020) at a pinned revision, with 2048 anchors selected by the same deterministic rule and four token positions $\{31, 63, 95, 127\}$ captured in the same forward pass.

Uncertainty. All intervals are 95% percentile intervals from 10,000 paired bootstrap replicates over the 1024 held-out raw anchors, with one shared anchor draw per replicate across every architecture, cell and arm, so paired comparisons remain paired and within-model dependence between the four depths of one architecture is preserved rather than assumed away. Probes are never refitted inside the bootstrap: fixed predictions are rescored under draw multiplicities. The intervals therefore quantify anchor sampling, not variation across pretraining runs, whose seeds are fixed by the released checkpoints, and the 16 cells are 4 architectures $\times$ 4 depths rather than 16 independently sampled model families. Where a pooled result could be driven by a single architecture we report the per-architecture split alongside the pooled value, with leave-one-architecture-out checks and the four distinct randomness roles in Appendix D.

Protocol discipline. Each study carries a protocol, thresholds and a verdict rule committed to its repository before the corresponding results existed, with a machine-checked record of that ordering;

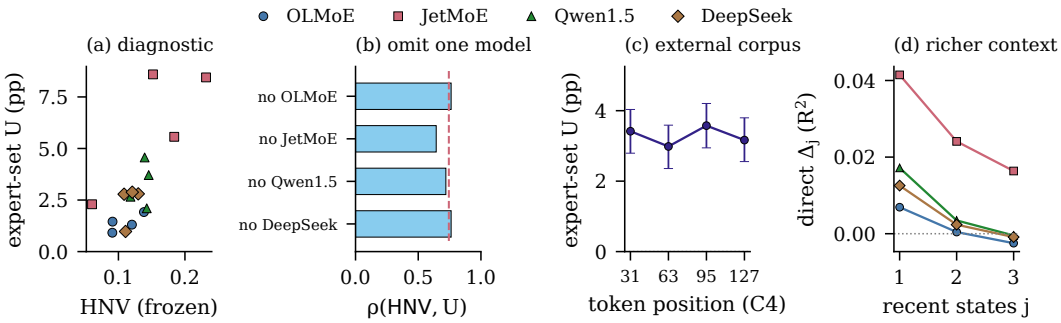

Figure 2: Four results the tables cannot convey. **(a)** The headroom-normalized diagnostic HNV, computed on one anchor manifold, against expert-set utility U realized on a disjoint manifold ($\rho = 0.744$, 95% CI $[0.565, 0.859]$). **(b)** The same correlation with each architecture omitted in turn over the remaining 12 cells; the dashed line marks the value over all 16 cells. **(c)** Expert-set utility on C4 at four token positions from one forward pass, with 95% bootstrap intervals. **(d)** The direct increment as the conditioning state widens to j recent score vectors: JetMoE remains positive while the three wider routers reach or cross zero, which is why the pooled residual at j=3 holds in only 5 of 16 cells. Panel (d) fixes older history at the single state S_{m-4} and continues no series from the three-layer H used elsewhere.

Table 1: Primary grid, both tasks, per architecture and pooled. Δ_G and U are the direct increments over the recent-only arm, whereas Ψ and U^{al} are the increments over the same-capacity misaligned arm. The two halves are measured on non-overlapping anchor manifolds, so a row does not report two metrics on one sample of text. Pooled rows are the equal-weight mean over the 16 cells, the frozen aggregation, and pooled CIs are the frozen bootstrap intervals, not differences of arm-wise intervals. Per-cell values and the remaining CIs: Tables 5 and 6.

	(a) router scores, R^2					**(b) expert sets**, Recall@k				
model	G	$[G;H]$	$[G;\pi(H)]$	Δ_G	Ψ	recent	+hist.	+π(h.)	U	U^{al}
OLMoE	0.772	0.791	0.765	+0.018	+0.025	0.758	0.772	0.752	+0.014	+0.020
JetMoE	0.264	0.371	0.261	+0.108	+0.110	0.517	0.579	0.512	+0.062	+0.067
Qwen1.5	0.703	0.736	0.695	+0.033	+0.041	0.618	0.650	0.607	+0.033	+0.043
DeepSeek	0.745	0.768	0.737	+0.023	+0.031	0.648	0.671	0.639	+0.024	+0.032
Pooled	0.621	0.667	0.615	**+0.045**	+0.052	0.635	0.668	0.627	**+0.033**	+0.041
95% CI, pooled Δ_G: $[+0.0435, +0.0474]$; pooled U: $[+3.01, +3.61]$ pp.										
$\Delta_G > 0$ and $U > 0$ in 4/4 cells per architecture, 16/16 pooled.										

"preregistered" means exactly that arrangement throughout, never registration in an external registry (Appendix D). One exception: the headroom concern motivating Equation 1 was formulated after the absolute-Ψ result had been observed, making the headroom analysis a post-hoc robustness study with its own frozen rule.

5 Results

5.1 Residual predictive structure beyond the previous layer

Across the 16 cells, adding older selection history to the complete score vector of the previous layer raises held-out R^2 by $\Delta_G = +0.0455$ on average (95% CI $[0.0435, 0.0474]$), positive in all 16 cells (Table 1a). The alignment-specific value is slightly larger at $\Psi = +0.0519$, also 16/16, the misaligned control performing slightly worse than G alone as expected of a same-width block of correctly distributed but wrongly assigned features.

The effect varies by nearly an order of magnitude across architectures (Table 1a), and JetMoE, the largest at $+0.108$, is also the model whose previous-layer state is least predictive on its own ($A = 0.264$ against 0.70–0.77). Across cells the two move together, $\text{Spearman}(A, \Psi) = -0.900$

(95% CI $[-0.947, -0.844]$): where the previous layer already explains the next one well, older history adds less. Part of this trade-off is geometric, since a cell with B near 1 has little headroom for an absolute increment, but normalizing by that headroom (Equation 1) weakens the relationship without removing it, giving $\mathrm{Spearman}(A, \mathrm{HNV}) = -0.229$ (95% CI $[-0.294, -0.109]$) with $\mathrm{HNV} > 0$ in 16/16 cells. Since the per-architecture rank correlations are not all negative, we read this as a cross-cell pattern rather than a per-model law.

One alternative explanation is that history only appears useful because a linear probe cannot exploit the recent state fully. A local control weakens it: in JetMoE at layer 12, a capacity-matched nonlinear recent-state predictor improves on the linear recent probe by $+0.0014$, whereas adding history to it is worth $+0.1644$ at matched parameter count. This is one model at one layer with its own training setup (Appendix H).

5.2 Expert sets, and a diagnostic that predicts where history helps

R^2 on scores is a proxy for what a routing system would act on, so we repeat the comparison with the selected set as the target. Adding older history raises Recall@native-k from 0.6350 to 0.6681, an increment of $U = +3.31$ pp (95% CI $[3.01, 3.61]$ pp) positive in 16/16 cells, and $+4.08$ pp against the shuffled arm, also 16/16 (Table 1b). This is a deliberately simple linear probe's increment, not a system-level gain; §5.5 evaluates it against a strong predictor.

The more useful result is that the increment is predictable in advance. HNV_i, frozen from the score-level study, ranks the 16 cells by their realized expert-set utility with $\mathrm{Spearman}(\mathrm{HNV}, U) = +0.744$ (95% CI $[0.565, 0.859]$), that utility measured on anchors beginning strictly after every anchor used to compute HNV; omitting any single architecture leaves the correlation between $+0.64$ and $+0.76$ (Figure 2a,b).

Using the diagnostic to place a history budget. Since the ranking precedes the expert-set measurement, it can decide *which* cells receive history. Giving history to the 8 of 16 cells ranked highest by HNV yields a mean Recall@native-k of 0.65859 over all 16 cells, against 0.63503 with history nowhere and 0.66814 everywhere: this half-budget retains $(0.65859 - 0.63503)/(0.66814 - 0.63503) \approx 71.2\%$ of the average gain of full deployment, exceeding the mean over random half-budget assignments (0.65157) by 0.702 pp. The ranking comes from the score-level study on the first anchor manifest and the recall values from the disjoint expert-set manifest, so selection and evaluation share no text. What the comparison establishes is that a cheap diagnostic places a budget usefully, not that HNV is the best rule for doing so: raw Ψ (0.65816) and lowest recent-only baseline A (0.65776) do about as well, the hindsight-best choice reaches only 0.65962, and the analysis was designed after the per-cell gains were known, so it demonstrates the link on the measured grid rather than validating it prospectively (all budgets and rules: Table 8). Half the *cells* is also not half the compute, and HNV tracks gain size rather than the token on which it falls, correlating -0.10 with per-token strict-win frequency, so effect size and budget ranking are its demonstrated uses.

5.3 Both effects reproduce on a new corpus and at four token positions

Everything above was developed on WikiText at a single token position. Repeating the frozen protocols on C4 (Raffel et al., 2020), with a new anchor manifest and four token positions captured in the same forward pass, reproduces both effects: the score-level direct increment is $+0.0486$ (95% CI $[0.0463, 0.0510]$) and the expert-set utility is $+3.29$ pp (95% CI $[2.97, 3.61]$ pp), each positive in all 16 architecture–position cells, with positive means at all four positions (Figure 2c). JetMoE remains much the largest, at $+0.133$ against $+0.016$–$+0.027$ for the other three. The full ranking does not carry over: WikiText orders the rest Qwen1.5 > DeepSeek > OLMoE, C4 orders them OLMoE > Qwen1.5 > DeepSeek, all within 0.011 R^2 on C4. What transfers is the presence of the effect in every cell and the separation of JetMoE, not an ordering among the wider routers. With the four models held fixed, this is transfer across corpora and positions.

5.4 History value under richer recent conditioning

How much history is worth depends on how much recent state the predictor already has, so we vary that directly. Holding the older-history input fixed at a single earlier selection state S_{m-4}, we

Table 2: Four quantities that are easily conflated, each against the comparator relative to which it is measured. Because the comparators differ, these gains are not additive components of one total and must not be differenced against one another. The first three are hindsight oracles bounding complementarity, not achievable performance; only the last operates without observing the target. F is the gate-reading predictor, RR the HCR-B reranker, RH recent-plus-history, REC a recent-only third arm, $\pi(\mathrm{RH})$ shuffled history. Per-cell: Table 12.

Comparison	Exact comparator	Gain (pp)	95% CI (pp)	Type
Raw oracle, history added	$O(F, \mathrm{RR})$, no history	+2.7102	$[+2.5640, +2.8600]$	oracle
Matched oracle, vs. recent-only	$O(F, \mathrm{RR}, \mathrm{REC})$	+0.5587	$[+0.4557, +0.6643]$	oracle
Alignment-controlled oracle	$O(F, \mathrm{RR}, \pi(\mathrm{RH}))$, 20 ctrls	+0.6324	$[+0.5355, +0.7304]$	oracle
Runtime selector, 3 predictors	HCR-B, operating baseline	+0.1831	$[+0.1099, +0.2569]$	runtime

widen the conditioning state to G_j, the j most recent score vectors. The pooled direct increment falls from $+0.0195$ to $+0.0076$ and $+0.0031$, with intervals $[0.0181, 0.0209]$, $[0.0066, 0.0086]$ and $[0.0024, 0.0039]$, and the alignment-specific value likewise to $+0.0087$ at G_3. This series uses a one-state history input where §5.1 uses the three-state H, so the two are separate protocols and $+0.0031$ is not the endpoint of a decay curve beginning at $+0.0455$.

The architecture split is what matters here. At G_3, 5 of 16 cells have positive Δ_3, with per-architecture means $+0.0164$ for JetMoE (4/4 positive) against -0.0025, -0.0005 and -0.0009 for OLMoE, Qwen1.5 and DeepSeek (Figure 2d); the pooled interval excludes zero because the JetMoE effect carries an equally weighted mean of 16 cells. Three recent score states thus absorb the older-history signal in the three wider routers but not in the narrow one, which measures how the value of history depends on conditioning. Ψ_3 meanwhile stays positive for all four architectures, aligned history still beating misaligned history at $j = 3$ even where it no longer beats the recent-only predictor (Appendix G).

5.5 What history adds on top of a strong predictor

The comparisons so far use history against a recent-only probe of the same family, whereas a deployment would face a strong predictor. We therefore evaluate against two strong baselines on the same held-out universe: a fixed-budget adaptation of a gate-reading rule reaching 0.8595 Recall@native-k, and a candidate reranker reaching 0.8629. Both sit far above the 0.6681 of the history probe because they read the gate parameters of the target layer itself, an input the history predictors never receive, so these are predictors with different information access rather than a weaker and a stronger fit to the same inputs.

What history retains under that competition is complementarity on particular tokens. A pointwise oracle over the strong predictors plus the history predictor reaches 0.8981 against 0.8629 for the reranker alone, and $+2.71$ pp of that headroom is reachable only through the history arm rather than the two strong predictors. That raw figure still credits history with value that *any* additional predictor would supply, so attribution needs a matched third arm: substituting a recent-only predictor of the same family raises the two-predictor oracle by $+2.15$ pp on the same universe. Net of that generic diversity, history contributes $+0.56$ pp matched and $+0.63$ pp under the alignment control, both with intervals excluding zero, so the history-specific complementarity is real and small. Measured against different comparators, these quantities are not additive parts of one total (Table 2).

A runtime selector converts little of that bound. Trained with grouped cross-validation on 18 runtime-observable features and never shown the target set, it reaches $+0.1831$ pp over the reranker. As a fraction of the $+3.5164$ pp by which the three-predictor oracle exceeds HCR-B, that realized gain is $(\mathrm{Recall}_{\text{FRH}} - \mathrm{Recall}_{\text{HCR-B}})/(O_{\text{FRH}} - \mathrm{Recall}_{\text{HCR-B}}) = 0.1831\,\text{pp}/3.5164\,\text{pp} = 5.2\%$ (95% CI $[3.2, 7.2]\%$); the denominator is that total oracle margin, not the matched $+0.56$ pp above. Essentially all of the gain arises from admitting the history arm, a selector over the two strong predictors alone gaining nothing (-0.0013 pp), but of the five feasibility conditions fixed in advance it clears two and misses three (both effect-size floors and the 20% recovery floor), so the frozen outcome is that competence is not predictable well enough to convert this oracle gap. With no symmetric recent-only or shuffled-history selector arms, the $+0.1831$ pp is attributable to adding the history arm but not to its alignment.

A related boundary comes from the same direction: when history is conditioned on the behaviour of the gate-reading predictor, the three preregistered tests of a conditioned history gain all came out negative, including a test of whether HNV transfers to that setting (Appendix F). HNV predicts utility for the expert-set probe family on which it was validated; it does not clearly generalize to a gate-reading strong predictor.

6 DISCUSSION

What the measurements establish. Conditioning on the complete score vector of the previous router, older expert selections remain predictive of the next routing decision under the probe families we fit, the increment depends on those selections belonging to that token, and it transfers from continuous scores to the discrete set the model selects. It is also anticipatable and reproducible, from a headroom-normalized quantity computed on separate text and on an unseen corpus at four token positions. Routing in these four models is thus not well described as a sequence of independent per-layer decisions, on the evidence of decodability under the probes used here (§3) across 16 cells drawn from four checkpoints.

What these measurements help decide. Two uses follow from the numbers rather than from extrapolation. The per-cell results say where history features earn their place and where recent state already suffices, and HNV anticipates that ranking well enough to place a half-budget of history retaining about 71% of the full average recall gain; since simpler rankings do about as well, what this establishes is the link between a cheap diagnostic and realized gain rather than the superiority of one rule. Richer conditioning then makes the dependence quantitative: how much history adds is a function of how much recent state is present.

How far the explanation currently reaches. Two gaps are open. The distance between the $+0.56$ pp of history-specific oracle complementarity and the 5.2% of the oracle margin our runtime selector realized is not explained by anything we ran. And the architecture differences are observations, not mechanisms: JetMoE exceeds the other three by roughly an order of magnitude on every history quantity and alone retains a positive direct residual under three recent score states, while also having by far the narrowest router ($E = 8$, $k = 2$, against 60–64 elsewhere) and the least self-predictive recent state. That a narrow router encodes less per layer, leaving more of the next decision distributed over earlier ones, is a plausible reading and remains untested, since four checkpoints cannot separate router width from sparsity, depth and corpus.

Implications for analysing MoE routing. Two methodological points generalize beyond this dataset: the control condition determines the answer, and distinguishing raw oracle, matched oracle and runtime performance changes the conclusion by an order of magnitude on the same data, so analyses reporting only the first overstate what routing structure can deliver. The width reading is also directly testable, since nothing in our protocol depends on E beyond the history block width: sweeping E with depth, sparsity fraction and data fixed should make the direct increment decline with E, a comparison for which our frozen per-cell tables give the baseline.

7 CONCLUSION

Conditioned on the full score vector of the previous router, older aligned expert selections retain measurable incremental predictive value, on every cell of a frozen four-architecture grid and on an unseen corpus, for both router scores and the expert set the model selects. That value is neither uniform nor unconditional: it varies by nearly an order of magnitude across architectures, a diagnostic computed on separate text anticipates where it is largest well enough to place a half-budget of history retaining about 71% of the full average gain, and richer recent state absorbs most of it in the wider routers. Two problems follow, one applied and one mechanistic: closing the distance between measured complementarity and what a runtime rule realizes, and testing whether router width governs the size of the effect.

REPRODUCIBILITY STATEMENT

All four backbones are public checkpoints pinned by revision hash, and both corpora are public datasets pinned by revision; the exact identifiers, revisions, target layers, anchor-selection rule and preprocessing order are given in Appendix B, and the statistical conventions in Appendix D. Every number in the main text is reproduced per cell in Appendix C, and Appendix J maps each study to the section or appendix reporting it.

The anonymous artifact at `https://anonymous.4open.science/r/moe-routing-anonymous-artifact-FD5A` contains the analysis implementations for every reported study (router-state capture, probe fitting, the expert-set predictors, the candidate reranker, the competence selector, the oracle and complementarity analyses, the shuffled-history controls, the paired bootstrap and the two-stage table aggregation), together with the per-study unit and pre-result check suites, the frozen per-cell result tables, and the derived arrays required to regenerate each figure. A CPU-only path (`./run.sh smoke`) runs the synthetic validator of each study on planted-signal and null fixtures, so the analysis code can be exercised without a GPU or network access; `./run.sh tests` runs the check suites. Raw router captures and the full raw-span manifests are not bundled because of their size, and `./run.sh reproduce` regenerates them from the public checkpoints in named stages. Checks that assert properties of the original per-study repositories, such as frozen commit hashes, are skipped with an explicit reason rather than deleted, so the runner reports what it is and is not verifying. Figure scripts read the frozen tables and write, alongside each figure, a CSV of exactly the values plotted.

An anonymous interactive view of the same material is available at `https://anonymous-artifact.net/`. It provides a token-level routing explorer over the captured states, described in Appendix I, a navigable graph of the studies with per-cell provenance and frozen source records, and the definitions used here. It reads the same frozen numbers reported in this paper and serves as a navigational aid: every quantitative claim above is self-contained in this PDF and no conclusion requires visiting the site.

AI USE STATEMENT

Large language models were used substantially in this work, in the following roles.

Review of experiment code. LLM assistance was used to review the experiment code written by the authors, including the capture, probe, control, bootstrap and aggregation implementations and the per-study unit and pre-result check suites. All of it was written and executed by the authors, and the numerical results reported here are the output of those executions rather than of any description produced by a model.

Interpretation of results. LLM assistance was used to reason about what the measured quantities do and do not license, in particular the separation of probe decodability from conditional mutual information and Markov order, and the separation of raw oracle, matched oracle and runtime quantities.

Figures and writing. The figures were produced by scripts that read frozen result tables; those scripts were written by the authors and reviewed with LLM assistance, and no figure content was generated by an image model. The manuscript text was revised with LLM assistance.

The authors reviewed all AI-assisted output, verified the numbers against the frozen study artifacts, and take full responsibility for the content of this paper, including its claims, its citations and its statistical statements.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou,

Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

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 Johnson, Katie Millican, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, 2022. arXiv:2202.01169.

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*, 2024.

Artyom Eliseev and Denis Mazur. Fast inference of mixture-of-experts language models with offloading. *arXiv preprint arXiv:2312.17238*, 2023.

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. arXiv:2101.03961.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024.

Samy Jelassi, Clara Mohri, David Brandfonbrener, Alex Gu, Nikhil Vyas, Nikhil Anand, David Alvarez-Melis, Yuanzhi Li, Sham M. Kakade, and Eran Malach. Mixture of parrots: Experts improve memorization more than reasoning. In *International Conference on Learning Representations*, 2025.

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.

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer SAEs. In *International Conference on Learning Representations*, 2025.

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. arXiv:2006.16668.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023.

Wengang Li, Lingqi Zhang, Toshio Endo, and Mohamed Wahib. Understanding cross-layer contributions to mixture-of-experts routing in LLMs. In *International Conference on Learning Representations*, 2026.

Ziyue Li and Tianyi Zhou. Your mixture-of-experts LLM is secretly an embedding model for free. In *International Conference on Learning Representations*, 2025.

Jingcong Liang, Siyuan Wang, Miren Tian, Yitong Li, Duyu Tang, and Zhongyu Wei. Not all models suit expert offloading: On local routing consistency of mixture-of-expert models. In *International Conference on Learning Representations*, 2026.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. A closer look into mixture-of-experts in large language models. *arXiv preprint arXiv:2406.18219*, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024. Spotlight.

Zihan Qiu, Zeyu Huang, Shuang Cheng, Yizhi Zhou, Zili Wang, Ivan Titov, and Jie Fu. Layerwise recurrent router for mixture-of-experts. *arXiv preprint arXiv:2408.06793*, 2024.

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. arXiv:1910.10683.

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. arXiv:1701.06538.

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

Zunhai Su, Qingyuan Li, Hao Zhang, Weihao Ye, Qibo Xue, Yulei Qian, Ngai Wong, and Kehong Yuan. Unveiling super experts in mixture-of-experts large language models. In *International Conference on Learning Representations*, 2026.

Jinghan Yao, Quentin Anthony, Aamir Shafi, Hari Subramoni, and Dhabaleswar K. Panda. Exploiting inter-layer expert affinity for accelerating mixture-of-experts model inference. *arXiv preprint arXiv:2401.08383*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

A Complete study index

This paper reports one line of work comprising 30 recorded studies. The table below indexes all of them: the question each asked, its type, its frozen outcome, and where it appears. "Frozen" means that the study carries a preregistration committed before its results, a machine-checked record that its checks ran before any evaluation statistic existed, and a verdict rule recorded as fixed in advance. Negative and closed results are listed with the same status as positive ones and are not restated as successes.

Aliases. HCR-B is arm B of HCR-E1, not a separate study. C0 is C0-RPC and C1 is C1-RCE. H2 is the coverage-versus-utility branch within the pair-utility study, closed by a recorded futility stop. RAC-E1 yields two estimands (A: matched complementarity, B: direct increment) from one study on one universe, so their uncertainties are not independent replications.

Shared universes. RSS-P4 and HCR-P5 share one anchor manifest byte-for-byte. HAP-P6, LHR-P6.7, HCR-E1, SRS-E1, C0-RPC, C1-RCE, RAC-E1 and MRS-E1 share a second, disjoint manifest. ECV-E1 uses an independent C4 manifest. Studies on a shared manifest are not independent replications of one another, and we do not count them as such anywhere in this paper.

Table 3: Complete index of the 30 recorded studies: question asked, study type, frozen outcome, and where in this paper each is reported. "Frozen" means that the study carries a protocol and verdict rule committed before its results, with a machine-checked record of that ordering. Negative and closed outcomes are listed with the same status as positive ones.

ID	Type	Question	Frozen outcome	Where
REDV-V1	explor.	Do rejected experts carry delayed value?	Uninterpretable: no matched control	H
REDV-V2	closed	Same, with a matched rejection control	Closed	H
DREV-P0	closed	Delayed value of rejected evidence	Closed	H
EIPC-P0	explor.	Is expert identity preserved across layers?	Provenance preserved	H
XEC-P0	closed	Cross-layer expert-cache intervention	Closed	H
EPD-P0	descr.	Path versus content in expert provenance	Descriptive	H
RMO-P0	descr.	How long a history window is decodable?	Descriptive, post-hoc on EPD-P0	H
RMC-P0	frozen	Does history dependence replicate cross-model?	Replicated	H
NHD-P0	frozen	Does the gain survive a nonlinear recent probe?	Survives	H
FRS-P1	frozen	Does it survive the *full* recent router state?	Survives; $\Delta \geq +0.128$	B
RHP-P2	descr.	How does value vary with lag?	Descriptive lag profile	G
CAMR-P3	frozen	Does it replicate on a wider router?	**THIRD-MODEL-NOT-REPLICATED**	B
RSS-P4	frozen	Value beyond G under a same-capacity control	SUFFICIENCY-HISTORY-TRADEOFF	§5, C
HCR-P5	frozen	Is the trade-off a headroom artifact?	HEADROOM-ROBUST-TRADEOFF	§5, C
HAP-P6	frozen	Does HNV predict expert-set utility?	HNV-PREDICTS-HISTORY-UTILITY	§5, C
FHR-P6.5	frozen	Does history help a gate-reading predictor?	3 negative verdicts	§5.5, F
FHD-P6.6	descr.	Why did P6.5 fail?	Descriptive diagnosis	F
LHR-P6.7	frozen	Can history rerank locally?	LOCAL-HISTORY-RERANKER-NOT-CONFIRMED	F
H2	closed	Coverage versus utility	Futility stop	F
SUG-E1	explor.	Can swap utility be gated?	Does not alter P6.7	F
HPU-E1	closed	Pair-level history utility	Closed	F
HCR-E1	frozen	Candidate reranking (defines HCR-B)	Frozen; supplies the strong baseline	§5.5, F
ACB-E0	frozen	Are adaptive candidate budgets feasible?	ADAPTIVE-BUDGET-NOT-FEASIBLE	F
SRS-E1	frozen	Setwise versus candidate-wise scoring	SETWISE-RECENT-NOT-PROMISING	F
C0-RPC	frozen	Is there oracle complementarity?	COMPLEMENTARITY-SUFFICIENT	§5.5, F
C1-RCE	frozen	Can a runtime selector realize it?	**COMPETENCE-NOT-PREDICTABLE**	§5.5, F
RAC-E1-A	frozen	Matched complementarity attribution	HISTORY-ATTRIBUTABLE-...-SUPPORTED	§5.5, F
RAC-E1-B	frozen	Direct increment beyond G	Supported, $+0.0455$	§5, C
ECV-E1	frozen	Does it hold on C4 and other positions?	Both supported; POSITION-ROBUST	§5, E
MRS-E1	frozen	Does it survive three recent score states?	Both labels hold; 5/16 cells	§5, G

Scope of individual entries. REDV-V1 lacked a matched control and its apparent effect is not interpretable; it is recorded, not used. CAMR-P3 remains permanently THIRD-MODEL-NOT-REPLICATED under its own protocol; RSS-P4 does not reinterpret it, and its role here is to motivate why a same-capacity control is required. The three negative verdicts of FHR-P6.5 include the finding that HNV does not clearly generalize to gate-reading-conditioned utility, which is stated in §5 rather than only here. The small positive gain of C1-RCE sits under a failed feasibility gate.

B MODELS, DATA AND PROTOCOL

B.1 BACKBONES AND PINNED REVISIONS

Table 4: The four frozen backbones with their pinned revision hashes, routed depth L, expert count E and native top-k. Revisions are pinned, never resolved from a moving branch.

key	model id	L	E	k
olmoe	`allenai/OLMoE-1B-7B-0125`	16	64	8
revision	`9b0c1aa87e34a20052389dce1f0cf01da783f654`			
jetmoe	`jetmoe/jetmoe-8b`	24	8	2
revision	`d8fd02ccf7911aa8148a63c7984ffd2e465b0352`			
qwen	`Qwen/Qwen1.5-MoE-A2.7B`	24	60	4
revision	`1a758c50ecb6350748b9ce0a99d2352fd9fc11c9`			
deepseek	`deepseek-ai/deepseek-moe-16b-base`	28	64	6
revision	`521d2bc4fb69a3f3ae565310fcc3b65f97af2580`			

DeepSeek uses `trust_remote_code=True` with the remote-code revision pinned; the others do not. Inference is FP16, in `eval()` mode, under `inference_mode()`, with parameters frozen: no quantization, no adapters, no training, no weight updates and no activation editing. Captures ran on idle Tesla V100-SXM2-32GB GPUs (`torch` 2.4.1+cu121, `transformers` 4.45.1, `scikit-learn` 1.3.2). The per-architecture capture cost for the external-corpus study was 35.6–111.0 s at 14.7–37.1 GiB peak memory; the project-level total across all studies was not recorded.

B.2 ROUTER IDENTIFICATION AND SCORE EXTRACTION

Target layers are $\arg\min_m |m/L - f|$ for $f \in \{0.3, 0.5, 0.7, 0.9\}$, ties to the smaller m: OLMoE $(5, 8, 11, 14)$, JetMoE $(7, 12, 17, 22)$, Qwen1.5 $(7, 12, 17, 22)$, DeepSeek $(8, 14, 20, 25)$. Indices count routed layers; the leading dense feed-forward block of DeepSeek is excluded, so its routed range is $2..28$.

The routed component is identified by module identity rather than by name matching: the attention-side mixture-of-attention router of JetMoE is excluded (192 module ids), and the two shared experts of DeepSeek are excluded from every state, as is its dense block. Scores are recomputed from the input of each gate with an unbiased linear map in the dtype of the model, never read from a convenience router-logits API. The audit verifies, over 960–972 rows per model, that the recomputed scores reproduce the native top-k selection of the model exactly (maximum absolute deviation 0.0), that gate-returned indices match where the API exposes them, and that the gate module ids are distinct from attention and shared-expert module ids. OLMoE, Qwen1.5 and DeepSeek select via `topk(softmax(logits))` and JetMoE via top-k on raw logits; since the softmax is strictly monotone, pre-softmax logits are the native decision score in both cases.

B.3 ANCHORS AND SPLITS

The development corpus is `Salesforce/wikitext`, `wikitext-103-raw-v1`, `train`, at revision `b08601e04326c79dfdd32d625aee71d232d685c3`; raw `text` fields are joined by `"\n\n"` in row order with no normalization, giving a 541,897,031-character corpus verified by SHA256. Candidate spans are non-overlapping 4096-character windows at stride 4096 in increasing offset order; a candidate is eligible only if all four tokenizers yield at least 128 tokens with `add_special_tokens=False`. The first 3072 eligible spans are taken and split into 2048 fitting

and the next 1024 held-out anchors, disjoint. Each sample keeps its first 128 tokens and is evaluated at position 127 (zero-based). All four models tokenize byte-identical raw text.

The expert-set manifest is built by the same rule, but its candidates begin strictly after the maximum character end-offset of every anchor in the first manifest, so the two span sets are disjoint. This is what makes the HNV-to-utility correlation a cross-manifest result. The external study uses `allenai/c4`, config `en`, split `validation`, at revision `1588ec454efa1a09f29cd18ddd04fe05fc8653a2`: the whole split (364,608 documents) is scanned, 44,383 documents have at least 4096 codepoints, the first 2048 eligible candidates in ascending document order are taken with 0 rejected for short tokenization, and the split is 1024 fitting / 1024 held-out with no document or span overlap. Positions $\{31, 63, 95, 127\}$ come from the same forward pass, so any difference between them is a position effect rather than a capture difference.

B.4 Preprocessing and probes

Scores are centered within sample, then standardized per coordinate using fitting-half statistics only. History is standardized on its $3E$ raw binary coordinates (fitting-half only), reduced by `PCA` to E components fitted on the fitting half only (randomized solver, seed 20260927), then standardized again. Zero-variance coordinates are scaled by 1.0 and counted; none occurred. The probe is multi-output `Ridge` with $\alpha = 1.0$ throughout, with no α search, no alternative history lengths, no PCA-component search, no router probabilities and no hidden states. The metric is uniform-average multi-output R^2 on the held-out half, skipping coordinates with zero total variance. No held-out quantity enters any fitted transform, and because there is no hyperparameter search there is no opportunity to select a model on the evaluation split.

B.5 The same-capacity control

Exactly 20 fixed permutations per cell, with master seeds 41001–41020. For permutation p, the already-preprocessed history matrix has its *rows* permuted: fitting rows with seed p, held-out rows with seed $p + 100000$, so the two permutations are independently generated but deterministic. Coordinates within a row are never shuffled, the PCA is not refitted, and G is untouched. Each control design is exactly $2E$ wide, matching the aligned design in width, active coordinate count and per-coordinate marginals. An independent ridge probe is fitted per permutation and B_i is the mean over all 20; no permutation is ever selected or dropped.

Why not a zero block. The alternative capacity control pairs $[G; H_E]$ against $[G; \mathbf{0}_E]$, with G numerically identical in the first E coordinates and both designs exactly $2E$ wide. Nominal width then matches, but fitted capacity does not: the zero block contributes no fitted columns, so only the history arm pays the ridge estimation cost of E additional columns. CAMR-P3 quantified that asymmetry on synthetic data before any real statistic existed. Its null case (the same design with no planted history) gave $\Delta_{\text{full}} = -0.02088$ (95% CI $[-0.02398, -0.01823]$) at $E = 60$ with 512 fitting rows, with fitting R^2 rising $0.808 \to 0.834$ while held-out R^2 fell $0.757 \to 0.736$. That is the estimator behaving correctly rather than a defect, and its own protocol recorded it as rendering the real zero-block comparison conservative. For our purposes it means that a zero-block increment mixes history value with a width penalty of magnitude comparable to the effect we are measuring, which is why the control here permutes rows instead. Note that the value is a *synthetic null on synthetic data*, not a measurement of real noise features on the real grid.

B.6 Two prior controls that shaped this design

CAMR-P3 tested whether the history effect replicates on a wider router using a zero-block capacity control and returned **THIRD-MODEL-NOT-REPLICATED** (0 of 2 targets). That verdict stands unchanged under its own protocol and is not reinterpreted by any later study. Its measured zero-block estimation asymmetry is the direct reason why the present control permutes rows instead.

FRS-P1 asked whether the effect survives conditioning on the *full* recent router state rather than a reduced one, in JetMoE at two layers, with both arms held to exactly 16 dimensions ($[G; \mathbf{0}_8]$ versus $[G; H_8]$) so that capacity is equal by construction. Held-out R^2 rose from $+0.304$ to $+0.432$ at layer 12 ($\Delta = +0.1285$, 95% CI $[+0.0985, +0.1634]$) and from $+0.160$ to $+0.362$ at layer 20

($\Delta = +0.2021$, 95% CI $[+0.1682, +0.2389]$). This is a single-model result on a narrow router and is reported as such; the main-text grid is what establishes the cross-architecture picture.

C PER-CELL RESULTS

Every table in this appendix is generated directly from the frozen study artifacts by the script shipped in the anonymous repository, so no value is transcribed by hand.

C.1 SCORE-LEVEL CONDITIONING, ALL 16 CELLS

A is held-out R^2 under G alone, R under the aligned design $[G; H]$, B the mean over the 20 same-capacity misaligned controls, and $\Delta_G = R - A$ the direct increment. Pooled: mean $\Delta_G = +0.045468$, 95% CI $[+0.043528, +0.047449]$, 16/16 cells positive; mean $\Psi = R - B = +0.051914$, 16/16 positive. Per-architecture mean Δ_G: OLMoE $+0.018273$ (range $+0.010831$–$+0.024684$), JetMoE $+0.107635$ ($+0.051434$–$+0.153739$), Qwen1.5 $+0.032797$ ($+0.026462$–$+0.046749$), DeepSeek $+0.023166$ ($+0.017000$–$+0.026139$).

Table 5: Score-level conditioning, all 16 model–depth cells. A, R and B are held-out R^2 under G alone, the aligned design $[G; H]$, and the mean over the 20 same-capacity misaligned controls; $\Delta_G = R - A$.

model	m	m/L	E	k	A	R	B	Δ_G
OLMoE	5	0.312	64	8	0.6441	0.6669	0.6335	+0.0228
OLMoE	8	0.500	64	8	0.7777	0.8024	0.7707	+0.0247
OLMoE	11	0.688	64	8	0.8415	0.8563	0.8366	+0.0148
OLMoE	14	0.875	64	8	0.8265	0.8373	0.8211	+0.0108
JetMoE	7	0.292	8	2	0.0924	0.1438	0.0888	+0.0514
JetMoE	12	0.500	8	2	0.3288	0.4825	0.3265	+0.1537
JetMoE	17	0.708	8	2	0.3125	0.4366	0.3097	+0.1241
JetMoE	22	0.917	8	2	0.3214	0.4227	0.3192	+0.1013
Qwen1.5	7	0.292	60	4	0.5943	0.6411	0.5829	+0.0467
Qwen1.5	12	0.500	60	4	0.7822	0.8086	0.7760	+0.0265
Qwen1.5	17	0.708	60	4	0.7695	0.7966	0.7628	+0.0271
Qwen1.5	22	0.917	60	4	0.6670	0.6980	0.6575	+0.0309
DeepSeek	8	0.286	64	6	0.7992	0.8162	0.7934	+0.0170
DeepSeek	14	0.500	64	6	0.7473	0.7734	0.7396	+0.0261
DeepSeek	20	0.714	64	6	0.6849	0.7104	0.6752	+0.0256
DeepSeek	25	0.893	64	6	0.7490	0.7730	0.7418	+0.0239

C.2 HEADROOM NORMALIZATION AND EXPERT-SET UTILITY, ALL 16 CELLS

$\mathrm{HNV} = \Psi/(1 - B)$ from Equation 1. Recall columns are Recall@native-k for the expert-set probe on the disjoint manifest, and U is their difference. All 16 headroom denominators are positive with minimum $1 - B = 0.163396$; no value was clamped, and no cell or architecture was removed. Pooled $U = +0.033106$, 95% CI $[+0.030067, +0.036115]$, 16/16 positive. $\mathrm{Spearman}(\mathrm{HNV}, U) = +0.744118$, 95% CI $[+0.564706, +0.858824]$.

C.3 EXPERT-SET PREDICTORS, AVERAGED OVER THE FOUR DEPTHS OF EACH MODEL

Recall@native-k for every predictor evaluated on the shared held-out universe. Frequency, persistence and transition are zero-cost statistical baselines. The FATE-style and SpecPrefetch-style columns are fixed-budget *adaptations* of published rules rather than exact reproductions; the required adaptations are enumerated in the artifact. Both read the gate parameters of the target layer, which the history predictors never see, and the FATE-style rule exceeds recent-plus-history in 16 of 16 cells. Recent-plus-history is a deliberately simple probe whose role is to isolate the marginal value of older history under a fixed budget, not to compete as an expert predictor. Oracle is the native selection itself and exists only as the upper bound.

Table 6: Headroom normalization and expert-set utility, all 16 cells. Ψ and $1 - B$ are in held-out R^2; HNV $= \Psi/(1 - B)$ is a unitless fraction; the recall columns and U are Recall@native-k on the disjoint manifold. All 16 headroom denominators are positive, minimum $1 - B = 0.163396$, and nothing is clamped.

model	m	Ψ	$1 - B$	HNV	Recall(rec.)	Recall(+hist.)	U
OLMoE	5	+0.0334	0.3665	0.0911	0.7120	0.7266	+0.0145
OLMoE	8	+0.0317	0.2293	0.1383	0.7483	0.7675	+0.0192
OLMoE	11	+0.0197	0.1634	0.1203	0.8022	0.8153	+0.0131
OLMoE	14	+0.0163	0.1789	0.0908	0.7695	0.7787	+0.0092
JetMoE	7	+0.0550	0.9112	0.0603	0.4253	0.4482	+0.0229
JetMoE	12	+0.1561	0.6735	0.2317	0.5659	0.6504	+0.0845
JetMoE	17	+0.1270	0.6903	0.1839	0.5649	0.6206	+0.0557
JetMoE	22	+0.1035	0.6808	0.1520	0.5112	0.5972	+0.0859
Qwen1.5	7	+0.0581	0.4171	0.1394	0.5706	0.6162	+0.0457
Qwen1.5	12	+0.0326	0.2240	0.1455	0.6648	0.7019	+0.0371
Qwen1.5	17	+0.0338	0.2372	0.1426	0.6731	0.6941	+0.0210
Qwen1.5	22	+0.0405	0.3425	0.1181	0.5623	0.5889	+0.0266
DeepSeek	8	+0.0229	0.2066	0.1107	0.6886	0.6984	+0.0098
DeepSeek	14	+0.0338	0.2604	0.1298	0.6493	0.6772	+0.0280
DeepSeek	20	+0.0352	0.3248	0.1085	0.6437	0.6715	+0.0278
DeepSeek	25	+0.0312	0.2582	0.1208	0.6087	0.6375	+0.0288

Table 7: Expert-set predictors, Recall@native-k averaged over the four depths of each model. Frequency, persistence and transition are zero-cost statistical baselines. FATE* and SpecPre* are fixed-budget *adaptations* of published rules, not exact reproductions, and both read the gate parameters of the target layer, which the history predictors never see. Oracle is the native selection itself and exists only as the upper bound.

model	Freq.	Persist.	Trans.	FATE*	SpecPre*	Rec.	Hist.	Rec.+H	Rec.+π(H)	Oracle
OLMoE	0.2449	0.1201	0.6045	0.8714	0.1430	0.7580	0.7028	0.7720	0.7518	1.0000
JetMoE	0.2751	0.2626	0.4637	0.8523	0.2911	0.5168	0.4994	0.5791	0.5116	1.0000
Qwen1.5	0.0997	0.0640	0.5025	0.8505	0.0746	0.6177	0.5916	0.6503	0.6071	1.0000
DeepSeek	0.1569	0.0957	0.5119	0.8628	0.1092	0.6476	0.6124	0.6712	0.6388	1.0000

C.4 SELECTIVE RETENTION BY HNV

Mean Recall@native-k over all 16 cells when history is supplied to only the highest-ranked cells under a budget. History nowhere gives 0.63503, history everywhere 0.66814.

Table 8: Selective history retention. Mean Recall@native-k over all 16 cells when history is supplied only to the highest-ranked cells under a budget. History nowhere gives 0.63503, history everywhere 0.66814. HNV beats the random-selection mean at every budget but does not beat both simple heuristics at 4/16 or 12/16, and all four selectable rules fall within 0.00867 of each other.

budget	HNV	raw Ψ	lowest A	random mean	oracle utility
4/16	0.65148	0.65202	0.65060	0.64334	0.65202
8/16	0.65859	0.65816	0.65776	0.65157	0.65962
12/16	0.66349	0.66434	0.66382	0.65982	0.66523

HNV beats the random-selection mean at all three budgets but does not beat both simple heuristics at 4/16 or 12/16. At any one budget the four selectable rules fall within 0.00867 of one another and each sits close to the oracle-utility bound. This table therefore does not establish HNV as a superior allocation rule, and §5 states as much rather than reporting only the favourable comparison for the diagnostic.

C.5 SECONDARY ABLATIONS

Mean over the 16 cells: aligned versus misaligned history $+0.040794$ (16/16 positive, all 20 controls retained); history-only versus recent-plus-history $+0.066587$ in favour of including recent scores; binary recent state versus continuous router logits -0.029928, so discretizing the recent state loses more than older history adds; history lengths 1, 2 and 3 give 0.65276, 0.66424, 0.66822.

D STATISTICAL DEFINITIONS AND CHECKS

D.1 UNITS, AGGREGATION AND WHAT THE INTERVALS COVER

Score-level quantities are held-out uniform-average multi-output R^2 computed within a cell, then averaged with equal weight across the 16 cells. Expert-set quantities use a two-stage convention: the mean over the 1024 held-out anchors within each cell in `float32`, then an equal-weight mean across the 16 cells in `float64`. The two stages matter, since a single-pass grand mean over all anchor-cell pairs gives a slightly different headline, so the convention is fixed and applied uniformly. Cells are never weighted by expert count or by native k.

All intervals are 95% percentile intervals from 10,000 paired bootstrap replicates over the 1024 held-out raw anchors, with one shared anchor draw per replicate reused across every architecture, cell, arm and predictor. Probes are never refitted inside the loop; fixed held-out predictions are rescored under draw multiplicities, and agreement with the direct R^2 definition was verified as a runtime gate (maximum absolute deviation 9.99×10^{-16}). Seeds: 314159 for the score-level grid, the headroom analysis and the complementarity, attribution and selector studies; 271828 for the expert-set and external studies.

Scope of the intervals. The bootstrap resamples *anchors*. The intervals therefore quantify sampling of held-out text only. They do not quantify variation across pretraining runs, across checkpoints, or across MoE architectures. Four distinct randomness roles appear in this work and none is a repetition of model training: backbone pretraining seeds are fixed by the released checkpoints and never varied; probe fitting is deterministic given the fitting half; the 20 shuffle seeds generate the same-capacity null; the bootstrap resamples anchors.

Dependence between cells. The 16 cells are 4 architectures $\times$ 4 depths, so four cells within one architecture are dependent. Using one shared anchor draw across all cells preserves that dependence in the resampling rather than assuming independence. We report no per-architecture p-values: four depths cannot support four separate significance tests, and per-architecture rank correlations are

labeled descriptive. Counts such as "16/16 cells positive" are counts of positive per-cell *point estimates*, not 16 significance tests.

D.2 Correlation statistics and leave-one-architecture-out

Cross-cell relationships use Spearman rank correlation, not Pearson, and the directional hypothesis is fixed before the statistic is computed.

Table 9: Cross-cell rank statistics with their frozen bootstrap intervals and their role in the argument. Spearman throughout, never Pearson; the directional hypothesis was fixed before each statistic was computed. The two descriptive rows carry no interval because neither enters any verdict.

statistic	point	95% CI	role
$\mathrm{Spearman}(A, \Psi)$	-0.900000	$[-0.947059, -0.844118]$	primary, score level
$\mathrm{Spearman}(A, \mathrm{HNV})$	-0.229412	$[-0.294118, -0.108824]$	primary, headroom-controlled
$\mathrm{Spearman}(A, \mathrm{RER})$	-0.202941	$[-0.302941, -0.120588]$	secondary
$\mathrm{Spearman}(\mathrm{HNV}, U)$	$+0.744118$	$[+0.564706, +0.858824]$	primary, cross-manifest
$\mathrm{Spearman}(\mathrm{HNV}, D_i)$	$+0.547059$	—	descriptive: gain magnitude
$\mathrm{Spearman}(\mathrm{HNV}, W_i)$	-0.095659	—	descriptive: strict-win frequency

The last two rows are the basis for the statement in §5 that HNV tracks how large a gain is but not where it occurs, and they are also why HNV was deliberately excluded from the feature set of the runtime selector: a selector needs to know *which* predictor wins on the current token, which is the quantity HNV does not track.

Leaving out one architecture at a time, $\mathrm{Spearman}(\mathrm{HNV}, U)$ over the remaining 12 cells is $+0.762$ (omit OLMoE), $+0.643$ (omit JetMoE), $+0.720$ (omit Qwen1.5) and $+0.762$ (omit DeepSeek). These are reported without p-values and without a verdict, as preregistered. A sensitivity analysis that resamples the held-out anchors of both studies independently gives $+0.698727$, 95% CI $[+0.523529, +0.852941]$.

D.3 Overlap between the diagnostic and its target

HNV is computed on the score-level manifest; the utility it predicts is measured on a manifest whose spans begin strictly after every span of the first. The correlation is therefore between quantities measured on non-overlapping text. The two quantities do share a construction: both involve the misaligned baseline B, HNV through its denominator and the utility through its predictor family. That shared dependence is why the headroom normalization was validated separately, below. HNV was copied verbatim from the frozen headroom study into the utility study, never recomputed or redefined, and the reconstruction of the per-anchor predictions of the earlier study reproduced the frozen HNV values to 0.0.

D.4 Synthetic validation of the normalization

Before any real statistic existed, the headroom study ran four synthetic cases with a-priori tolerances. The decisive one holds the fractional closure constant at $h = 0.25$ while sweeping the baseline recent share from 0.10 to 0.65 across 8 cells: pure ceiling geometry, with no real trade-off. There $\mathrm{Spearman}(A, \Psi) = -1.000$ with a Ψ spread of 0.1452, whereas $\mathrm{Spearman}(A, \mathrm{HNV}) = -0.381$ with an HNV spread of 0.0198. A second case with a genuine fractional trade-off gives -1.000 under both. A null case with zero history share gives $\max|\mathrm{HNV}| = 0.00218$ against a tolerance of 0.02. Eight structural invariants were checked, including the HNV algebra, its reduction to Ψ at zero baseline, row-only permutation, and independence of the fitting and held-out permutations. No scientific parameter was tuned from a synthetic outcome, and no verdict threshold was reused as a code tolerance.

This shows that the normalization does not manufacture a strong negative trend out of ceiling geometry alone. It does not show that all baseline and headroom influence has been removed, and we make no such claim.

D.5 PREREGISTRATION STATUS, STATED PRECISELY

Each study carries a preregistration committed before its results, a `pre_result_checks` record confirming that the checks ran before any evaluation statistic existed, and a verdict rule flagged as frozen in advance. Two qualifications belong here rather than in the main text. First, the motivating concern of the headroom study was formulated *after* the absolute-Ψ result had been seen, so although its own rule was frozen before its own statistics, it is a post-hoc robustness analysis of an existing finding rather than an independent confirmation of it. Second, where this paper writes "preregistered" it means exactly the arrangement just described: a protocol, threshold or verdict rule committed to the repository of the study before the corresponding statistic existed, with a machine-checked record of that ordering. It does not mean registration in an external registry in the clinical-trial sense, and no claim of external registration is made anywhere.

E EXTERNAL-CORPUS AND TOKEN-POSITION VALIDATION

This study validates already-frozen methods on a corpus that was never used to develop them. Both probe families were imported from the frozen source repositories and the entire design, including the target-layer rule and the four token positions, was fixed before the first forward pass. Nothing was developed here, and its scope is explicitly a corpus and position check on the *same* four models.

Design. One target layer per architecture, chosen mechanically as the second-lowest layer of the frozen four-layer grid of that architecture (OLMoE m=8, JetMoE m=12, Qwen1.5 m=12, DeepSeek m=14), never by performance, and verified against the frozen grid before any result (50/50 checks pass). Four 0-based causal positions $\{31, 63, 95, 127\}$, all captured in the same forward pass, so any difference between positions is a position effect and not a capture or corpus difference. Cells are 4 architectures $\times$ 4 positions $= 16$.

Score-level structure. Overall mean $\Delta_G = +0.048554$, 95% CI $[+0.046256, +0.050974]$; mean $\Psi = +0.061101$, 95% CI $[+0.058819, +0.063586]$; levels $R^2(G) = 0.633943$, $R^2([G;H]) = 0.682497$, $R^2([G;\pi(H)]) = 0.621396$. Positive in 16/16 cells for both quantities. Architecture means for Δ_G: OLMoE $+0.026603$, JetMoE $+0.133443$, Qwen1.5 $+0.017749$, DeepSeek $+0.016420$. JetMoE is again an order of magnitude above the rest, which is the part of the pattern that transfers. The ordering among the other three does *not* transfer: on WikiText they run Qwen1.5 $(+0.0328)$ > DeepSeek $(+0.0232)$ > OLMoE $(+0.0183)$, whereas here they run OLMoE > Qwen1.5 > DeepSeek, all within 0.011 R^2 of one another. We report the separation of JetMoE as reproduced and make no claim that the full ranking is stable across corpora.

Expert-set utility. Overall Recall@native-k rises from 0.627169 (recent only) to 0.660027 (recent plus history), against 0.608978 for the misaligned arm. Mean $U = +0.032857$, 95% CI $[+0.029739, +0.036054]$; mean $U^{\text{align}} = +0.051048$, 95% CI $[+0.047984, +0.054131]$; both positive in 16/16 cells.

Position robustness. Per-position means, each averaged over the four architectures:

Table 10: External-corpus results by token position on C4, each mean taken over the four architectures. All four positions have positive mean Δ_G (held-out R^2) and positive mean U (Recall@native-k), which is the frozen condition for the position-robustness label. There is no monotone trend in position and none is claimed.

position	Δ_G	95% CI	U	95% CI
31	$+0.048730$	$[+0.044199, +0.053517]$	$+0.034170$	$[+0.027934, +0.040344]$
63	$+0.045375$	$[+0.040727, +0.050145]$	$+0.029846$	$[+0.023580, +0.035889]$
95	$+0.051092$	$[+0.046643, +0.055528]$	$+0.035756$	$[+0.029449, +0.042043]$
127	$+0.049018$	$[+0.044974, +0.053196]$	$+0.031657$	$[+0.025533, +0.037974]$

Corpus provenance. C4 (Raffel et al., 2020), config `en`, split `validation`, Hugging Face revision `1588ec454efa1a09f29cd18ddd04fe05fc8653a2` resolved before the manifest

Table 11: All 16 architecture–position cells on C4. $R^2(G)$ and Δ_G are held-out R^2 for the router-score task; the recall columns and U are Recall@native-k for the expert-set task. $\Delta_G > 0$ and $U > 0$ in 16/16 cells.

model	pos.	$R^2(G)$	Δ_G	Recall(rec.)	Recall(+hist.)	U
OLMoE	31	0.7299	+0.0260	0.7158	0.7308	+0.0150
OLMoE	63	0.7275	+0.0243	0.7130	0.7289	+0.0159
OLMoE	95	0.7118	+0.0273	0.7051	0.7191	+0.0140
OLMoE	127	0.7126	+0.0288	0.7029	0.7170	+0.0142
JetMoE	31	0.2893	+0.1348	0.5439	0.6226	+0.0786
JetMoE	63	0.3230	+0.1255	0.5586	0.6216	+0.0630
JetMoE	95	0.2963	+0.1414	0.5298	0.6177	+0.0879
JetMoE	127	0.3157	+0.1321	0.5586	0.6226	+0.0640
Qwen1.5	31	0.7675	+0.0175	0.6321	0.6606	+0.0286
Qwen1.5	63	0.7723	+0.0165	0.6311	0.6560	+0.0249
Qwen1.5	95	0.7712	+0.0182	0.6401	0.6619	+0.0217
Qwen1.5	127	0.7740	+0.0188	0.6235	0.6558	+0.0322
DeepSeek	31	0.7399	+0.0166	0.6265	0.6410	+0.0145
DeepSeek	63	0.7347	+0.0152	0.6219	0.6375	+0.0156
DeepSeek	95	0.7367	+0.0175	0.6209	0.6403	+0.0194
DeepSeek	127	0.7406	+0.0164	0.6108	0.6271	+0.0163

was created (a moving `main` was not used afterwards). The whole pinned split was scanned: 364,608 documents, 44,383 with at least 4096 codepoints, 2048 candidates examined in ascending document order with 0 rejected for short tokenization, giving exactly 2048 anchors with unique document ids and unique span hashes. The fitting and held-out halves share no document and no span. No quality filtering, no semantic filtering and no manual exclusions were applied. Raw text is not redistributed in the artifact; hashes and provenance are.

What this does not establish. The models did not change. This is evidence that the effect is not an artifact of WikiText or of token position 127; it is not evidence about MoE architectures beyond the four studied, and we do not describe it as architectural generalization anywhere.

F Strong predictors, oracle complementarity and the runtime selector

This appendix collects the studies that test routing history against predictors far stronger than a recent-only probe of the same family. Most of them returned negative verdicts, which are reported here as results.

F.1 The strong baselines

Two strong predictors define the operating point. A fixed-budget adaptation of a gate-reading rule reaches 0.859467 Recall@native-k (95% CI $[0.856867, 0.862042]$), and a history-conditioned candidate reranker, arm B of the reranking study and referred to throughout as HCR-B, reaches 0.862938 ($[0.860356, 0.865498]$). The history predictor reaches 0.668139 ($[0.664218, 0.672071]$). Both strong predictors read the gate parameters of the target layer; the history predictors never do. Pairwise over the 16,384 anchor-cell pairs, the gate-reading rule and HCR-B agree on 90.9%, while each exceeds the history predictor on about 62% and ties on about 23%.

F.2 Oracle complementarity

A pointwise oracle over the three predictors, choosing per token with hindsight, reaches 0.898102 ($[0.895962, 0.900253]$) against 0.871000 ($[0.868492, 0.873474]$) for an oracle over the two strong predictors alone. Total oracle headroom above HCR-B is $+0.035164$ ($[+0.033526, +0.036812]$), of which $+0.027102$ ($[+0.025640, +0.028600]$) is reachable only through the history arm. The history predictor is strictly best on 13.4% of anchor-cell pairs and is tied for best on a further 22.8%; where

it is strictly best, its mean advantage is $+0.2016$. An identity check confirms that the history-unique headroom matches $\mathrm{mean}(\max(0, u_H - \max(u_F, u_R)))$ to 6.7×10^{-9}. The frozen gate required headroom point estimates and interval lower bounds above fixed floors; all four conditions passed, giving COMPLEMENTARITY-SUFFICIENT. That outcome did not authorize the follow-up study, which required its own preregistration.

F.3 Matched attribution: how much of the oracle gap is history-specific

The raw $+2.71$ pp credits history with value that any additional predictor supplies. The attribution study adds two matched comparisons on the same universe. Substituting a recent-only third predictor of the same family raises the two-predictor oracle by $+0.021515$, generic third-predictor diversity. Removing it leaves $\Delta_{\text{matched}} = +0.005587$ (95% CI $[+0.004557, +0.006643]$). An oracle formed with the 20 shuffled-history controls instead of real history gives $\Delta_{\text{align}} = +0.006324$ ($[+0.005355, +0.007304]$); all 20 controls were retained (seeds 41001–41020, mean 0.891778, s.d. 2.19×10^{-4}). Both primary conditions passed, giving HISTORY-ATTRIBUTABLE-COMPLEMENTARITY-SUPPORTED. Δ_{matched} and Δ_{align} are the history-attributable quantities; Δ_{raw} is reported as secondary and is not attribution evidence.

Table 12: Oracle levels and the three attribution increments, per cell, in Recall@native-k. F is the gate-reading predictor, RR the HCR-B reranker, RH recent-plus-history, REC a recent-only third arm, and π(RH) the shuffled-history counterpart. The three increments are measured against different comparators and are not additive components of one total.

cell	$O(F, \mathrm{RR})$	$O(F, \mathrm{RR}, \mathrm{REC})$	$O(F, \mathrm{RR}, \mathrm{RH})$	$O(F, \mathrm{RR}, \pi(\mathrm{RH}))$	Δ_{raw}	Δ_{matched}	Δ_{align}
olmoe_m5	0.8256	0.8500	0.8534	0.8487	+0.0278	+0.0034	+0.0047
olmoe_m8	0.8779	0.8988	0.9010	0.8974	+0.0231	+0.0022	+0.0036
olmoe_m11	0.9111	0.9264	0.9292	0.9257	+0.0181	+0.0028	+0.0035
olmoe_m14	0.9102	0.9202	0.9221	0.9195	+0.0120	+0.0020	+0.0026
jetmoe_m7	0.8486	0.8726	0.8838	0.8732	+0.0352	+0.0112	+0.0106
jetmoe_m12	0.8799	0.9194	0.9355	0.9182	+0.0557	+0.0161	+0.0173
jetmoe_m17	0.8872	0.9189	0.9297	0.9181	+0.0425	+0.0107	+0.0116
jetmoe_m22	0.8525	0.8906	0.9019	0.8911	+0.0493	+0.0112	+0.0108
qwen_m7	0.8135	0.8403	0.8550	0.8391	+0.0415	+0.0146	+0.0159
qwen_m12	0.8948	0.9187	0.9219	0.9174	+0.0271	+0.0032	+0.0045
qwen_m17	0.8884	0.9128	0.9141	0.9124	+0.0256	+0.0012	+0.0017
qwen_m22	0.8660	0.8779	0.8801	0.8775	+0.0142	+0.0022	+0.0027
deepseek_m8	0.8551	0.8742	0.8758	0.8729	+0.0207	+0.0016	+0.0030
deepseek_m14	0.8771	0.8875	0.8905	0.8873	+0.0133	+0.0029	+0.0032
deepseek_m20	0.8714	0.8864	0.8879	0.8849	+0.0164	+0.0015	+0.0029
deepseek_m25	0.8766	0.8854	0.8879	0.8851	+0.0112	+0.0024	+0.0028

F.4 The runtime selector

The selector study asks whether runtime-observable signals can identify which frozen predictor is competent on the current token accurately enough to convert the oracle gap into real recall. Its answer was no.

Competence labels were taken verbatim from the frozen studies and verified to reproduce every frozen reference bitwise. The estimator sees exactly 18 features: 6 static geometry terms (k/E, m/L, a 4-way architecture one-hot), 6 per-predictor confidence terms (boundary margin and selected span, each scaled within the sample), and 6 pairwise top-k disagreement terms. It never sees an expert id, a full E-dimensional predictor vector, a target label, the true expert set, a hidden state, a token embedding or raw text. The model is a 1187-parameter MLP ($18 \to 32 \to 16 \to 3$, GELU, sigmoid) trained with uniform three-way MSE on continuous utilities rather than on a winner class, with 5 outer folds grouped by raw anchor id and an inner anchor-level split, three seeds averaged before any selector logic, and no cell of an evaluation anchor entering scaler fitting, training or early stopping. HNV was deliberately not a feature and no HNV ablation was trained: HNV tracks gain magnitude ($\rho \approx +0.547$) but not strict-win frequency ($\rho \approx -0.096$), and identity is what a selector requires.

Table 13: Runtime selector results. The first four rows are Recall@native-k levels; the last three are the derived quantities on which the frozen gate was evaluated. "Fraction of oracle gap recovered" is $(\mathrm{Recall}_{\mathrm{FRH}} - \mathrm{Recall}_{\mathrm{HCR\text{-}B}})/(O_{\mathrm{FRH}} - \mathrm{Recall}_{\mathrm{HCR\text{-}B}}) = 0.001831/0.035164$, i.e. the realized gain over the margin of the three-predictor oracle over HCR-B; the denominator is not the matched increment of §F. It is a unitless ratio, whereas the rows above it are Recall@native-k.

quantity	value	95% CI
HCR-B (operating baseline)	0.862938	[0.860356, 0.865498]
Selector over two strong predictors	0.862925	[0.860339, 0.865476]
Selector over all three	0.864769	[0.862144, 0.867358]
Oracle over all three	0.898102	[0.895962, 0.900253]
Gain vs HCR-B	+0.001831	[+0.001099, +0.002569]
Increment from admitting the history arm	+0.001844	[+0.001109, +0.002584]
Fraction of oracle gap recovered	0.052072	[0.031522, 0.072412]

The frozen gate had five conditions. Two passed (both interval lower bounds above zero) and three failed: the gain floor of $+0.005$, the history-actionable floor of $+0.002$, and the recovery floor of 0.20, with the entire recovery interval below it. The outcome is COMPETENCE-NOT-PREDICTABLE; no threshold was altered after results existed. Diagnostics: history wins on 13.4% of sample-cells, the selector chooses it on 3.0%, its recall of history wins is 0.0926, and its precision when choosing history is 0.418.

An attribution limit we did not resolve. The study ran two selector arms only, over the two strong predictors and over all three. It contains no symmetric arm in which the third predictor is recent-only or shuffled-history. Almost all of the $+0.1831$ pp comes from adding the history arm rather than from better arbitration between the strong pair, but without those symmetric controls the gain cannot be attributed specifically to *aligned older history* as opposed to the addition of a third, differently behaving predictor. We state the measured quantity and this limitation together.

F.5 Negative and closed branches on the strong-predictor side

FHR-P6.5 asked whether older history helps once the predictor is conditioned on the behaviour of a gate-reading rule. All three preregistered quantities came out negative: the conditioned history gain was -0.007624 (95% CI $[-0.009252, -0.006032]$), the alignment-specific version -0.002940 ($[-0.004307, -0.001547]$), and the test of whether HNV transfers to this setting returned HNV-DOES-NOT-CLEARLY-GENERALIZE-TO-FATE-UTILITY. This bounds the validated scope of HNV to the expert-set probe family, and §5 says so. **FHD-P6.6** diagnosed that failure descriptively (candidate coverage, miss ranks, margin quartiles, history boundary signal) without deriving a verdict, and confirmed that the three quantities keep their sign under every single-architecture omission.

LHR-P6.7 tested a local history reranker. Its primary quantity was -0.087072 (95% CI $[-0.089867, -0.084437]$): the reranker performed substantially *worse* than the baseline it was intended to improve, and the whole interval lies below zero. Its alignment control was positive ($+0.031164$), which is the informative combination, since aligned history beats shuffled history within a mechanism that nonetheless damages the baseline. The verdict is LOCAL-HISTORY-RERANKER-NOT-CONFIRMED with no partial-rescue alternative. **SUG-E1** and the swap-value exploration reused that data to ask whether the harmful swaps could be gated; they are exploratory and do not alter the verdict. **HPU-E1** and the **H2** coverage-versus-utility branch were closed, the latter by a recorded futility stop.

SRS-E1 compared setwise against candidate-wise scoring on a frozen candidate set and returned SETWISE-RECENT-NOT-PROMISING. **ACB-E0** asked whether adaptive candidate budgets are feasible and returned ADAPTIVE-BUDGET-NOT-FEASIBLE, with 2 of 3 frozen conditions holding; the third failed by $0.75k$ candidates, and the analytic mean oracle multiplier was 1.2943 against a required threshold of 2. **HCR-E1** is the study that produced HCR-B, and it altered no prior verdict.

None of these studies makes a latency, throughput, FLOP, memory-traffic or cache speedup claim, and neither does the main text.

G RICHER RECENT CONDITIONING AND THE LAG PROFILE

G.1 PROTOCOL, AND WHY IT IS NOT A CONTINUATION OF THE MAIN GRID

The richer-conditioning study widens the conditioning state while holding the older history input *fixed* at a single earlier selection state $H_{\text{old}} = S_{m-4}$:

$$G_1 = \mathbf{g}_{m-1}, \qquad G_2 = [\mathbf{g}_{m-1}; \mathbf{g}_{m-2}], \qquad G_3 = [\mathbf{g}_{m-1}; \mathbf{g}_{m-2}; \mathbf{g}_{m-3}],$$

with $\Delta_j = R^2([G_j; H_{\text{old}}]) - R^2(G_j)$ and $\Psi_j = R^2([G_j; H_{\text{old}}]) - R^2([G_j; \pi(H_{\text{old}})])$.

The history input here is one selection state; in the main grid it is three ($[S_{m-4}; S_{m-3}; S_{m-2}]$). The two protocols therefore differ in both arms, and Δ_1 is not the same estimand as Δ_G of the main grid. Plotting the main-grid result and this series as one curve would suggest a single protocol in which only the recent context varies, which is not what was run. Figure 2d shows this series alone, and no decay law is fitted, since three points cannot determine one.

This study used the same 16 cells, the same anchor ids and the same fitting/held-out split as the frozen grid, and ran no new MoE inference: it consumed stored captures that record scores by actual layer over the $m{-}4..m$ window. Preprocessing follows the frozen convention, verified bit-identical to the source studies. H_{old} is a single E-dimensional vector, so the PCA step that reduces a $3E$ stack to E in the main grid is not mechanically required and is not applied; this is recorded rather than silently varied. 34 unit tests and 46 pre-result checks passed before any statistic was computed, across 144 probe fits (3 window widths $\times$ 3 arms $\times$ 16 cells).

G.2 POOLED RESULTS

Table 14: Pooled richer-conditioning results. Δ_j is the direct increment over G_j alone and Ψ_j the increment over the same-capacity misaligned arm, both in held-out R^2, with j recent score vectors in the conditioning state and the older-history input fixed at the single state S_{m-4}. Intervals are the frozen paired bootstrap intervals.

quantity	$j = 1$	$j = 2$	$j = 3$
mean Δ_j	$+0.019532$	$+0.007589$	$+0.003138$
95% CI	$[+0.018129, +0.020912]$	$[+0.006578, +0.008560]$	$[+0.002359, +0.003867]$
mean Ψ_j	$+0.026207$	$+0.013374$	$+0.008671$
95% CI	$[+0.024764, +0.027667]$	$[+0.012349, +0.014415]$	$[+0.007883, +0.009477]$
cells with $\Delta_j > 0$	16/16	—	**5/16**

Both frozen labels hold at $j = 3$: the direct residual and the alignment-specific residual each have a positive pooled mean with an interval excluding zero. The ratio $\Delta_3/\Delta_1 = 0.1607$ is descriptive.

G.3 WHY THE POOLED RESIDUAL NEEDS THE ARCHITECTURE SPLIT BESIDE IT

Table 15: Richer conditioning by architecture, in held-out R^2. Only JetMoE keeps a positive mean Δ_3; the other three architectures average slightly negative, so the positive pooled residual at $j{=}3$ is carried by one architecture. Ψ_3 stays positive for all four, which is why the direct and alignment-specific increments must be read separately.

model	mean Δ_1	mean Δ_2	mean Δ_3	mean Ψ_3	cells $\Delta_3 > 0$
OLMoE	$+0.006917$	$+0.000433$	-0.002473	$+0.004139$	0/4
JetMoE	$+0.041464$	$+0.024100$	$+0.016371$	$+0.018489$	4/4
Qwen1.5	$+0.017183$	$+0.003467$	-0.000466	$+0.006939$	1/4
DeepSeek	$+0.012565$	$+0.002358$	-0.000879	$+0.005118$	0/4

Only 5 of 16 cells have positive Δ_3, and three of the four architectures have negative mean Δ_3. The pooled interval excludes zero because the 16-cell mean is equally weighted and the JetMoE effect is large enough to carry it. For OLMoE, Qwen1.5 and DeepSeek, three recent score states absorb the

older-history signal that one recent state leaves visible. Reporting only the pooled $+0.0031$, or only the uniformly positive Ψ_3, would misdescribe the result, so the main text carries the cell count and the split.

Note also that Ψ_3 remains positive in the pool for every architecture while Δ_3 does not. Aligned history still beats distribution-matched misaligned history at $j = 3$, but it no longer reliably beats the recent-only predictor. That is exactly the distinction §3 keeps separate, and this is where it matters most.

G.4 ALL 16 CELLS

Table 16: Richer-conditioning results, all 16 cells, in held-out R^2. Δ_j is positive in 16/16 cells at j=1 but only 5/16 at j=3; the older-history input is fixed at S_{m-4} throughout, so this is a different protocol from the three-layer H of the main grid.

model	m	Δ_1	Δ_2	Δ_3	Ψ_1	Ψ_2	Ψ_3
OLMoE	5	+0.0045	-0.0003	-0.0054	+0.0157	+0.0100	+0.0050
OLMoE	8	+0.0103	+0.0019	-0.0015	+0.0178	+0.0082	+0.0046
OLMoE	11	+0.0085	+0.0009	-0.0012	+0.0147	+0.0063	+0.0042
OLMoE	14	+0.0044	-0.0007	-0.0018	+0.0098	+0.0039	+0.0027
JetMoE	7	+0.0275	+0.0208	+0.0137	+0.0312	+0.0248	+0.0177
JetMoE	12	+0.0396	+0.0118	+0.0094	+0.0421	+0.0133	+0.0111
JetMoE	17	+0.0586	+0.0346	+0.0237	+0.0612	+0.0374	+0.0259
JetMoE	22	+0.0402	+0.0293	+0.0187	+0.0412	+0.0303	+0.0193
Qwen1.5	7	+0.0248	+0.0054	-0.0003	+0.0365	+0.0160	+0.0097
Qwen1.5	12	+0.0122	+0.0024	-0.0003	+0.0191	+0.0076	+0.0048
Qwen1.5	17	+0.0159	+0.0031	+0.0001	+0.0237	+0.0098	+0.0065
Qwen1.5	22	+0.0158	+0.0030	-0.0014	+0.0255	+0.0119	+0.0067
DeepSeek	8	+0.0093	+0.0017	-0.0008	+0.0154	+0.0070	+0.0045
DeepSeek	14	+0.0146	+0.0028	-0.0001	+0.0215	+0.0087	+0.0054
DeepSeek	20	+0.0109	+0.0021	-0.0018	+0.0204	+0.0096	+0.0057
DeepSeek	25	+0.0155	+0.0028	-0.0009	+0.0236	+0.0092	+0.0049

G.5 POST-RESULT CHANGE, RECORDED RATHER THAN HIDDEN

One post-result edit exists in this study: a report-layout constant governing the number of report items was corrected from 22 to 26, because the original value undercounted the per-window items and the architecture summaries. It selects no cell, arm, seed, probe or statistic and changes no number above; the run was not repeated. The method files, probe, checks and tests are byte-identical to their pre-result state, no architecture or cell was removed, no window width was added or dropped, and no inference was rerun.

G.6 THE LAG PROFILE

An earlier descriptive study profiled predictive value against lag, adding one older selection state at a time. Its cross-target descriptive summary was near lags $+0.10392$, far lags $+0.06146$, and near minus far $+0.04246$. It produced no binary verdict, and its Spearman coefficients are descriptive with no p-values used for any label. It is reported here for completeness and no claim in the main text rests on it.

H EARLY EVIDENCE, EXPLORATORY PILOTS AND CLOSED BRANCHES

These studies preceded or sit beside the main grid. They are grouped by the scientific question they asked rather than by date. Several are exploratory or were closed by their own rules; those are marked and are not used to support any claim in the main text.

H.1 Cross-model replication and mechanism disambiguation

RMC-P0 asked whether predictive routing dependence beyond the immediately preceding layer replicates in an architecture distant from the one where it was first observed, with the verdict rule fixed before any R^2 existed. In JetMoE at layers 12 and 20, $\Delta_{\text{RMC}} = R^2([\text{recent}; \text{history}]) - R^2(\text{recent})$ was $+0.142748$ (95% CI lower bound $+0.109977$) and $+0.205277$ (lower bound $+0.168911$), both above the preregistered threshold of 0.02, and all six frozen conditions held: **REPLICATED**. This study established the phenomenon that the main grid later measured under a same-capacity control.

NHD-P0 asked whether that gain survives a *nonlinear* recent-state predictor, which is the obvious alternative explanation: perhaps older history only appears useful because a linear probe cannot exploit the recent state fully. On the frozen JetMoE artifacts at layer 12 the held-out R^2 values were: linear recent probe $+0.20972$; nonlinear recent predictor of the same input width $+0.21111$; capacity-matched nonlinear recent predictor $+0.21809$ (1317 parameters); nonlinear predictor given history $+0.38249$ (1320 parameters, matched mechanically to within 3 parameters before any held-out result). Two increments follow, against two different baselines, and they must not be conflated. Nonlinearity alone bought $+0.00139$ over the linear recent probe, which we write δ_{nl}, comparing like input width. Adding history bought $+0.16440$ over the *capacity-matched* nonlinear recent predictor, its δ_{hist}. (Against the unmatched nonlinear recent predictor the history gain is $+0.17137$; the matched figure is the conservative one and the one we quote.) Three seeds, early stopping on a split carved out of the fitting half, held-out data never used for stopping or architecture choice. The history gain is therefore not a substitute for missing nonlinearity in the recent state: nonlinearity is worth roughly a hundredth of what history is worth here.

CAMR-P3 attempted the same replication on a wider router using a zero-block capacity control and returned **THIRD-MODEL-NOT-REPLICATED** (0 of 2 targets). That verdict is permanent under its own protocol. Its measured zero-block estimation asymmetry ($\Delta_{\text{full}} = -0.02088$ on the synthetic null at $E = 60$, where only the history arm pays for E additional fitted columns) is the direct reason why the main grid uses row-permuted controls, and it is also why the wider-router architectures reappear with positive Ψ in the main grid: under a same-capacity control, aligned history outperforms distribution-matched shuffled history in those models, which is a narrower statement than the one CAMR-P3 tested and does not reinterpret its verdict.

H.2 Provenance: what carries the signal

EIPC-P0 asked whether explicit expert provenance, which standard MoE aggregation discards by fusing expert outputs, carries more information about future router states than the same history after early fusion. It found that it does, by $+0.109$ R^2 at layer 8 and $+0.175$ at layer 12 in OLMoE, also against an identity-destroyed control. **EPD-P0** then decomposed that signal into path versus content components; it is explicitly exploratory, builds no method, and reports descriptively. **RMO-P0** is a post-hoc pilot on the already-observed EPD-P0 data asking whether a strong selection-path signal is mostly trivial local persistence; it is CPU-only and exploratory, and produces no formal verdict. **XEC-P0** tested whether a cross-layer expert cache is actionable and was **closed**: the earlier provenance result did not show that a cache improves anything, and this study did not establish that it does.

H.3 Rejected experts: an uninterpretable pilot and its control

REDV-V1 asked whether near-miss experts rejected by top-k at one layer carry incremental information about counterfactual routing regret at the next. Its design lacked a matched control, so an apparent effect cannot be separated from the additional capacity that its design added. The frozen record marks it **uninterpretable**, and it is listed here for completeness and used as evidence for nothing. **REDV-V2** added a matched rejection control and the branch was **closed**. **DREV-P0**, a screening pilot asking whether rejected evidence pays off several layers later rather than immediately, was also **closed**.

Recording these as closed matters for reading the rest of the paper: the history signal we do measure comes from *selected* experts, and the parallel question about rejected experts was tested and did not produce usable evidence.

H.4 Full recent state

FRS-P1 is described in Appendix B. Its role is to rule out the explanation that the effect exists only because the conditioning state was a reduced or binarized version of the recent router state: with both arms held to exactly 16 dimensions and G bitwise identical in the first 8 coordinates of both, held-out R^2 still rose by $+0.1285$ and $+0.2021$ at the two layers tested. It is a single-model result and is not the basis of any cross-architecture claim.

H.5 Interpreting this group as a whole

Two of these studies (RMC-P0, NHD-P0) supply the positive groundwork for the main grid, one (FRS-P1) closes an obvious alternative explanation, one (CAMR-P3) is a negative result whose control artifact reshaped the main design, one (EIPC-P0) locates the signal in provenance, and five are exploratory or closed. We report them as a set of results with different statuses rather than as a narrative of progress: the closed and uninterpretable entries are outcomes of the same protocol discipline that produced the positive ones.

I Interactive browsing of cross-layer expert selections

The anonymous interactive artifact at https://anonymous-artifact.net/ allows a reader to follow a single token through the routed layers of a frozen MoE backbone and to see which experts were selected at each one. Selections are browsed by architecture, corpus, sample anchor, token position, target layer, and the span of preceding layers to display, so the conditioning structure examined in this paper can be inspected one token at a time rather than only in aggregate. Code, configurations and frozen result tables are available separately at https://anonymous.4open.science/r/moe-routing-anonymous-artifact-FD5A; the site is a browsing interface over the same frozen material.

Displayed nodes are the native expert selections stored in the frozen captures, read directly as binary occupancy rather than recomputed from scores in the browser, so what a reader sees is the selection made by the model itself. The published display data was checked against the frozen sources: across 60 token traces, the expert selections, router scores and sample coordinates agree with the frozen records. That check covers the traceability and fidelity of the displayed traces, not the full experiment or every interface feature. Highlighted nodes encode recorded native expert selections, whereas the all-to-all visual connectors between adjacent displayed selection sets support cross-layer comparison; the fainter background curves are schematic and serve the same comparative purpose.

The example in Figure 3 shows OLMoE on WikiText at anchor 2048, token position 127, with target layer $m = 14$. Layers 6 through 14 are displayed, of which the eight preceding layers form the browsing span rather than the model input: the main analysis conditions on $G = \mathbf{g}_{13}$ with history $H = [S_{10}; S_{11}; S_{12}]$. The node view presents discrete selections, so the continuous score vector G is not fully rendered there.

J Coverage map and figure provenance

J.1 Every study has a location

J.2 Figure provenance

Figure 1 is a hand-drawn schematic of the design and plots no data, and Figure 3 is a screen capture of the browsing interface described in §I. Every data figure is produced by a script that reads frozen artifacts and writes a CSV of exactly the plotted values beside the figure. No figure content was produced by a generative image model.

Appendix tables are generated by a companion script from the same artifacts: Table 5 from RSS-P4; Table 6 from HCR-P5 and HAP-P6; Table 7 from HAP-P6 `main_comparison.csv`; Table 11 from the ECV-E1 per-cell CSVs; Table 12 from RAC-E1 `experiment_a.json`; and Table 16 from the MRS-E1 cell table. Tables 1 and 2 in the main text are generated by the same script set.

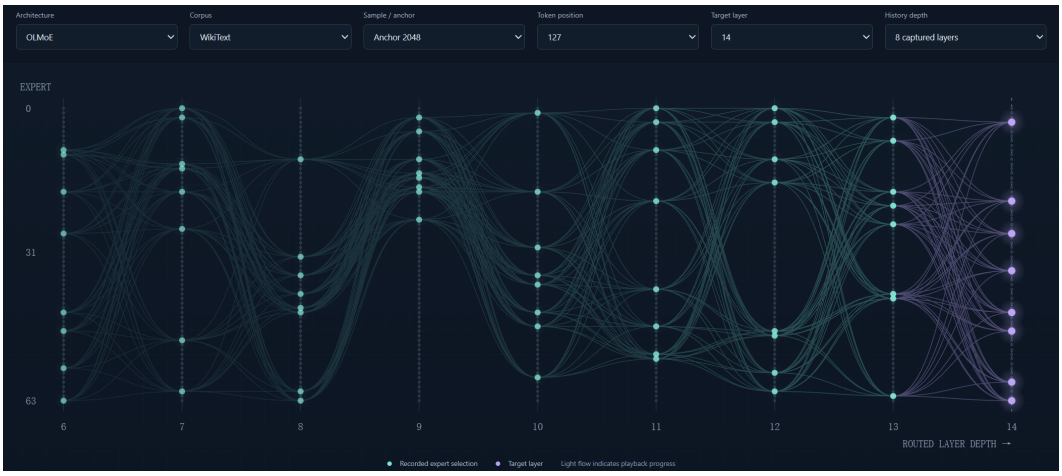

Figure 3: Interactive browsing of cross-layer expert selections for one token: OLMoE on WikiText, anchor 2048, token position 127, target layer $m = 14$, with recorded layers 6–14 displayed. Cyan nodes mark the experts selected at each layer under the native top-k rule of the model; purple nodes mark the selection at the target layer. Node values are read from the frozen captures as binary occupancy. Connectors between adjacent displayed selection sets are visual aids for comparing sets across layers, and equal expert indices at different layers do not denote the same expert.

Table 17: Coverage map: every recorded study and where its quantities are reported.

Study	Reported in	Which quantities
RSS-P4	§5, C	A, R, B, Ψ per cell; $\mathrm{Spearman}(A, \Psi)$
RAC-E1-B	§5, C	mean Δ_G and CI; per-architecture ranges
HCR-P5	§5, C, D	HNV definition, per-cell HNV, $\mathrm{Spearman}(A, \mathrm{HNV})$, RER, synthetic validation
HAP-P6	§5, C, D	recall levels, U, $\mathrm{Spearman}(\mathrm{HNV}, U)$, baselines, budget table, ablations, leave-one-out
ECV-E1	§5, E	external Δ_G, Ψ, U, U^{align}, per-position and per-cell
MRS-E1	§5, G	$\Delta_{1,2,3}$, $\Psi_{1,2,3}$, per-cell, architecture split, post-result edit
C0-RPC	§5.5, F	oracle levels, total and history-unique headroom, win rates, D_i, W_i
RAC-E1-A	§5.5, F	Δ_{raw}, Δ_{matched}, Δ_{align}, generic diversity, per cell
C1-RCE	§5.5, F	selector recalls, gain, recovery, 5-condition gate, features, diagnostics, attribution limit
HCR-E1	§5.5, F	HCR-B operating baseline
FHR-P6.5	§5, F	three negative verdicts incl. HNV non-transfer
FHD-P6.6	F	descriptive diagnosis, omission checks
LHR-P6.7	F	Δ_{LHR}, alignment control, verdict
SRS-E1	F	setwise vs candidate-wise outcome
ACB-E0	F	3-condition gate, oracle multiplier
SUG-E1, HPU-E1, H2	F	exploratory gating, pair utility, futility stop
FRS-P1	B, H	full-recent-state control, two layers
CAMR-P3	B, H	negative verdict; zero-block estimation asymmetry
RMC-P0	H	cross-model Δ_{RMC}, replication verdict
NHD-P0	H	nonlinear recent control, capacity-matched gain
EIPC-P0, EPD-P0, RMO-P0, XEC-P0	H	provenance, decomposition, order pilot, cache closure
REDV-V1/V2, DREV-P0	H	uninterpretable pilot and its closures
RHP-P2	G	descriptive lag profile

Table 18: Figure provenance: each figure, its companion CSV of plotted values where one exists, and the frozen artifacts it reads.

Figure or table	Companion data	Source artifacts
1	— (uploaded vector figure)	schematic of the design in §3; carries no plotted data
2	`fig1_results.csv`	HAP-P6 `hnv_vs_history_utility.csv`, `results.json`; ECV-E1 `external_per_position.csv`; MRS-E1 cell table
1	generated in place	RSS-P4 `cell_results.csv`; RAC-E1 `experiment_b.json`; HAP-P6 `cell_results.csv`, `results.json`
2	generated in place	RAC-E1 `experiment_a.json`; C1-RCE `results.json`
3	— (interface screen capture)	display data of the browsing interface, read from the frozen captures as binary occupancy (§I)

J.3 TERMINOLOGY

Router scores are pre-softmax gate logits. *Selection history* means the binary native top-k indicator vectors of earlier routed layers for the same token. *Expert-set recall* is Recall@native-k against the set selected by the model itself. *Direct increment* compares aligned history to the recent-only predictor; *alignment-specific value* compares it to the same-capacity misaligned control. *Conditional complementarity* refers to per-token oracle gains over a strong predictor, always reported with the matched and alignment controls beside the raw figure. *Oracle* always denotes a hindsight upper bound, never an achievable rule. Internal study identifiers are used only in this appendix and in the reproducibility mapping.