MoE Routing Is Not Token-Local: How Expert Influence Propagates Across Tokens & Layers

Anonymous authors
Paper under double-blind review

Abstract

In Sparse Mixture-of-Experts (MoE) models, routing is typically analyzed at the level of individual tokens. We instead trace how an expert's computation on one token affects the routing decisions of other tokens. We show that routing influence propagates across tokens and layers through two causal channels, a direct same-token channel and a context-mediated inter-token channel. Our evaluation on five MoE models reveals that: (i) the two channels combine approximately linearly in routing-margin space before the first routing divergence; (ii) they propagate differently with depth, as direct influence acts immediately and fades while context-mediated influence accumulates downstream; and (iii) the two channels are typically redundant rather than synergistic, with either pathway alone often sufficient to cross a routing boundary. To formalize these findings, we examine MoE routing boundaries and introduce a first-order, margin-normalized routing leverage. Routing leverage makes explicit that whether a Top-K boundary is crossed depends on how an expert's contribution aligns with the contrast direction between the competing experts, relative to the margin separating them, and it predicts the layer at which a token's routing first diverges, with accuracy increasing as more of the causal path is included. Discrete Top-K selection turns these continuous, additive effects into nonlinear changes in computation, and the first divergence marks the onset of a routing cascade.[1]

1 Introduction

Sparse Mixture-of-Experts (MoE) architectures have been widely integrated into the design of frontier large language models (LLMs), such as DeepSeek-V4.1-Flash (DeepSeek-AI, 2026), Qwen3.8-Max (Qwen Team, 2026), and Mistral Small 4 (Mistral AI, 2026). Their effectiveness stems from sparsity: by routing each token to only a small subset of experts, MoE models keep per-token computation nearly constant while scaling to hundreds of billions or even trillions of parameters, decoupling computational cost from representational capacity. This sparsity is realized by the router, which selects the experts for each token and thereby determines which computation the model applies to it. Understanding an MoE's computation therefore requires knowing not only what individual experts compute, but also how their outputs influence future routing decisions across the network.

Existing interpretability work has studied expert specialization (Herbst et al., 2026), expert importance (Su et al., 2026), and the expert paths followed by individual tokens (Gu et al., 2026; Yoon et al., 2026). These approaches take a token-local view, interpreting experts by their incoming tokens and token trajectories by their selected expert sequences.

This perspective is useful for specialization and token-level computation, but it leaves the cross-token structure of routing unexamined. That attention transports information across tokens follows from the architecture. What has not been established is how much of a routing decision it accounts for, how that compares with what the token receives through its own residual stream, and how those two influences combine and propagate with depth.

[1] Code available at `https://anonymous.4open.science/r/MoE_influence-4377`.

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.

As illustrated in Figure 1, ablating an expert at a single token position causes later tokens that never routed to the ablated expert to diverge from their original trajectory, meaning that their Top-K expert set differs from the clean forward pass at some later layer. Across five models, expert ablations change the downstream routing of a median 52–81% of tokens that never routed to the ablated experts, and in a separate patching experiment, restoring the clean keys and values of the context removes every such routing change. Routing therefore cannot be fully characterized as a collection of independent token-level decisions. Interpreting an expert by the tokens it receives omits the routing it changes at tokens it never receives, and reading a token's expert sequence as a property of that token omits the upstream experts that shaped it.

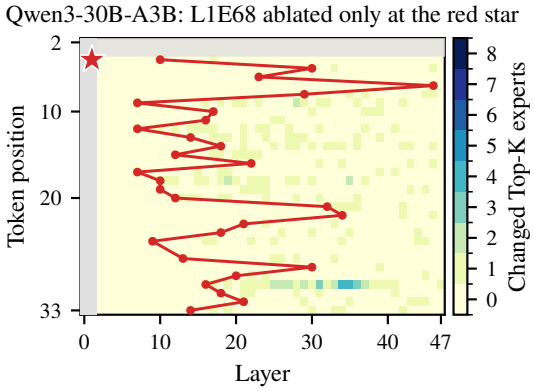


Figure 1: Cross-token routing influence spread. The red line marks the frontier of the first routing divergences.

The main contributions of this work are as follows:

- We quantify how far an expert's influence reaches beyond the token on which it executes, and causally isolate the two channels that carry it: a direct same-token channel and a context-mediated inter-token channel.
- We provide a mechanistic account of how these channels propagate and interact, showing that their effects combine approximately linearly in routing-margin space before the first Top-K divergence, after which discrete routing changes initiate a nonlinear routing cascade.
- We introduce *routing leverage*, a margin-normalized estimator that formulates the interplay of perturbation magnitude, alignment with the router's contrast direction, and boundary vulnerability, and predicts from clean-pass quantities alone where clean-path propagation first breaks into a routing divergence and the routing cascade begins.

2 Related Work

Expert importance and routing paths. Expert importance is typically assigned from properties of the expert itself, such as its activation magnitude (Su et al., 2026), load (Muzio et al., 2024), or sensitivity to quantization (Chitty-Venkata et al., 2025). Gu et al. (2026) study the sequences of experts traversed by individual tokens, showing that routing paths can exhibit meaningful structure and semantic information. A complementary line of work studies routing counterfactually, asking whether the trained router selects an optimal expert under a fixed compute budget (Yoon et al., 2026). This is a normative question about which route should be chosen, whereas we ask what determines the route actually taken.

Influences on routing. A closely related line of work studies how routing changes when the context of a token is changed: Olson et al. (2025) show that adding context which alters the meaning of a fixed token can route that token to different experts. Li et al. (2026) decompose routing scores into contributions from upstream layers, attention heads and experts, using variance over expert scores to quantify their influence. We instead intervene on an internal component, causally isolating the effect of an upstream expert and tracing how its influence reaches other routers through the inter-token channel. Ahrac et al. (2026) study how router and expert directions become geometrically coupled during training, the geometry on which our criterion operates: routing leverage projects an expert's contribution onto the contrast direction between two competing experts and asks whether it consumes the margin between them.

3 Preliminaries

MoE-enhanced transformers. Let N denote the number of experts per layer, L the number of transformer layers, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_S \in \mathbb{R}^d$ an input sequence of length S. We consider a causal MoE-enhanced transformer in which each layer applies a self-attention block followed by an MoE

block. For token i at layer ℓ, the respective block inputs are $\boldsymbol{a}^{\ell}_{\text{in},i} = \text{LN}^{\ell}_1(\boldsymbol{x}^{\ell}_i)$ and $\boldsymbol{m}^{\ell}_{\text{in},i} = \text{LN}^{\ell}_2(\boldsymbol{x}^{\ell}_i + \boldsymbol{a}^{\ell}_{\text{out},i})$. The resulting layer recurrence is

$$\boldsymbol{x}^{\ell+1}_i = \boldsymbol{x}^{\ell}_i + \boldsymbol{a}^{\ell}_{\text{out},i} + \boldsymbol{m}^{\ell}_{\text{out},i}, \qquad \ell = 0, \ldots, L-1, \tag{1}$$

where $\boldsymbol{x}^0_i = \boldsymbol{x}_i$. Following standard RMSNorm, inputs $\boldsymbol{x} \in \mathbb{R}^d$ are transformed as $\text{LN}^{\ell}_b(\boldsymbol{x}) = \boldsymbol{\gamma}^{\ell}_b \odot \frac{\boldsymbol{x}}{\text{RMS}(\boldsymbol{x})}$, with $b \in \{1, 2\}$, learned parameters $\boldsymbol{\gamma}^{\ell}_b \in \mathbb{R}^d$, and root mean square $\text{RMS}(\boldsymbol{x})$.

For simplicity, we assume multi-head self-attention (MHSA), with per-head queries, keys and values $\boldsymbol{q}^{\ell,h}_i, \boldsymbol{k}^{\ell,h}_j, \boldsymbol{v}^{\ell,h}_j$ and attention weights $p^{\ell,h}_{i,j}$. Appendix A formally defines $\boldsymbol{a}^{\ell}_{\text{out},i}$ and $\boldsymbol{m}^{\ell}_{\text{out},i}$.

Routing. We consider a linear router $\boldsymbol{W}^{\ell}_G \in \mathbb{R}^{N \times d}$. Defining $\boldsymbol{g}^{\ell,n} := (\boldsymbol{W}^{\ell}_G)^{\top}_{n,:}$, the routing logits, the logit of expert n, and the corresponding routing weights are

$$\boldsymbol{z}^{\ell}_i = \boldsymbol{W}^{\ell}_G \boldsymbol{m}^{\ell}_{\text{in},i}, \qquad z^{\ell,n}_i = \boldsymbol{g}^{\ell,n} \cdot \boldsymbol{m}^{\ell}_{\text{in},i}, \qquad w^{\ell,n}_i = \text{softmax}(\boldsymbol{z}^{\ell}_i)_n, \tag{2}$$

where the weights may optionally be renormalized over the selected Top-K experts. The K selected experts are the indices of the K largest entries of $\boldsymbol{z}^{\ell}_i$, denoted by TopK^{ℓ}_i, and the block output $\boldsymbol{m}^{\ell}_{\text{out},i}$ is the sum of their outputs $\boldsymbol{e}^{\ell,n}_{\text{out},i}$, each weighted by its routing weight.

Routing score decomposition. Because the router reads the residual stream, each routing logit can be decomposed into contributions from the components that wrote into it. Following Li et al. (2026), the routing logit of expert (ℓ, n) for token i can be written as

$$\begin{aligned} z^{\ell,n}_i &= \boldsymbol{g}^{\ell,n} \cdot \text{LN}^{\ell}_2(\boldsymbol{x}^0_i + \sum_{c=0}^{\ell} \boldsymbol{a}^{c}_{\text{out},i} + \sum_{c=0}^{\ell-1} \boldsymbol{m}^{c}_{\text{out},i}) \\ &= \boldsymbol{g}^{\ell,n} \cdot \overline{\text{LN}}^{\ell}_{2,i}(\boldsymbol{x}^0_i) + \sum_{c=0}^{\ell} \boldsymbol{g}^{\ell,n} \cdot \overline{\text{LN}}^{\ell}_{2,i}(\boldsymbol{a}^{c}_{\text{out},i}) + \sum_{c=0}^{\ell-1} \boldsymbol{g}^{\ell,n} \cdot \overline{\text{LN}}^{\ell}_{2,i}(\boldsymbol{m}^{c}_{\text{out},i}), \end{aligned} \tag{3}$$

where $\overline{\text{LN}}^{\ell}_{2,i}(\boldsymbol{c}) = \boldsymbol{\gamma}^{\ell}_2 \odot \frac{\boldsymbol{c}}{\text{RMS}(\boldsymbol{r}^{\ell}_i)}$ fixes the denominator at the original residual input $\boldsymbol{r}^{\ell}_i = \boldsymbol{x}^0_i + \sum_{c=0}^{\ell} \boldsymbol{a}^{c}_{\text{out},i} + \sum_{c=0}^{\ell-1} \boldsymbol{m}^{c}_{\text{out},i}$. This turns RMSNorm into a linear map, allowing the routing logit to be decomposed additively across residual components. The decomposition is exact at $\boldsymbol{r}^{\ell}_i$.

4 Causal channels of routing influence

An expert's output is added to its token's residual stream, creating two possible routes by which it may affect downstream routing: along the same token's residual path, or through causal self-attention to subsequent token positions. We isolate these routes by controlling the token positions at which the expert's output is removed, allowing us to separate direct same-token influence from context-mediated inter-token influence.

Interventions. We ablate an expert by zeroing its output, without removing it from the Top-K or renormalizing the routing weights. We say that token i *flips*[2] at MoE block ℓ if its Top-K set differs from that of the clean forward pass for the first time at ℓ. Under causal attention, an ablation reaches the router of a token i through only two routes: its own residual stream, the *direct same-token channel*, or the streams of earlier tokens via attention, the *context-mediated inter-token channel*. We separate them by ablating an expert u at layer ℓ^u, whose weighted output at token j is $w^u_j \boldsymbol{e}^u_{\text{out},j}$, at a set of positions P:

- **Direct**: $P = \{i\}$, isolating the direct channel.
- **Context**: $P = \{j < i : u \in \text{TopK}^{\ell^u}_j\}$, isolating the context-mediated channel.
- **Joint**: $P = \{j \leq i : u \in \text{TopK}^{\ell^u}_j\}$, both together.

The direct and context interventions are applied separately to the same token, so routing at i may diverge under both. When several experts are ablated together, P is defined per expert.

Setup. We evaluate five MoE models covering diverse architectures and scales: OLMoE (Muennighoff et al., 2025), DeepSeek-V2-Lite (Liu et al., 2024), Qwen3-30B-A3B (Yang et al., 2025),

[2]We use the terms "flip" and "first routing divergence" interchangeably throughout.

Mixtral-8x7B (Jiang et al., 2024), and Phi-3.5-MoE (Abdin et al., 2024) (Appendix B). Unless stated otherwise, all interventions use the same 500 C4 prompts (Raffel et al., 2020) (Appendix M), each truncated to 128 tokens, and run in `float32`, which reduces the numerical noise that would otherwise flip near-tied Top-K boundaries. The ablated experts vary in depth and load (Appendix E). Those marked * belong to coalitions identified with an expert-to-expert influence graph (Appendix C), whose joint ablation raises NLL by 3.9–33.9× the sum of the members' individual ablations; reporting them alongside unmarked experts tests whether experts whose removal damages the model also exert unusual routing influence.

4.1 Nonlocal routing influence

Although attention ensures that an ablation perturbs the hidden states of attending tokens, Top-K routing changes only when these perturbations cross routing boundaries. The prevalence of such crossings therefore determines the extent to which expert influence propagates beyond the perturbed token. We quantify this effect as the fraction of tokens that flip under the context intervention.

Table 1: Context-mediated routing flips.

Model	Ablations	Routing flips	Median first flip (layers)	Restored by KV patch
Mixtral-8x7B	6	51.8%	11	100%
Phi-3.5-MoE	8	59.3%	12	100%
OLMoE	9	60.0%	4	100%
DeepSeek-V2-Lite	5	71.0%	3	100%
Qwen3-30B-A3B	10	80.6%	9.5	100%

As shown in Table 1, such context-mediated flips are common: across models, a median of 51.8% to 80.6% of these tokens change their routing, with the first flip occurring 3 to 12 layers after the ablation. Patching the keys and values of the attended tokens back to their clean values restores every context-mediated routing decision (Section 4.3). These results show that expert computation is not token-local: the attention-mediated channel propagates expert influence across tokens, substantially altering the routing of tokens that never selected the expert.

4.2 A taxonomy of routing flips

Section 4.1 considered tokens whose routing can only be affected through the context pathway. For tokens that selected the ablated expert, both pathways are available. We therefore classify flips as direct, context-mediated, or requiring both pathways. For each sequence, we uniformly sample one token preceded by another token that also selected the expert and apply all three interventions.

Table 6 (Appendix F) partitions the outcomes into six mutually exclusive categories according to which interventions change the token's routing. *Direct only* denotes tokens that flip under the direct intervention but not the context one, *Context only* the converse, and *Both alone* tokens that flip under both; all three also flip under the joint intervention. *Both required* flips only under the joint intervention, and *No change* under none.

The context route remains substantial even when a direct route is present: *Context only* and *Both alone* together account for more than 70% of analyzed tokens in 10 of 18 ablations and more than 95% in six. The two channels are therefore usually redundant routes to a routing divergence, rather than synergistic ones that must act together. If their combined effect were typically needed to cross a routing margin, *Both required* would be common; instead, it remains below 3% in every ablation, while *Both alone* is the largest category in 12 of 18 and *Direct only* in one. Thus, either route alone typically suffices to induce a routing divergence.

4.3 Attention rescue

Attention is the only architectural pathway between tokens, but MoE implementations batch tokens by their selected experts, so rerouting one token can in principle alter the numerical computation at another. We rule this out by applying the context intervention and restoring the attended tokens' keys and values to their clean values. Table 7 (Appendix G) reports the share of context-mediated flips recovered by patching both, only the values, or only the keys.

Restoring both removes every context-mediated flip across all five models and introduces none, confirming that the effect is mediated by attention. Patching either alone recovers only part of the

flips: the stronger restores 12%–64%, with values outweighing keys in four of five models. Partial patches also introduce some new flips, since mixing clean and perturbed ones is itself a perturbation.

Thus, context-mediated influence depends on both what earlier tokens contribute and how strongly they are attended to. We arrive at the causal chain: ablating expert u removes $w_j^u \boldsymbol{e}_{\mathrm{out},j}^u$ from the residual stream of each context token $j < i$ that selected it, perturbing its keys and values, $\delta \boldsymbol{x}_j \rightarrow \delta \boldsymbol{k}_j, \delta \boldsymbol{v}_j$. At the first layer after the ablation, the residual stream of the analyzed token i remains clean, so these perturbed keys and values are the only pathway to it. Through causal attention, they perturb its attention output, $\rightarrow \delta \boldsymbol{a}_{\mathrm{out},i}$, which enters its residual stream and MoE input, $\rightarrow \delta \boldsymbol{x}_i \rightarrow \delta \boldsymbol{m}_{\mathrm{in},i}$, eventually crossing its Top-K boundary.

5 Propagation of routing influence

We next disentangle the direct and context pathways at the routing boundary, quantifying their individual contributions and their combined effect up to the first routing divergence.

5.1 Channel separability before routing divergence

The joint intervention removes the writes of both channels together. If the margin displacement is approximately linear in the removed writes, the displacements of the two channels should therefore add, up to the first routing divergence.

For an analyzed token i and layer ℓ, let $m_i^\ell(a,b) = z_i^{\ell,a} - z_i^{\ell,b}$ be the routing margin between the lowest-ranked selected expert a and the highest-ranked unselected expert b. We measure the margin displacement under intervention as $\delta = m_i^\ell(a,b) - \tilde{m}_i^\ell(a,b)$. Since a and b are fixed from the clean run, all three interventions are directly comparable, and additivity predicts $\delta_{\mathrm{joint}} \approx \delta_{\mathrm{dir}} + \delta_{\mathrm{ctx}}$. We quantify the additivity error as a fraction of the margin, $|\delta_{\mathrm{dir}} + \delta_{\mathrm{ctx}} - \delta_{\mathrm{joint}}| / m_i^\ell(a,b)$.

The measured displacements agree closely with the additive prediction (Table 8, Appendix H): in 20 of 22 ablations, the median error is at most 2.1% and the 90th percentile remains below 20.1%. The Pearson correlation between $\delta_{\mathrm{dir}} + \delta_{\mathrm{ctx}}$ and δ_{joint} exceeds 0.99 in all models except Phi. Channel additivity also holds under multi-expert ablations, where the median error across all four evaluated expert sets remains bounded by that of the worst individual expert.

These results show that, despite the nonlinear computation along the path, the direct and context-mediated channels combine approximately linearly in continuous margin space; the dominant nonlinearity arises when the Top-K operation converts this displacement into a discrete route flip. This also explains the taxonomy of Section 4.2: because either channel alone often displaces the boundary far enough, flips rarely require both.

5.2 Layerwise propagation dynamics

We next examine how the two channels propagate across layers. Because direct influence enters a token's own residual stream at the ablated layer while context-mediated influence must first travel from earlier positions through attention, we expect the share of tokens that diverge under the direct intervention to peak immediately and decay, and the context share to emerge gradually, accumulate downstream, and depend on how many layers remain below the ablation.

To test this we ablate one expert at each of five layers per model, all chosen at the same routing load, so depth is the only quantity varying (Appendix E). Figure 2 reports both channels as a function of the relative depth of the ablation.

In all 23 ablations in which the context-mediated channel induced a flip at all, its median first flip is at least as deep as the direct one, strictly so in twenty. Among the 21 ablations with at least three layers below them, the share of tokens diverging under the direct intervention peaks immediately in 16 and has decayed by the last layer in 19, whereas the context share accumulates after the first layer in 19, starting from zero in three of them and otherwise growing to 1.4–8× its initial value (Table 5).

Downstream depth limits the context channel: as the ablation moves deeper, the share of tokens diverted by the context channel alone falls in every model, from 5.6–20.2% at the shallowest ablation

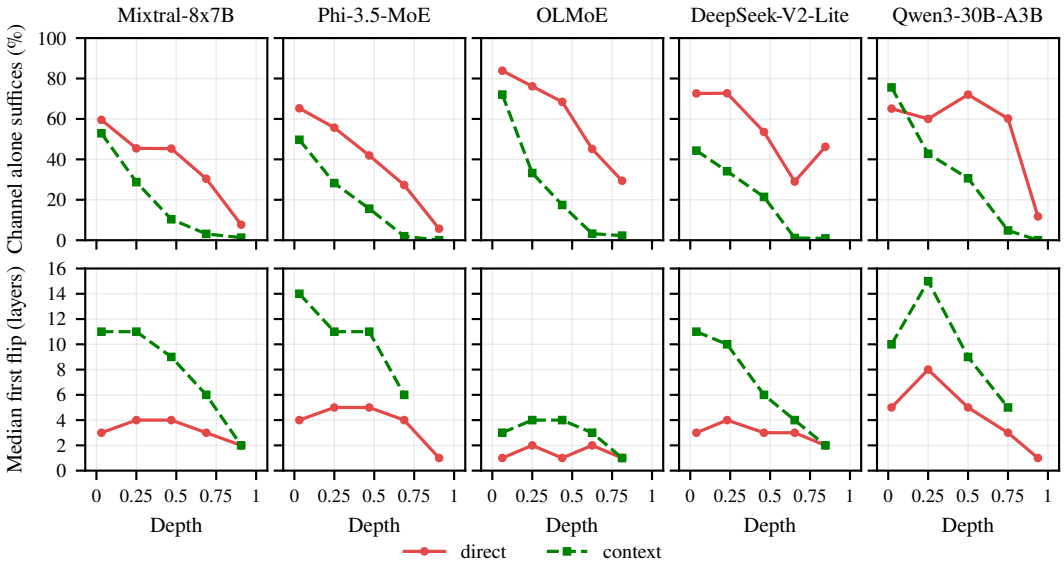


Figure 2: Depth dependence of direct and context-mediated influence. Top: token share diverging under each single-channel intervention. Bottom: median first flip relative to ablated layer.

to at most 0.4% at the deepest. The direct channel shows no such decline; where it eventually drops, only two layers remain, and all categories shrink together.

5.3 Context scaling

All results so far are obtained for a fixed context length of 128 tokens. Scaling the number of perturbed context tokens shows no systematic variation in the balance between the channels, the delay between them, or the accuracy of the decomposition (Appendix I).

6 Margin-normalized routing leverage

We now formalize when a routing decision changes. Recall that the routing logit of expert n at layer ℓ for token i is given by $z_i^{\ell,n} = \boldsymbol{g}^{\ell,n} \cdot \boldsymbol{m}_{\text{in},i}^{\ell}$. Let a be a selected expert and b an unselected expert for that token and layer. By definition of the Top-K selection, it must hold that:

$$z_i^{\ell,a} - z_i^{\ell,b} = (\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}_{\text{in},i}^{\ell} > 0, \tag{4}$$

where $\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}$ is the contrast direction between experts and $m_i^{\ell}(a,b) = z_i^{\ell,a} - z_i^{\ell,b}$ is the routing margin. If a perturbation causes $m_i^{\ell}(a,b) < 0$, b overtakes a and the Top-K changes, and the Top-K is unchanged as long as every such margin remains positive. The routing boundary $m_i^{\ell}(a,b) = 0$, where they are tied, is therefore a hyperplane in router-input space, with one side routing to a and the other to b. If we ablate an upstream expert u, its missing write reaches the boundary through the direct and context-mediated channels, perturbing the router input and with it the margin, $\tilde{m}_i^{\ell}(a,b) = (\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \tilde{\boldsymbol{m}}_{\text{in},i}^{\ell}$. Whether b has overtaken a is not known in advance, but crossing has an equivalent condition:

$$(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \tilde{\boldsymbol{m}}_{\text{in},i}^{\ell} < 0 \iff \frac{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot (\boldsymbol{m}_{\text{in},i}^{\ell} - \tilde{\boldsymbol{m}}_{\text{in},i}^{\ell})}{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}_{\text{in},i}^{\ell}} > 1. \tag{5}$$

We see that crossing a downstream routing boundary depends on two quantities. The numerator is the projection of the missing router input onto the contrast direction between the competing experts, so impact depends on alignment with that direction, not on magnitude alone. The denominator is the routing margin between those experts: the smaller it is, the more susceptible the decision is to perturbation. Routing influence therefore emerges from the interplay of magnitude, alignment and routing-margin vulnerability.

This motivates the following definition: for token i and router at layer ℓ, with upstream expert ablation u, we define

$$J^{\ell}_{u,i} = \max_{a \in \mathrm{TopK}^{\ell}_i,\, b \notin \mathrm{TopK}^{\ell}_i} \frac{[(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot (\boldsymbol{m}^{\ell}_{\mathrm{in},i} - \tilde{\boldsymbol{m}}^{\ell}_{\mathrm{in},i})]_+}{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}^{\ell}_{\mathrm{in},i}}, \tag{6}$$

which we refer to as the *margin-normalized routing leverage*. It evaluates the most susceptible boundary, and the token's Top-K changes if and only if $J^{\ell}_{u,i} > 1$.

Because the numerator is linear in the router-input perturbation, the criterion extends to a coalition of experts by summing their perturbations before projection, and to perturbations entering at different token positions by grouping terms within the same sum. Section 5.1 measures how closely the true displacements of the two channels add; Appendix J gives the coalition form.

We have thus far deferred estimating the perturbation $\tilde{\boldsymbol{m}}^{\ell}_{\mathrm{in},i}$. We build it up in three steps that progressively widen the causal pathways taken into account. Comparing the three shows how much each additional pathway contributes to predicting a flip, and how far in depth that prediction extends.

6.1 The three estimators

Assume that ablating an upstream expert u removes its weighted output, $w^u_i \boldsymbol{e}^u_{\mathrm{out},i}$, from token i's residual stream. The router at layer ℓ reads that stream through the normalizer LN^{ℓ}_2. We approximate how its output changes when the write goes missing by a first-order expansion at the clean input, writing $\widehat{\mathrm{LN}}^{\ell}_{2,i}$ for the normalizer's Jacobian there (Appendix J.1). This gives $J^{(0)}$, which accounts only for the removed write and formally assumes that (i) the removed write reaches layer ℓ unchanged along the residual stream, (ii) all attention blocks up to and including layer ℓ output their clean forward-pass values, and (iii) all intermediate selected experts produce their clean outputs.

$J^{(1)}$ relaxes assumption (ii) at layer ℓ, letting the attention block there read the perturbed context. Ablating u removes $w^u_j \boldsymbol{e}^u_{\mathrm{out},j}$ from the residual stream of every token $j \le i$ that selected it, and linearizing the attention block around its clean forward pass expresses the resulting change in the attention output, $\delta \boldsymbol{a}^{\ell}_{\mathrm{out},i}$, as the sum of a value route and an attention weight route (Appendix J.4). The MoE block at layer ℓ is then short by $w^u_i \boldsymbol{e}^u_{\mathrm{out},i} + \delta \boldsymbol{a}^{\ell}_{\mathrm{out},i}$. The estimators are organized by which computations respond to the ablation, whereas the channels of Section 4 are defined by the positions at which the write is removed: seeding an estimator with the write at position i alone gives the direct channel, and at positions $j < i$ alone the context-mediated one.

$J^{(\infty)}$ relaxes (ii) and (iii) at every intervening layer. $J^{(1)}$ still holds them fixed there, although for $\ell > \ell^u + 1$ perturbed streams are re-read by subsequent attention blocks and re-processed by intervening MoE layers. We formalize it as a *clean-path propagated leverage*: stream deltas accumulate recursively across layers under the condition that the sequence's routing trajectory remains on the clean forward path, each layer advancing all tokens jointly through attention, normalizer, expert response and residual update (Appendix J.5).

All three estimates therefore take the form $\boldsymbol{m}^{\ell}_{\mathrm{in},i} - \tilde{\boldsymbol{m}}^{\ell}_{\mathrm{in},i} \approx \widehat{\mathrm{LN}}^{\ell}_{2,i}(\boldsymbol{s})$ and differ only in the terms $\boldsymbol{s}$ collects:

$$\boldsymbol{s} = \begin{cases} w^u_i \boldsymbol{e}^u_{\mathrm{out},i}, & \text{for } J^{(0)}, \\ w^u_i \boldsymbol{e}^u_{\mathrm{out},i} + \delta \boldsymbol{a}^{\ell}_{\mathrm{out},i}, & \text{for } J^{(1)}, \\ w^u_i \boldsymbol{e}^u_{\mathrm{out},i} + \sum_{c=\ell^u+1}^{\ell} \delta \boldsymbol{a}^{c}_{\mathrm{out},i} + \sum_{c=\ell^u+1}^{\ell-1} \delta \boldsymbol{m}^{c}_{\mathrm{out},i}, & \text{for } J^{(\infty)}, \end{cases} \tag{7}$$

where the first term is the write removed at ℓ^u, the second is attention-mediated and the third expert-mediated. Substituting into Equation 6 gives the three estimators (Appendix J.2).

$J^{(\infty)}$ thus provides a complete first-order account of depth-wise perturbation propagation, subject to linearizing normalizers, attention, activation functions and softmaxes at clean-pass values, and to the Top-K assignments of every token $j \le i$ remaining identical to the clean pass. That last condition is not merely a modeling constraint, since it separates two regimes of behavior. As long as no token $j \le i$ changes its Top-K routing, the model operates in the *clean-path regime*, where $J^{(\infty)}$ tracks the continuous accumulation of perturbation across layers to first order. The first routing flip among context tokens marks the onset of a discrete *nonlinear routing cascade*, beyond which the ablated pass executes a different set of experts.

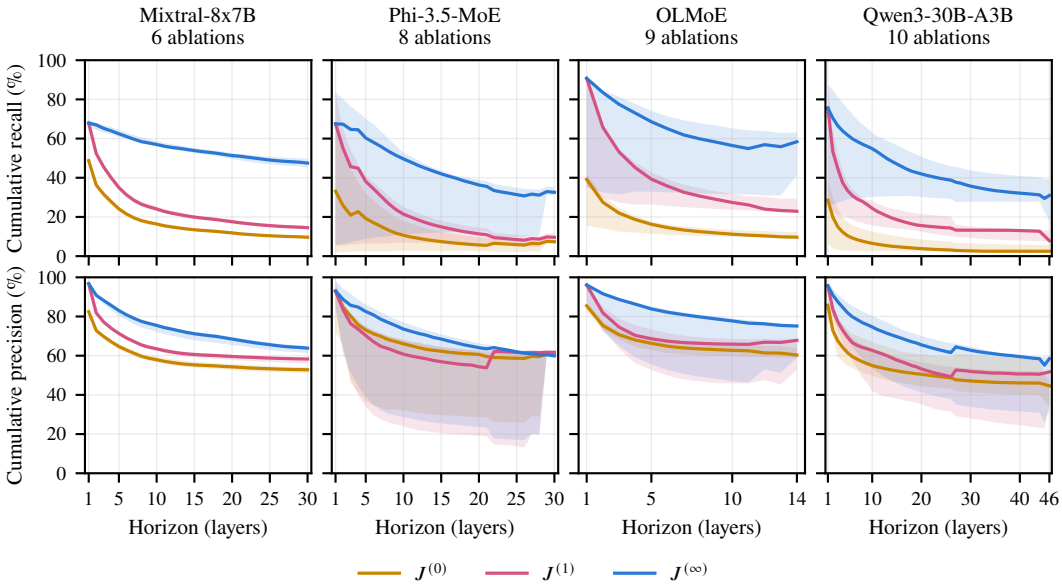


Figure 3: Precision and recall of routing leverage, accumulated over layers past the ablation. Lines show medians across ablations; shaded bands show interquartile ranges.

6.2 Predicting the first routing divergence

We evaluate each estimator against the true ablated trajectory by predicting the layer of a token's first routing divergence. A prediction is correct only if the first layer at which the estimator exceeds one coincides exactly with the layer of the true first divergence. Precision is the fraction of predictions that are correct, and recall the fraction of actual divergences predicted correctly. Results are reported in Figure 3, with per-hop tables (hops count layers past the ablation) in Appendix K.[3]

Recall rises from $J^{(0)}$ to $J^{(1)}$ to $J^{(\infty)}$ in all 33 ablations without exception, so every widening of the causal pathway buys predictive reach. Across each model's ablations, median recall over all hops rises from 2.5–9.7% for $J^{(0)}$ to 30.2–51.3% for $J^{(\infty)}$. $J^{(0)}$ is usually correct when it fires, with first-hop precision of roughly 70% or more throughout, but its recall never exceeds 17.7%, since it cannot fire at tokens not routed to u, which diverge only through attention (Section 4.1).

Both metrics are weakened by a small number of ablations, which also widen the interquartile bands in Figure 3. They are mostly members of the starred coalitions of Section 4, which largely coincide with the experts of largest activation magnitude, and what they share is that they remove most of the residual stream at $j^\star$, the position later tokens attend to most, which is typically the sink token. When nearly the whole stream is removed, the perturbation is no longer local, so the first-order expansion around the clean input fails (Appendix J.6), and attention carries that error to every token that reads $j^\star$. The failure therefore arises where $\|\boldsymbol{x}_{j^\star} - w^u_{j^\star} \boldsymbol{e}^u_{\mathrm{out},j^\star}\| \ll \|\boldsymbol{x}_{j^\star}\|$, which holds for six of the 33 ablations. In these, the removed write is nearly parallel to the stream at $j^\star$ (cosine similarity at least 0.999) and amounts to 86–100% of its norm, leaving 1–14% of it.

To test whether the write at $j^\star$ causes these failures, we ablate the expert at every position except $j^\star$ and recompute both the predictions and the true divergences. Precision rises on all nine ablations that include one of these experts, by 13.6 to 53.2 points, and recall on eight of them by 13.9 to 34.3 points, while the precision of the other 24 changes by at most 4.9 points (Appendix K).

6.3 Routing leverage against its components

We test the contribution of each component of $J^{(\infty)}$ by scoring tokens with each component independently. At every hop past the ablation, we retain tokens that remain on their clean route and ask which diverge at that hop. For token i at that hop, with a the weakest selected and b the strongest

[3] DeepSeek-V2-Lite is omitted because our implementation of $J^{(1)}$ and $J^{(\infty)}$ does not cover its latent attention, which has no separate key and value projections.

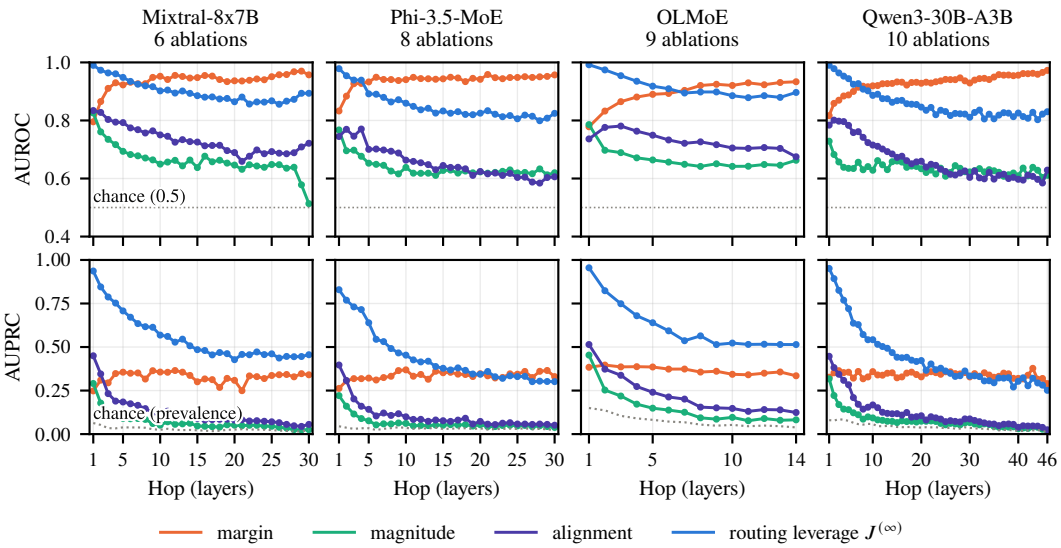


Figure 4: Discriminative power of the routing leverage and its components, by hop. Lines show the median across a model's ablations. Hops with fewer than 20 token flips are omitted.

unselected expert, we score margin as $-m_i^\ell(a,b)$, magnitude as $|\delta \boldsymbol{m}_{\text{in},i}^\ell|$, with $\delta \boldsymbol{m}_{\text{in},i}^\ell$ the perturbation predicted by $J^{(\infty)}$, alignment as $(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \delta \boldsymbol{m}_{\text{in},i}^\ell$, and leverage as $J_{u,i}^{(\infty),\ell}$. All four scores are computed from the same clean forward pass using the same predicted perturbation, isolating the scoring rule. We evaluate each score using AUROC and AUPRC, with chance levels of 0.5 and the fraction of tokens that diverge at that hop, respectively.

Figure 4 shows that alignment scores above magnitude in every model beyond the first hop: what matters is the perturbation's component along the contrast direction, not its overall size. $J^{(\infty)}$ in turn scores above alignment at every hop, both because it normalizes by the margin and because it evaluates the most susceptible boundary (Appendix L). At the hop immediately after the ablation, $J^{(\infty)}$ separates diverging from surviving tokens almost perfectly, with AUROC between 97.8% and 99.2%, while no single component exceeds 83.4%. AUPRC gives the same ordering: $J^{(\infty)}$ scores highest in 28 of 33 ablations, and each of the other five contains an expert whose write at $j^\star$ breaks the expansion (Section 6.2).

The margin alone is a strong ranker at depth, reaching 91.0% to 94.2% from the fifth hop on and overtaking $J^{(\infty)}$ there. The decline of $J^{(\infty)}$ there coincides with the end of the clean-path regime of Section 6.1. Restricted to tokens whose context has not yet diverged, $J^{(\infty)}$ holds at 93.7% to 100% over the same hops, above the margin in every model but Phi (Appendix L).

7 Conclusion

Expert computation in MoE models is not confined to the token on which it executes. An expert's computation can influence routing through two causal channels: the direct same-token channel through its own residual stream, and the context-mediated inter-token channel through attention, which alters the routing of tokens that never selected the expert. Before the first Top-K change, the two channels combine approximately linearly in routing-margin space. Routing leverage predicts the layer of the first routing divergence from clean-pass quantities alone, with longer causal pathways extending the range over which the prediction remains accurate. The first routing divergence thus marks the boundary between approximately linear perturbation propagation, captured by the clean-pass linearization, and the onset of a nonlinear routing cascade.

Our results suggest that the effects of quantization and pruning in MoE models depend not only on local statistics but also on routing sensitivity. Routing leverage identifies computations whose perturbation is most likely to induce downstream routing divergence, while routing-margin vulnerability may provide a useful target for router training or regularization.

REPRODUCIBILITY STATEMENT

Every result in this paper can be reproduced from the code linked in the abstract, which contains the customized model implementations the interventions require, the measurement scripts that produce the records, and the analysis scripts that turn them into each figure and table. Section 4 states the models, prompts and numerical precision used throughout, and Appendix B gives the architectural specifications of each model. Appendices C and D describe how the ablated experts were selected, and Appendix E reports their routing loads. The derivations behind the three estimators are given in Appendix J, and the per-ablation numbers behind Figures 3 and 4 in Appendices K and L.

AI USE STATEMENT

In this work, we used generative AI tools for code generation and as an editorial aid for writing and grammar. We have not used generative AI tools for research ideation: the research questions, the experimental design, the analyses and all claims are the authors' own. We have reviewed all AI-assisted work; generated code was reviewed by the authors and its outputs verified against the results reported here, and all AI-assisted text was checked and revised by the authors. We take full responsibility for the final content of this work.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

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

Krishna Teja Chitty-Venkata, Jie Ye, and Murali Emani. MoPEQ: Mixture of mixed precision quantized experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025.

CodeParrot. CodeParrot: A dataset of Python files from GitHub, 2022. URL `https://huggingface.co/datasets/codeparrot/codeparrot-clean`.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. September 2026. URL `https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf`.

Zijin Gu, Tatiana Likhomanenko, Vimal Thilak, Jason Ramapuram, and Navdeep Jaitly. Path-constrained mixture-of-experts. *arXiv preprint arXiv:2603.18297*, 2026.

Jeremy Herbst, Stefan Wermter, and Jae Hee Lee. The expert strikes back: Interpreting mixture-of-experts language models at expert level. In *Forty-third International Conference on Machine Learning*, 2026.

Hugging Face. Open R1: A fully open reproduction of DeepSeek-R1, 2025. URL `https://github.com/huggingface/open-r1`.

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, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

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

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, et al. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Mistral AI. Cheaper, better, faster, stronger. https://mistral.ai/news/mixtral-8x22b, 2024.

Mistral AI. Introducing Mistral Small 4. Mistral AI, March 2026. URL https://mistral.ai/news/mistral-small-4/.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. OLMoE: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alexandre Muzio, Alex Sun, and Churan He. SEER-MoE: Sparse expert efficiency through regularization for mixture-of-experts. *arXiv preprint arXiv:2404.05089*, 2024.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. Probing semantic routing in large mixture-of-expert models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18263–18278, 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. Qwen Blog, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

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.

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 *The Fourteenth International Conference on Learning Representations*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

APPENDIX

A ATTENTION AND MOE BLOCKS

We state the two block computations of Section 3 in full, in the notation used throughout.

Attention block. For simplicity, we assume multi-head self-attention (MHSA) with H heads and per-head dimension $d_h = d/H$. For head h, queries, keys, and values are $\boldsymbol{q}_i^{\ell,h} = \boldsymbol{W}_Q^{\ell,h}\boldsymbol{a}_{\text{in},i}^{\ell}$, $\boldsymbol{k}_j^{\ell,h} = \boldsymbol{W}_K^{\ell,h}\boldsymbol{a}_{\text{in},j}^{\ell}$, and $\boldsymbol{v}_j^{\ell,h} = \boldsymbol{W}_V^{\ell,h}\boldsymbol{a}_{\text{in},j}^{\ell}$, with projection matrices $\boldsymbol{W}_Q^{\ell,h}, \boldsymbol{W}_K^{\ell,h}, \boldsymbol{W}_V^{\ell,h} \in \mathbb{R}^{d_h \times d}$. Under causal attention, the attention weights and the block output are

$$p_{i,j}^{\ell,h} = \underset{t \le i}{\text{softmax}}\left(\frac{\boldsymbol{q}_i^{\ell,h} \cdot \boldsymbol{k}_t^{\ell,h}}{\sqrt{d_h}}\right)_j, \qquad \boldsymbol{a}_{\text{out},i}^{\ell} = \sum_{h=1}^{H}\sum_{j \le i} \boldsymbol{W}_O^{\ell,h} p_{i,j}^{\ell,h} \boldsymbol{v}_j^{\ell,h}, \tag{8}$$

where $\boldsymbol{W}_O^{\ell,h} \in \mathbb{R}^{d \times d_h}$.

Expert computation. Each expert n computes

$$\boldsymbol{e}_{\text{out},i}^{\ell,n} = \boldsymbol{W}_{\text{down}}^{\ell,n}\left[\sigma(\boldsymbol{W}_{\text{gate}}^{\ell,n}\boldsymbol{m}_{\text{in},i}^{\ell}) \odot (\boldsymbol{W}_{\text{up}}^{\ell,n}\boldsymbol{m}_{\text{in},i}^{\ell})\right], \tag{9}$$

where $\boldsymbol{W}_{\text{gate}}^{\ell,n}, \boldsymbol{W}_{\text{up}}^{\ell,n} \in \mathbb{R}^{d_{\text{ff}} \times d}$, $\boldsymbol{W}_{\text{down}}^{\ell,n} \in \mathbb{R}^{d \times d_{\text{ff}}}$, and $\sigma(\cdot)$ is element-wise. The MoE block output aggregates the selected expert contributions as

$$\boldsymbol{m}_{\text{out},i}^{\ell} = \sum_{n \in \text{TopK}_i^{\ell}} w_i^{\ell,n} \boldsymbol{e}_{\text{out},i}^{\ell,n}. \tag{10}$$

B ARCHITECTURAL SPECIFICATIONS OF ANALYZED MODELS

The routing experiments of Sections 4–6 run in `float32`, which doubles memory requirements and excludes the three largest models; they are therefore conducted on the five smaller ones. The expert-selection analysis of Appendix C measures continuous quantities rather than a discrete boundary, and runs in `bfloat16`, which accommodates all eight models, among them Mixtral-8x22B (Jiang et al., 2024; Mistral AI, 2024), Qwen3-235B-A22B (Yang et al., 2025) and DeepSeek-V2 (Liu et al., 2024).

Table 2: MoE models analyzed.

Model	Total params	Activated params	Layers	Number of routed experts	TopK
Mixtral-8x7B	47B	13B	32	8	2
Mixtral-8x22B	141B	39B	56	8	2
Phi-3.5-MoE	42B	6.6B	32	16	2
OLMoE	7B	1B	16	64	8
Qwen3-30B-A3B	30B	3B	48	128	8
Qwen3-235B-A22B	235B	22B	94	128	8
DeepSeek-V2-Lite	16B	2.4B	27	64	6
DeepSeek-V2	236B	21B	60	160	6

C EXPERT-TO-EXPERT INFLUENCE GRAPHS

The ablations in Section 4 are selected with a first-order expert-to-expert routing influence graph, built from the routing score decomposition of Equation 3.

Graph definition. For an upstream MoE block $c < \ell$, the linearity of $\overline{\mathrm{LN}}^{\ell}_{2,i}$ further gives

$$\boldsymbol{g}^{\ell,n} \cdot \overline{\mathrm{LN}}^{\ell}_{2,i}(\boldsymbol{m}^{c}_{\mathrm{out},i}) = \sum_{m \in \mathrm{TopK}^{c}_{i}} \boldsymbol{g}^{\ell,n} \cdot \overline{\mathrm{LN}}^{\ell}_{2,i}(w^{c,m}_{i} \boldsymbol{e}^{c,m}_{\mathrm{out},i}), \tag{11}$$

so that the contribution of upstream expert (c, m) to the routing logit of downstream expert (ℓ, n) is

$$\boldsymbol{g}^{\ell,n} \cdot \overline{\mathrm{LN}}^{\ell}_{2,i}(w^{c,m}_{i} \boldsymbol{e}^{c,m}_{\mathrm{out},i}), \qquad c < \ell, \tag{12}$$

which captures how one expert's contribution changes the routing score of an expert at a downstream layer.

We represent this influence as a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ whose vertices are experts and whose edges run forward across layers:

$$\begin{aligned} &\mathcal{V} = [L] \times [N], \\ &\mathcal{E} = \{(c, m) \to (\ell, n) \mid c < \ell\}. \end{aligned} \tag{13}$$

The constraint $c < \ell$ makes the graph acyclic, since an expert can only influence routers at later layers.

For an upstream expert (c, m) and a downstream expert (ℓ, n), we remove the contribution from Equation 12 from the downstream expert's routing logit while leaving all other logits unchanged:

$$\tilde{z}^{\ell,n'}_{i,(c,m)} = \begin{cases} z^{\ell,n}_{i} - \boldsymbol{g}^{\ell,n} \cdot \overline{\mathrm{LN}}^{\ell}_{2,i}\left(w^{c,m}_{i} \boldsymbol{e}^{c,m}_{\mathrm{out},i}\right), & n' = n, \\ z^{\ell,n'}_{i}, & n' \neq n. \end{cases} \tag{14}$$

We recompute TopK^{ℓ}_{i} from $\tilde{\boldsymbol{z}}^{\ell}_{i,(c,m)}$, and an expert outside the Top-K has routing weight zero, in $w^{\ell,n}_{i}$ and $\tilde{w}^{\ell,n}_{i,(c,m)}$ alike. The edge weight is then the expected absolute change in the downstream routing weight, over tokens for which the source expert was selected:

$$\mathcal{W}(c, m \to \ell, n) = \mathbb{E}_{i}\left[\left|w^{\ell,n}_{i} - \tilde{w}^{\ell,n}_{i,(c,m)}\right| \mid m \in \mathrm{TopK}^{c}_{i}\right]. \tag{15}$$

For example, suppose expert (ℓ, n) contributes to token i's hidden state with routing weight 0.4 on the clean pass. If removing the contribution of an upstream expert (c, m) reduces this weight to 0.05, the edge $(c, m) \to (\ell, n)$ carries a value of 0.35. This indicates that the upstream expert has a strong influence on the downstream routing decision, even though (ℓ, n) remains in the Top-K.

This estimate is obtained by holding the intervening computation fixed. It therefore identifies where an expert's contribution can reach a downstream router, rather than predicting the outcome of a full ablation once subsequent Top-K decisions are allowed to change. It also excludes the context-mediated channel. The resulting graph can thus be viewed as a first-order routing skeleton, which the model subsequently amplifies through the nonlinear cascades induced by routing flips. Because each edge removes the contribution from a single downstream logit, it measures how sensitive that expert's routing weight is to the contribution. We use the graph only to select the experts we ablate.

Graph-based influence ranking. Ranking experts by the total outgoing influence of their vertex,

$$\inf(c, m) = \sum_{\ell > c} \sum_{n \in [N]} \mathcal{W}(c, m \to \ell, n), \tag{16}$$

gives the selection used throughout. Table 3 reports the NLL increase when the highest-ranked set of each model is ablated, both jointly and member by member. Where that set holds more than one expert, the joint effect exceeds the sum of the individual ones by between $3.9\times$ and $33.9\times$, which is why we mark members of these coalitions with *. OLMoE is the only model in which ablating even its highest-ranked experts leaves NLL essentially unchanged. Nevertheless, its routing is affected as strongly as in the other models. Ablations of OLMoE experts change the downstream routing of

Table 3: Superlinear expert ablation effect on C4. $\Delta\mathrm{NLL}$ is relative to the clean run. *Indiv.* sums the $\Delta\mathrm{NLL}$ from individual ablations, while *Joint / Indiv.* is their ratio.

Model	**#Ablated experts**	**Joint ΔNLL**	**Indiv. ΔNLL**	**Joint / Indiv.**
Mixtral-8x7B	1	$\mathbf{0.30}$	—	—
Mixtral-8x22B	2	$\mathbf{1.43}$	0.20	$\mathbf{7.07}\times$
Phi-3.5-MoE	3	$\mathbf{1.33}$	0.34	$\mathbf{3.91}\times$
Qwen3-30B-A3B	4	$\mathbf{1.76}$	0.05	$\mathbf{33.87}\times$
Qwen3-235B-A22B	2	$\mathbf{1.93}$	0.19	$\mathbf{10.38}\times$
DeepSeek-V2-Lite	2	$\mathbf{0.25}$	0.04	$\mathbf{5.77}\times$
DeepSeek-V2	3	$\mathbf{0.09}$	0.02	$\mathbf{5.70}\times$
OLMoE	3	0.01	0	—

a median 60.0% of the tokens that never selected the ablated experts (Table 1). Routing influence and model-wide importance are therefore separable. Whereas the super-expert criterion of Su et al. (2026) ranks experts by activation magnitude, $\inf(\cdot)$ ranks them by how much they shape routing decisions at downstream layers (Appendix D). Our analysis therefore characterizes a mechanism that the super-expert criterion does not capture.

Figures 5–7 show influence graphs of Mixtral-8x7B on three tasks, each computed from 500 prompts. Almost all of the edges are negligible, so we draw the subset $\mathcal{E}_Q$ whose magnitude exceeds the Q-quantile of the edge-weight distribution $\{|\mathcal{W}(e)| : e \in \mathcal{E}\}$, with $Q = 99.9\%$ throughout, together with the strongest outgoing edges of the most influential experts.

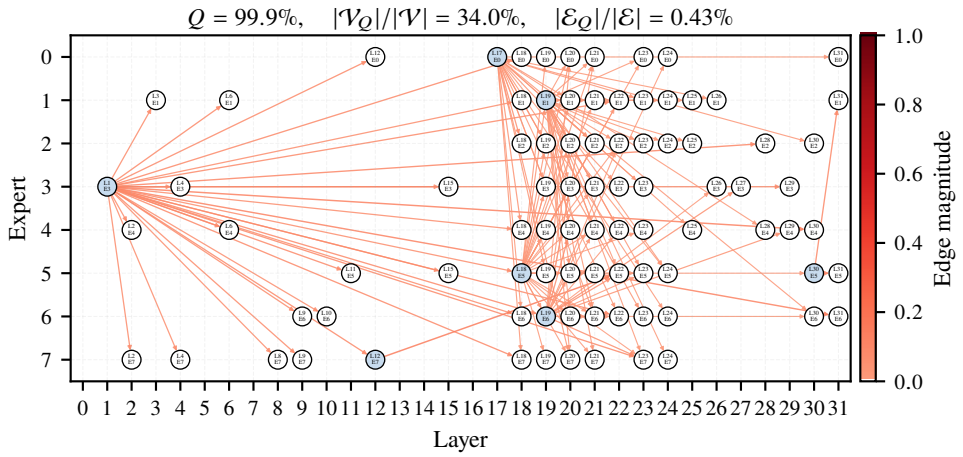


Figure 5: Expert-to-expert influence graph for Mixtral-8x7B on C4.

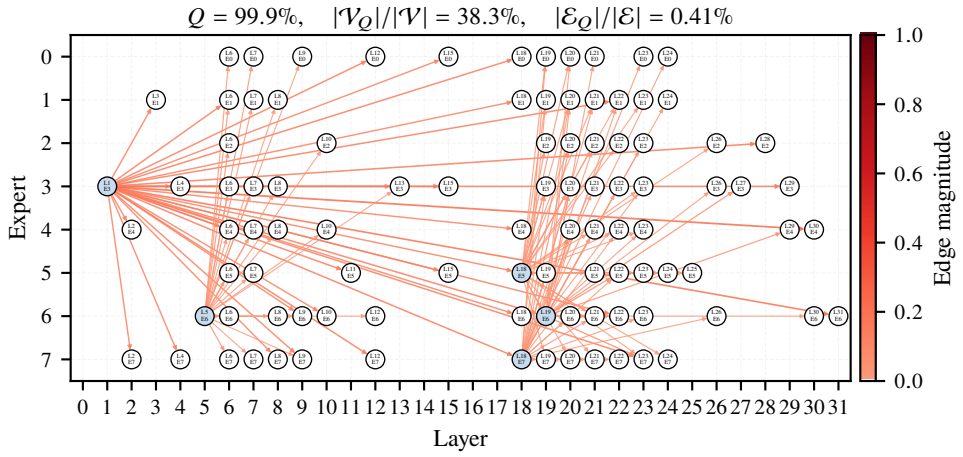


Figure 6: Expert-to-expert influence graph for Mixtral-8x7B on OpenR1-Math (Hugging Face, 2025).

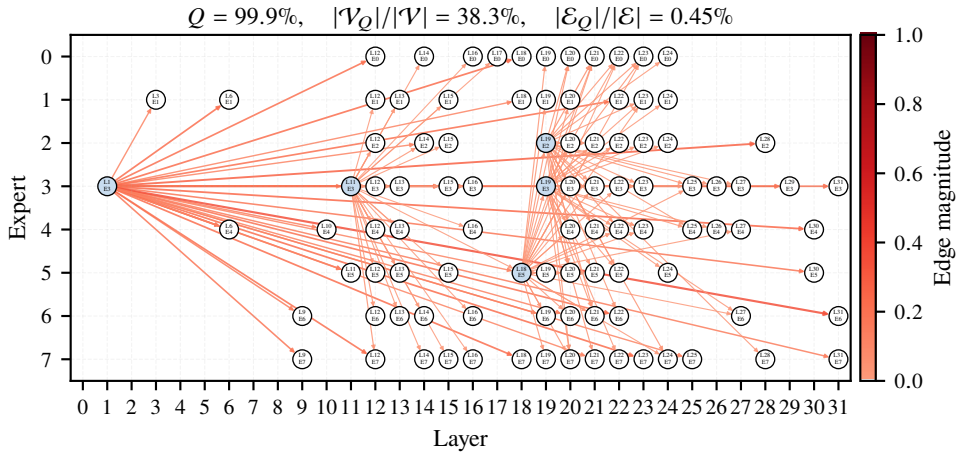


Figure 7: Expert-to-expert influence graph for Mixtral-8x7B on CodeParrot (CodeParrot, 2022).

D INFLUENTIAL VERSUS SUPER EXPERTS

Su et al. (2026) define an expert as *super* when its activation magnitude a, the largest ℓ_∞ norm its down-projection output attains over the tokens routed to it, satisfies $a > P_{99.5}$ and $a > \frac{1}{10} a_{\max}$, and, following the default of their released code, when the expert sits in the first three quarters of the model's depth, $\ell < 0.75L$. We rank experts instead by their total outgoing influence (Equation 16).

The depth condition corrects for *outlier experts*: experts that pass both magnitude criteria but sit in the final layers, where too few layers remain for their write to alter the computation that follows. In $\inf(\cdot)$ this depth prior is intrinsic, since the sum runs over the routers an expert can still reach; the experts it ranks highest are those whose edges extend far down the network and arrive at many downstream experts at once (e.g. Mixtral-8x7B L1E3 in Figures 5–7).

Table 4 ranks each model's experts by activation magnitude and by $\inf(\cdot)$, ablates the highest-ranked under each, and reports the resulting NLL; the *Ranking* column gives the rank a set holds under either criterion. Experts marked * belong to a superlinear coalition, and underlined experts rank highly by $\inf(\cdot)$ without being super experts. Rows labelled *random* report the median over five sets of randomly chosen experts, of the same size as that model's coalition and drawn uniformly over its MoE layers.

The two rankings largely agree at the top, in a different order, and $\inf(\cdot)$ additionally selects false negatives of the magnitude criterion, underlined in the table, whose activations fall below its threshold. Ablated alone these experts leave NLL unchanged, but they are not inert: adding Phi-3.5-MoE L10E0 to that model's two super experts raises the joint increase from 0.80 to 1.33 nats, although its own ablation raises it by 0.00. The reason is that $\inf(\cdot)$ measures how much an expert's write moves downstream routing weights, so all three quantities of Section 6.3 enter it at once: magnitude, alignment with the downstream router direction, and margin vulnerability.

Table 4: Expert ablation effect on C4. *joint* is the NLL increase with all listed experts ablated together, *indiv.* the sum of the increases with each ablated alone, both in nats, and *interact.* their ratio, where $1\times$ is exactly additive.

Model	Ranking	Ablated experts	PPL	NLL	NLL increase		
					joint	indiv.	interact.
Mixtral-8x7B	baseline	—	6.88	1.93			
	act 1 / inf 1	L1E3*	9.29	2.23	**0.30**	—	—
	random	1 expert	6.89	1.93	0.00	—	—
Mixtral-8x22B	baseline	—	6.19	1.82			
	act 1 / inf 2	L1E7*	7.31	1.99	0.17	—	—
	act 2 / inf 1	L0E2*	6.42	1.86	0.04	—	—
	act / inf top-2	L1E7*+L0E2*	25.79	3.25	**1.43**	0.20	**7.07**×
	random	2 experts	6.20	1.82	0.00	—	—
Phi-3.5-MoE	baseline	—	7.67	2.04			
	act 1 / inf 2	L5E10*	9.40	2.24	0.20	—	—
	act 2 / inf 1	L3E7*	8.79	2.17	0.14	—	—
	act 3 / inf 3	L10E0*	7.66	2.04	0.00	—	—
	act / inf top-2	L3E7*+L5E10*	17.01	2.83	**0.80**	0.34	**2.34**×
	act / inf top-3	L3E7*+L5E10*+L10E0*	28.95	3.37	**1.33**	0.34	**3.91**×
	random	3 experts	7.68	2.04	0.00	—	—
Qwen3-30B-A3B	baseline	—	12.15	2.50			
	act 1 / inf 1	L2E92*	12.20	2.50	0.00	—	—
	act 2	L1E68*	12.47	2.52	0.03	—	—
	act 3 / inf 2	L3E82*	12.41	2.52	0.02	—	—
	inf 3	L21E69*	12.15	2.50	0.00	—	—
	act top-3	L2E92*+L1E68*+L3E82*	68.65	4.23	**1.73**	0.05	**33.47**×
	act top-3 + inf 3	L2E92*+L1E68*+L3E82*+L21E69*	70.27	4.25	**1.76**	0.05	**33.87**×
	random	4 experts	12.16	2.50	0.00	—	—
Qwen3-235B-A22B	baseline	—	9.30	2.23			
	act 1 / inf 1	L3E120*	9.40	2.24	0.01	—	—
	act 2 / inf 2	L2E39*	11.09	2.41	0.18	—	—
	act / inf top-2	L3E120*+L2E39*	64.10	4.16	**1.93**	0.19	**10.38**×
	random	2 experts	9.30	2.23	0.00	—	—
DeepSeek-V2-Lite	baseline	—	8.91	2.19			
	act 1	L3E54*	9.27	2.23	0.04	—	—
	act 2	L4E38*	8.94	2.19	0.00	—	—
	act top-2 / inf top-4	L3E54*+L4E38*	11.43	2.44	**0.25**	0.04	**5.77**×
	random	2 experts	8.92	2.19	0.00	—	—
DeepSeek-V2	baseline	—	6.85	1.92			
	act 1 / inf 1	L18E96*	6.86	1.93	0.00	—	—
	act 2 / inf 2	L21E94*	6.96	1.94	0.02	—	—
	act 3 / inf 4	L20E48*	6.86	1.93	0.00	—	—
	act top-3 / inf top-4	L18E96*+L21E94*+L20E48*	7.53	2.02	**0.09**	0.02	**5.70**×
	inf 3	L12E62	6.85	1.92	0.00	—	—
	inf top-3	L18E96*+L21E94*+L12E62	7.15	1.97	**0.04**	0.02	**2.66**×
	random	3 experts	6.85	1.92	0.00	—	—
OLMoE	baseline	—	10.47	2.35			
	inf top-3	L2E30+L1E9+L4E14	10.53	2.35	0.01	—	—
	act top-4	L1E9+L4E27+L1E18+L2E30	10.59	2.36	0.01	—	—
	random	3 experts	10.50	2.35	0.00	—	—

E LOAD-MATCHED EXPERTS

Table 5: Routing load of each ablated expert and the range of loads at its layer. *Direct* gives the share of tokens diverging under the direct intervention one layer past the ablation and at the last layer, *Context* the context share one layer past the ablation and at its peak; † marks direct shares that peak later.

Model	Ablation	Load	Layer range	Direct	Context
Mixtral-8x7B	L1E2	24.75%	22.18–30.86%	19.50 → 1.66%	3.32 → 6.02%
	L8E0	24.68%	23.68–26.39%	14.59 → 1.48%	0.85 → 5.50%
	L15E2	25.88%	15.36–30.04%	11.99 → 2.24%	0.61 → 2.03%
	L22E4	24.88%	21.72–26.94%	6.04† → 0.83%	0.42 → 1.04%
	L29E6	24.97%	19.72–29.04%	3.41† → 5.12%	0.21 → 1.49%
Phi-3.5-MoE	L1E1	13.15%	5.55–18.29%	18.74 → 3.79%	1.26 → 5.68%
	L8E6	13.89%	0.71–26.72%	9.42 → 3.00%	0.43 → 3.00%
	L15E13	12.80%	1.53–22.72%	6.91† → 4.09%	0.00 → 2.81%
	L22E12	12.29%	2.16–23.70%	3.08† → 3.52%	0.00 → 1.32%
	L29E14	12.86%	0.79–22.66%	3.95 → 1.75%	0.00 → 0.00%
OLMoE	L1E7	12.03%	6.93–21.32%	53.76 → 7.53%	18.28 → 24.95%
	L4E20	11.89%	3.68–29.37%	35.09 → 8.94%	5.96 → 8.72%
	L7E32	11.93%	2.57–38.45%	35.32 → 14.13%	4.42 → 4.42%
	L10E49	11.95%	2.86–21.21%	18.75 → 11.81%	0.23 → 1.85%
	L13E12	12.03%	2.36–32.08%	16.97 → 16.29%	1.58 → 1.58%
DeepSeek-V2-Lite	L1E37	8.71%	4.13–28.46%	18.49 → 7.57%	2.23 → 7.35%
	L6E63	8.66%	4.96–15.28%	16.71 → 11.06%	1.41 → 5.88%
	L12E35	8.69%	4.96–19.98%	13.39 → 4.17%	1.49 → 5.36%
	L17E26	8.54%	6.00–18.15%	7.03† → 3.52%	0.22 → 0.88%
	L22E3	8.67%	5.42–19.37%	20.54 → 13.54%	0.45 → 0.45%
Qwen3-30B-A3B	L1E15	2.97%	0.04–40.13%	12.40 → 1.94%	3.10 → 13.57%
	L12E49	3.03%	0.00–39.18%	5.52† → 5.52%	3.45 → 8.97%
	L24E9	3.00%	0.00–40.27%	15.05 → 3.76%	1.61 → 6.45%
	L36E7	3.08%	0.00–33.95%	19.72 → 7.27%	0.00 → 1.73%
	L45E126	3.02%	0.00–47.73%	7.06 → 4.71%	0.00 → 0.00%

Section 5.2 varies the layer at which an expert is ablated while holding its routing load fixed, and Appendix I and Section 4.3 reuse each model's layer-1 expert from that set. Load has to be controlled because it sets how many tokens an ablation touches. Holding the expert index fixed would not control it: expert n at an early layer and expert n at a late one are unrelated, and their loads differ substantially. Matching each layer to its own median would not control it either, because the per-layer medians drift with depth. We therefore fix one target load per model, the median of the per-layer medians, and at each layer ablate the expert closest to it.

Table 5 reports the routing load of each ablated expert together with the range of loads at its layer, and the shares of tokens diverging under each single-channel intervention that Section 5.2 summarizes. Load is measured on the clean forward pass as the fraction of positions whose selected set contains that expert, over the same 500 C4 sequences of 128 tokens used throughout.

Within each model the ablated experts' loads span at most 1.6 percentage points across its five layers, while the loads within a single layer span tens of points, so depth is the only quantity changing appreciably.

F TAXONOMY OF ROUTING FLIPS

Table 6 gives the per-ablation breakdown behind Section 4.2.

Direct only denotes tokens that flip under the direct intervention but not the context one, *Context only* the converse, and *Both alone* tokens that flip under both; all three also flip under the joint

Table 6: A taxonomy of routing flips. All entries are percentages of analyzed tokens.

Model	Ablation	Direct only	Context only	Both alone	Both required	Cancelled	No change
Mixtral-8x7B	L1E3*	1.9%	75.3%	22.6%	**0.0%**	0.2%	0.0%
	L15E3	30.0%	2.9%	9.0%	**1.8%**	3.5%	52.8%
	L29E3	6.9%	1.0%	2.4%	**2.0%**	1.0%	86.7%
Phi-3.5-MoE	L3E7*	0.0%	44.1%	**55.7%**	**0.0%**	0.2%	0.0%
	L5E10*	9.9%	25.1%	**57.1%**	**0.8%**	1.2%	5.9%
	L10E0*	**50.1%**	0.7%	20.4%	**0.7%**	2.7%	25.4%
	L1E1	25.1%	9.5%	**40.2%**	**2.9%**	5.7%	16.6%
OLMoE	L1E9	11.1%	15.4%	**67.0%**	**0.6%**	1.9%	4.0%
	L2E30	8.7%	15.8%	**66.7%**	**1.1%**	2.4%	5.3%
	L4E14	28.3%	7.1%	**57.4%**	**1.7%**	2.8%	2.7%
	L13E30	22.5%	0.2%	0.7%	**0.7%**	0.4%	75.5%
DeepSeek-V2-Lite	L3E54*	0.0%	11.4%	**88.6%**	**0.0%**	0.0%	0.0%
	L4E38*	22.7%	4.3%	**65.5%**	**0.4%**	2.1%	5.0%
	L22E54	16.0%	0.0%	0.9%	**0.2%**	0.5%	82.4%
Qwen3-30B-A3B	L1E68*	0.0%	22.0%	**77.8%**	**0.0%**	0.2%	0.0%
	L2E92*	1.0%	44.8%	**54.3%**	**0.0%**	0.0%	0.0%
	L3E82*	3.3%	13.3%	**83.4%**	**0.0%**	0.0%	0.0%
	L21E69*	9.0%	18.5%	**56.9%**	**1.5%**	4.2%	9.9%

intervention. *Both required* flips only under the joint intervention, *Cancelled* under a single channel but not jointly, and *No change* under none.

G ATTENTION PATCHING BY COMPONENT

Table 7 gives the per-ablation breakdown behind Section 4.3.

Table 7: Fraction of context-mediated routing flips restored by attention patching.

Model	Ablation	KV	V	K
Mixtral-8x7B	L1E2	**100%**	44%	28%
Phi-3.5-MoE	L1E1	**100%**	47%	27%
OLMoE	L1E7	**100%**	40%	11%
DeepSeek-V2-Lite	L1E37	**100%**	64%	18%
Qwen3-30B-A3B	L1E15	**100%**	6%	12%

H CHANNEL ADDITIVITY PER ABLATION

Table 8 gives the per-ablation breakdown behind Section 5.1. Displacements are compared at every layer after the ablated one and strictly before the analyzed token's first Top-K change under any of the three interventions. Context tokens may diverge earlier, but causal masking keeps their states identical under the context and joint interventions, so what they contribute to token i is the same in both.

Table 8: Additivity of direct and context-mediated channels up to the first routing divergence.

Model	Ablation	$\rho(\delta_{\mathrm{dir}}+\delta_{\mathrm{ctx}}, \delta_{\mathrm{joint}})$	Median err. / margin	p90 err. / margin
Mixtral-8x7B	L1E3*	0.9971	3.15%	25.37%
	L15E3	0.9981	0.06%	1.24%
	L29E3	0.9990	0.30%	2.64%
Phi-3.5-MoE	L3E7*	0.9771	5.88%	28.08%
	L5E10*	0.9709	1.97%	20.07%
	L10E0*	0.9705	0.01%	3.40%
	L3E7*+L5E10*+L10E0*	0.9872	1.58%	19.84%
	L1E1	0.9402	0.60%	10.85%
OLMoE	L1E9	0.9979	0.47%	6.66%
	L2E30	0.9986	0.49%	5.48%
	L4E14	0.9979	1.18%	11.65%
	L13E30	1.0000	0.02%	0.26%
	L2E30+L1E9+L4E14	0.9975	0.56%	9.85%
DeepSeek-V2-Lite	L3E54*	0.9947	2.10%	19.85%
	L4E38*	0.9982	0.33%	4.83%
	L22E54	1.0000	$< 0.01\%$	0.02%
	L3E54*+L4E38*	0.9953	1.02%	16.12%
Qwen3-30B-A3B	L1E68*	0.9987	1.88%	13.28%
	L2E92*	0.9998	0.37%	1.98%
	L3E82*	0.9994	0.54%	3.50%
	L21E69*	0.9992	0.30%	2.51%
	L2E92*+L1E68*+L3E82*+L21E69*	0.9993	0.72%	10.77%

I CONTEXT SCALING

We ask whether the channel structure of Sections 5.1 and 5.2 depends on how much perturbed context a token has.

Because attention is causal, a token's routing decision depends strictly on its prefix. The critical variable for context scaling is therefore not the absolute sequence length, but n_{ctx}: the number of preceding tokens routed to the ablated expert through which the perturbation enters. We group analyzed tokens by n_{ctx}, extracting all groups from one run of 2048-token sequences per model. We ablate each model's layer-1 median-load expert, selected as in Section 5.2.

We expect routing behavior to remain stable across n_{ctx} due to how attention normalizes influence. Specifically, an expert's routing load remains approximately uniform across a sequence. For instance, if an expert processes 10% of tokens, a token at position 100 has 10 perturbed sources in its prefix, whereas a token at position 2000 has 200. However, because attention weights sum to one, the attention mechanism computes a weighted average over the prefix rather than an unnormalized sum. Consequently, the context channel's influence depends on the share of attention that falls on perturbed positions rather than on their absolute number (which grows linearly with context length); if attention were spread evenly, that share would equal the perturbed fraction of the prefix, which remains stable at 10%. The same tokens grouped by their position i in the sequence confirm that it is this share that stays fixed: the median n_{ctx} rises 19–36$\times$ from the earliest group to the latest, while the perturbed share n_{ctx}/i barely moves.

Table 9 shows no systematic variation with n_{ctx} in the channel balance, the delay between the channels, or the accuracy of the decomposition. The additive linear decomposition remains as accurate at several hundred context tokens as it is at a handful, with the Pearson correlation staying between 0.925 and 0.9999 and the error between 0.11% and 1.09% of the margin across all groups. Context-

Table 9: Channel structure against available context, in 2048-token sequences. The intervention columns give the share of tokens that diverge under each single-channel intervention, and the median first flip is measured in layers past the ablation. Groups holding fewer than 20 tokens are omitted.

Model	Ablation	n_{ctx}	Intervention		Median first route flip		$\rho(\delta_{\text{dir}}+\delta_{\text{ctx}},\delta_{\text{joint}})$	Median err. / margin
			direct	context	direct	context		
Mixtral-8x7B	L1E2	32–63	55.9%	47.1%	2	10	0.9630	0.42%
		64+	55.7%	47.0%	3	11	0.9958	0.36%
Phi-3.5-MoE	L1E1	16–31	72.4%	41.4%	2	13	0.9417	0.82%
		32–63	57.1%	38.1%	5	14	0.9583	0.80%
		64+	62.4%	43.5%	3	13	0.9253	0.53%
OLMoE	L1E7	16–31	75.0%	77.5%	1	2	0.9991	0.94%
		32–63	78.8%	83.8%	1	3	0.9985	0.66%
		64+	84.7%	83.5%	1	2	0.9986	1.09%
Qwen3-30B-A3B	L1E15	4–7	77.1%	85.7%	3	9	0.9995	0.43%
		8–15	60.0%	78.7%	7	9	0.9994	0.23%
		16–31	64.6%	80.8%	5	10	0.9986	0.28%
		32–63	69.1%	86.8%	5	9	0.9987	0.46%
		64+	70.9%	95.3%	5	7	0.9955	0.30%
DeepSeek-V2-Lite	L1E37	8–15	70.8%	41.7%	3	12	0.9998	0.13%
		16–31	76.2%	31.0%	4	10	0.9999	0.11%
		32–63	75.9%	57.0%	3	12	0.9999	0.12%
		64+	74.7%	49.1%	3	11	0.9997	0.18%

mediated divergences occur deeper than direct ones in all 17 groups, by between 1 and 11 layers. We initially expected this delay to shrink at higher n_{ctx}, assuming that more sources would accumulate influence faster per layer. It does not: no model shows a consistent decline. This suggests that the delay is set by how attention propagates influence across layers rather than by how much is transported. The balance between the channels is likewise unsystematic, with neither share moving in a consistent direction as n_{ctx} grows.

J Derivations of the leverage estimators

J.1 The layernorm expansion

If no other intermediate computations change en route to layer ℓ, the perturbed input to the target MoE block is $\tilde{\boldsymbol{m}}^{\ell}_{\text{in},i} = \text{LN}^{\ell}_2(\boldsymbol{x}^{\ell}_i + \boldsymbol{a}^{\ell}_{\text{out},i} - w^u_i \boldsymbol{e}^u_{\text{out},i}) \approx \boldsymbol{m}^{\ell}_{\text{in},i} - \widehat{\text{LN}}^{\ell}_{2,i}(w^u_i \boldsymbol{e}^u_{\text{out},i})$, a first-order Taylor expansion of LN^{ℓ}_2 around its clean input $\boldsymbol{v} := \boldsymbol{x}^{\ell}_i + \boldsymbol{a}^{\ell}_{\text{out},i}$. Hence $\widehat{\text{LN}}^{\ell}_{2,i}$ is the Jacobian of the normalizer at $\boldsymbol{v}$ applied to the missing write, which for RMSNorm reads

$$\widehat{\text{LN}}^{\ell}_{2,i}(\boldsymbol{w}) = \frac{\boldsymbol{\gamma}}{\text{RMS}(\boldsymbol{v})} \odot \left(\boldsymbol{w} - \boldsymbol{v} \, \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{d \, \text{RMS}(\boldsymbol{v})^2} \right). \tag{17}$$

The same expansion at LN^{ℓ}_1 gives the perturbation of the target attention block's input at every position $j \le i$ as $\widehat{\text{LN}}^{\ell}_{1,j}(w^u_j \boldsymbol{e}^u_{\text{out},j})$; if j was not routed to u, this write is zero and the perturbation is absent.

J.2 The estimators in explicit form

Substituting each perturbation estimate of Section 6 into Equation 6 gives the three estimators. With only the write missing from token i's own residual stream,

$$J^{(0),\ell}_{u,i} = \max_{a \in \text{TopK}^{\ell}_i,\ b \notin \text{TopK}^{\ell}_i} \frac{[(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \widehat{\text{LN}}^{\ell}_{2,i}(w^u_i \boldsymbol{e}^u_{\text{out},i})]_+}{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}^{\ell}_{\text{in},i}}. \tag{18}$$

Letting the attention block at layer ℓ read the perturbed context adds a second summand:

$$J_{u,i}^{(1),\ell} = \max_{a \in \mathrm{TopK}_i^\ell,\, b \notin \mathrm{TopK}_i^\ell} \left[\underbrace{\frac{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \widehat{\mathrm{LN}}_{2,i}^{\ell}(w_i^u \boldsymbol{e}_{\mathrm{out},i}^u)}{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}_{\mathrm{in},i}^\ell}}_{\text{removed write}} + \underbrace{\frac{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \widehat{\mathrm{LN}}_{2,i}^{\ell}(\delta \boldsymbol{a}_{\mathrm{out},i}^\ell)}{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}_{\mathrm{in},i}^\ell}}_{\text{attention-mediated}} \right]_+ . \tag{19}$$

Advancing the recurrence to the router of the target layer yields $\boldsymbol{m}_{\mathrm{in},i}^\ell - \tilde{\boldsymbol{m}}_{\mathrm{in},i}^\ell \approx \widehat{\mathrm{LN}}_{2,i}^{\ell}(\delta \boldsymbol{x}_i^\ell + \delta \boldsymbol{a}_{\mathrm{out},i}^\ell)$ and the multi-layer estimator

$$J_{u,i}^{(\infty),\ell} = \max_{a \in \mathrm{TopK}_i^\ell,\, b \notin \mathrm{TopK}_i^\ell} \frac{[(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot (\boldsymbol{m}_{\mathrm{in},i}^\ell - \tilde{\boldsymbol{m}}_{\mathrm{in},i}^\ell)]_+}{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}_{\mathrm{in},i}^\ell}. \tag{20}$$

J.3 Coalition form

Every estimate of Section 6.1 is linear in the removed writes, so the estimated perturbations of several ablated experts add at a fixed routing boundary. For a coalition of experts S this gives

$$J_{S,i}^{\ell} = \max_{a \in \mathrm{TopK}_i^\ell,\, b \notin \mathrm{TopK}_i^\ell} \frac{[(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \sum_{u \in S} (\boldsymbol{m}_{\mathrm{in},i}^\ell - \tilde{\boldsymbol{m}}_{\mathrm{in},i}^{\ell,u})]_+}{(\boldsymbol{g}^{\ell,a} - \boldsymbol{g}^{\ell,b}) \cdot \boldsymbol{m}_{\mathrm{in},i}^\ell}. \tag{21}$$

Perturbations entering at different token positions are the same sum grouped differently: the direct and context-mediated channels of Section 4 collect the terms of the analyzed position and of the positions preceding it.

J.4 Attention perturbation in $J^{(1)}$

By Appendix J.1, the input of the target attention block is perturbed at every position $j \le i$ by $\widehat{\mathrm{LN}}_{1,j}^{\ell}(w_j^u \boldsymbol{e}_{\mathrm{out},j}^u)$. Because the input stream is perturbed, it produces perturbed keys, queries, and values. The resulting change in the attention output is

$$\delta \boldsymbol{a}_{\mathrm{out},i}^\ell = \boldsymbol{a}_{\mathrm{out},i}^\ell - \tilde{\boldsymbol{a}}_{\mathrm{out},i}^\ell = \sum_{h=1}^{H} \sum_{j \le i} \boldsymbol{W}_O^{\ell,h} (p_{i,j}^{\ell,h} \boldsymbol{v}_j^{\ell,h} - \tilde{p}_{i,j}^{\ell,h} \tilde{\boldsymbol{v}}_j^{\ell,h}). \tag{22}$$

Writing $\delta \boldsymbol{v}_j^{\ell,h} = \boldsymbol{v}_j^{\ell,h} - \tilde{\boldsymbol{v}}_j^{\ell,h}$ and $\delta p_{i,j}^{\ell,h} = p_{i,j}^{\ell,h} - \tilde{p}_{i,j}^{\ell,h}$, we can expand the parenthetical term to first order by dropping the $\delta p\, \delta \boldsymbol{v}$ product:

$$p_{i,j}^{\ell,h} \boldsymbol{v}_j^{\ell,h} - \tilde{p}_{i,j}^{\ell,h} \tilde{\boldsymbol{v}}_j^{\ell,h} \approx \underbrace{p_{i,j}^{\ell,h}\, \delta \boldsymbol{v}_j^{\ell,h}}_{\text{value route}} + \underbrace{\delta p_{i,j}^{\ell,h}\, \boldsymbol{v}_j^{\ell,h}}_{\text{attention weight route}} . \tag{23}$$

Because values are a linear projection of the normalized input, the value route is simply $\delta \boldsymbol{v}_j^{\ell,h} \approx \boldsymbol{W}_V^{\ell,h} \widehat{\mathrm{LN}}_{1,j}^{\ell}(w_j^u \boldsymbol{e}_{\mathrm{out},j}^u)$.

The weight route requires linearizing the attention scores and the softmax operation. The attention scores are bilinear in queries and keys. Dropping the second-order $\delta \boldsymbol{q} \cdot \delta \boldsymbol{k}$ term yields the score perturbation:

$$\delta s_{i,j}^{\ell,h} = s_{i,j}^{\ell,h} - \tilde{s}_{i,j}^{\ell,h} \approx \frac{\boldsymbol{q}_i^{\ell,h} \cdot \delta \boldsymbol{k}_j^{\ell,h} + \delta \boldsymbol{q}_i^{\ell,h} \cdot \boldsymbol{k}_j^{\ell,h}}{\sqrt{d_h}}, \tag{24}$$

where the first term captures the effect of context token j's perturbed key $\delta \boldsymbol{k}_j^{\ell,h} \approx \boldsymbol{W}_K^{\ell,h} \widehat{\mathrm{LN}}_{1,j}^{\ell}(w_j^u \boldsymbol{e}_{\mathrm{out},j}^u)$, and the second captures token i's own perturbed query $\delta \boldsymbol{q}_i^{\ell,h} \approx \boldsymbol{W}_Q^{\ell,h} \widehat{\mathrm{LN}}_{1,i}^{\ell}(w_i^u \boldsymbol{e}_{\mathrm{out},i}^u)$. Rotary embeddings are a fixed linear map at each position, so they rotate $\delta \boldsymbol{k}_j^{\ell,h}$ with $\boldsymbol{k}_j^{\ell,h}$ and leave the expansion unchanged.

To compute the attention weight perturbation $\delta p_{i,j}^{\ell,h}$ without explicitly rerunning the softmax, we linearize around the clean forward-pass scores $s_{i,j}^{\ell,h}$. Applying the softmax Jacobian $\frac{\partial p_{i,j}}{\partial s_{i,t}} = p_{i,j}(\mathbf{1}_{j=t} - p_{i,t})$ yields:

$$\delta p_{i,j}^{\ell,h} \approx \sum_{t \le i} \frac{\partial p_{i,j}^{\ell,h}}{\partial s_{i,t}^{\ell,h}} \, \delta s_{i,t}^{\ell,h} = p_{i,j}^{\ell,h} \left(\delta s_{i,j}^{\ell,h} - \sum_{t \le i} p_{i,t}^{\ell,h} \, \delta s_{i,t}^{\ell,h} \right). \tag{25}$$

Thus, the perturbation of an attention weight is determined by the deviation of its score perturbation from the *attention-weighted average* score perturbation. This centering term arises directly from the softmax normalization and ensures that the perturbed weights still sum to 1 to first order. Notably, since these weight perturbations sum to zero, the weight route depends only on how the attended values deviate from their attention-weighted mean; it vanishes entirely when the attended values are all alike, regardless of how large the induced key perturbation is.

J.5 Multi-layer propagation in $J^{(\infty)}$

Starting from the residual-stream delta $\delta \boldsymbol{x}_j^c = \boldsymbol{x}_j^c - \tilde{\boldsymbol{x}}_j^c$, which the ablation sets to $\delta \boldsymbol{x}_j^{\ell^u+1} = w_j^u \boldsymbol{e}_{\mathrm{out},j}^u$ at every position routed to u and $\mathbf{0}$ elsewhere, each layer $c \in \{\ell^u + 1, \ldots, \ell - 1\}$ advances all tokens jointly:

1. **Attention propagation:** Using the $J^{(1)}$ formulation evaluated on the accumulated input deltas $\delta \boldsymbol{x}_j^c$, with LN_1^c expanded around the clean inputs $\boldsymbol{x}_j^c$, attention outputs shift by $\delta \boldsymbol{a}_{\mathrm{out},i}^c = \boldsymbol{a}_{\mathrm{out},i}^c - \tilde{\boldsymbol{a}}_{\mathrm{out},i}^c$.
2. **Normalizer & MoE input:** Expanding LN_2^c to first order around its clean input $\boldsymbol{x}_i^c + \boldsymbol{a}_{\mathrm{out},i}^c$ yields the MoE input delta: $\delta \boldsymbol{m}_{\mathrm{in},i}^c = \boldsymbol{m}_{\mathrm{in},i}^c - \tilde{\boldsymbol{m}}_{\mathrm{in},i}^c \approx \widehat{\mathrm{LN}}_{2,i}^c(\delta \boldsymbol{x}_i^c + \delta \boldsymbol{a}_{\mathrm{out},i}^c)$.
3. **MoE block output:** The MoE block output shifts by $\delta \boldsymbol{m}_{\mathrm{out},i}^c = \boldsymbol{m}_{\mathrm{out},i}^c - \tilde{\boldsymbol{m}}_{\mathrm{out},i}^c$, derived below.
4. **Residual update:** The stream delta updates via $\delta \boldsymbol{x}_i^{c+1} = \delta \boldsymbol{x}_i^c + \delta \boldsymbol{a}_{\mathrm{out},i}^c + \delta \boldsymbol{m}_{\mathrm{out},i}^c$.

Section 6.1 holds the router selection at every intervening layer c fixed to the clean Top-K set, so that the MoE output $\boldsymbol{m}_{\mathrm{out},i}^c$ is a function of $\boldsymbol{m}_{\mathrm{in},i}^c$ alone. Expanding $\boldsymbol{m}_{\mathrm{out},i}^c = \sum_{n \in \mathrm{TopK}_i^c} w_i^{c,n} \boldsymbol{e}_{\mathrm{out},i}^{c,n}$ to first order via the product rule yields:

$$\delta \boldsymbol{m}_{\mathrm{out},i}^c \approx \frac{\partial \boldsymbol{m}_{\mathrm{out},i}^c}{\partial \boldsymbol{m}_{\mathrm{in},i}^c} \, \delta \boldsymbol{m}_{\mathrm{in},i}^c = \sum_{n \in \mathrm{TopK}_i^c} \Big(\underbrace{w_i^{c,n} \, \delta \boldsymbol{e}_{\mathrm{out},i}^{c,n}}_{\text{expert output shift}} + \underbrace{\delta w_i^{c,n} \, \boldsymbol{e}_{\mathrm{out},i}^{c,n}}_{\text{routing weight shift}} \Big). \tag{26}$$

Both constituents are nonlinear functions of $\boldsymbol{m}_{\mathrm{in},i}^c$ and require first-order expansions:

- **Expert output shift:** Let $\boldsymbol{h}_{\mathrm{gate},i}^{c,n} = \boldsymbol{W}_{\mathrm{gate}}^{c,n} \boldsymbol{m}_{\mathrm{in},i}^c$ and $\boldsymbol{h}_{\mathrm{up},i}^{c,n} = \boldsymbol{W}_{\mathrm{up}}^{c,n} \boldsymbol{m}_{\mathrm{in},i}^c$ be the clean projections. Linearizing the activation σ elementwise and applying the product rule gives:

$$\delta \boldsymbol{e}_{\mathrm{out},i}^{c,n} \approx \boldsymbol{W}_{\mathrm{down}}^{c,n} \Big[\big(\sigma'(\boldsymbol{h}_{\mathrm{gate},i}^{c,n}) \odot \boldsymbol{W}_{\mathrm{gate}}^{c,n} \delta \boldsymbol{m}_{\mathrm{in},i}^c \big) \odot \boldsymbol{h}_{\mathrm{up},i}^{c,n} + \sigma(\boldsymbol{h}_{\mathrm{gate},i}^{c,n}) \odot \boldsymbol{W}_{\mathrm{up}}^{c,n} \delta \boldsymbol{m}_{\mathrm{in},i}^c \Big]. \tag{27}$$

- **Routing weight shift:** Router logits shift linearly by $\boldsymbol{g}^{c,n} \cdot \delta \boldsymbol{m}_{\mathrm{in},i}^c$. Expanding the router softmax via its Jacobian yields:

$$\delta w_i^{c,n} \approx w_i^{c,n} \left[\boldsymbol{g}^{c,n} \cdot \delta \boldsymbol{m}_{\mathrm{in},i}^c - \sum_{n'} w_i^{c,n'} (\boldsymbol{g}^{c,n'} \cdot \delta \boldsymbol{m}_{\mathrm{in},i}^c) \right]. \tag{28}$$

  Each routing weight shifts proportional to its logit's deviation from the weighted average logit shift.

J.6 Where the first-order expansion fails

Section 6.2 traces the prediction failures to ablations that remove nearly the whole residual stream at $j^\star$. The reason lies in the normalizer's Jacobian.

Let $\boldsymbol{w} := w_{j^\star}^u \boldsymbol{e}_{\mathrm{out},j^\star}^u$ be the write the ablation removes at $j^\star$, and let $\boldsymbol{v} := \boldsymbol{x}_{j^\star}^\ell$ be the clean input of the attention block's normalizer, through which the write reaches every token that attends to $j^\star$. Split the write into the part aligned with that input and the rest,

$$\boldsymbol{w} = \alpha \boldsymbol{v} + \boldsymbol{w}_\perp, \qquad \boldsymbol{w}_\perp \perp \boldsymbol{v}, \tag{29}$$

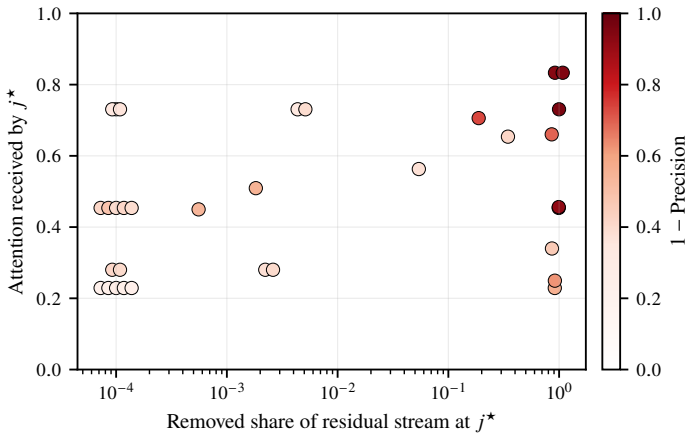


Figure 8: Prediction failure, as the share of incorrect $J^{(\infty)}$ predictions, against removed write size and attention received. $j^\star$ is the position receiving the most attention on the clean forward pass, measured at the layer following the ablation; each point is one ablation, reporting medians across sequences.

so that α is the fraction of the stream the ablation removes along the stream's own direction.

The first-order prediction ignores α, because the Jacobian keeps only the component of the write orthogonal to the stream. With $\mathrm{RMS}(\boldsymbol{v})^2 = \|\boldsymbol{v}\|^2/d$, it reads

$$\widehat{\mathrm{LN}}(\boldsymbol{w}) = \frac{\boldsymbol{\gamma}}{\mathrm{RMS}(\boldsymbol{v})} \odot \left(\boldsymbol{w} - \boldsymbol{v}\, \frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{v}\|^2} \right), \tag{30}$$

where the bracket is that component: it cancels $\alpha\boldsymbol{v}$ exactly, leaving $\widehat{\mathrm{LN}}(\boldsymbol{w}) = \boldsymbol{\gamma} \odot \boldsymbol{w}_\perp/\mathrm{RMS}(\boldsymbol{v})$, so the predicted perturbation is the same whether the ablation removes a tenth of the stream or all of it.

However, the true change does not ignore α. The ablated input is $\boldsymbol{v} - \boldsymbol{w} = (1-\alpha)\boldsymbol{v} - \boldsymbol{w}_\perp$, so as α approaches one its aligned part vanishes and it tends to $-\boldsymbol{w}_\perp$, which is orthogonal to $\boldsymbol{v}$. Clean and ablated inputs then point in unrelated directions while the prediction has not moved at all. The expansion therefore holds as long as $\boldsymbol{w}_\perp$ is small compared with the stream that remains, $(1-\alpha)\boldsymbol{v}$, and fails as α approaches one. A purely aligned removal ($\boldsymbol{w}_\perp = 0$, $\alpha < 1$) only rescales the input, which the normalizer ignores, so the prediction and the true change both vanish. Figure 8 places each ablation by the share of the stream it removes at $j^\star$ and by the attention $j^\star$ receives: the failing predictions are those at the right, where that share approaches one.

A second failure follows from the same condition. It leaves $j^\star$ with so little stream that the next write reroutes it, typically one layer after the ablation, which takes every token attending to it out of the clean-path regime of Section 6.1.

J.7 Architectural variants

Appendix A fixes one simple architecture for exposition. The models analyzed depart from it in several ways, each of which the formalism accommodates by instantiating terms it already contains. All reported experiments use the formalism adapted to each model's architecture.

Normalizer. Phi-3.5-MoE uses LayerNorm rather than RMSNorm. Its mean subtraction is linear, so the stream and the missing write are both centred before the Jacobian of Equation 17 is applied.

Query–key normalization. OLMoE normalizes the full query and key projections, Qwen3 each head separately. In both cases the norm adds a rank-one term to $\delta\boldsymbol{q}_i^{\ell,h}$ and $\delta\boldsymbol{k}_j^{\ell,h}$, over the dimension the model normalizes.

Attention variant. Perturbed keys and values are carried per key–value head and read by the query heads that share them, which covers grouped-query and multi-head self-attention alike and leaves

the value and weight routes of Appendix J.4 unchanged. DeepSeek's multi-head latent attention compresses keys and values into a shared latent and has no separate key and value projections. Our estimators do not implement this path, which is why Section 6.2 omits the model.

Biases. A bias is constant, so it cancels in a first-order difference and contributes no term.

Routing. The routing weights enter through the missing write $w_i^u \boldsymbol{e}_{\text{out},i}^u$ and through $\delta w_i^{c,n}$ in Appendix J.5. They are formed with each model's own rule: renormalized over the selected experts in Mixtral and Qwen3, unnormalized in OLMoE and in DeepSeek-V2, where selection is further restricted to the experts of the three highest-scoring of eight groups, and by SparseMixer's two masked softmaxes in Phi-3.5-MoE, whose masks are held at their clean-pass values.

Shared experts. The DeepSeek models run two shared experts per MoE layer alongside the routed ones. They are always active, so they form part of the block output and are untouched by ablating a routed expert; the interventions subtract only that expert's weighted output.

K First-flip predictions

Tables 10–12 give the per-hop precision and recall behind Figure 3, one table per estimator. Figure 9 and Tables 13–15 repeat this analysis with each expert ablated at every position except $j^\star$.

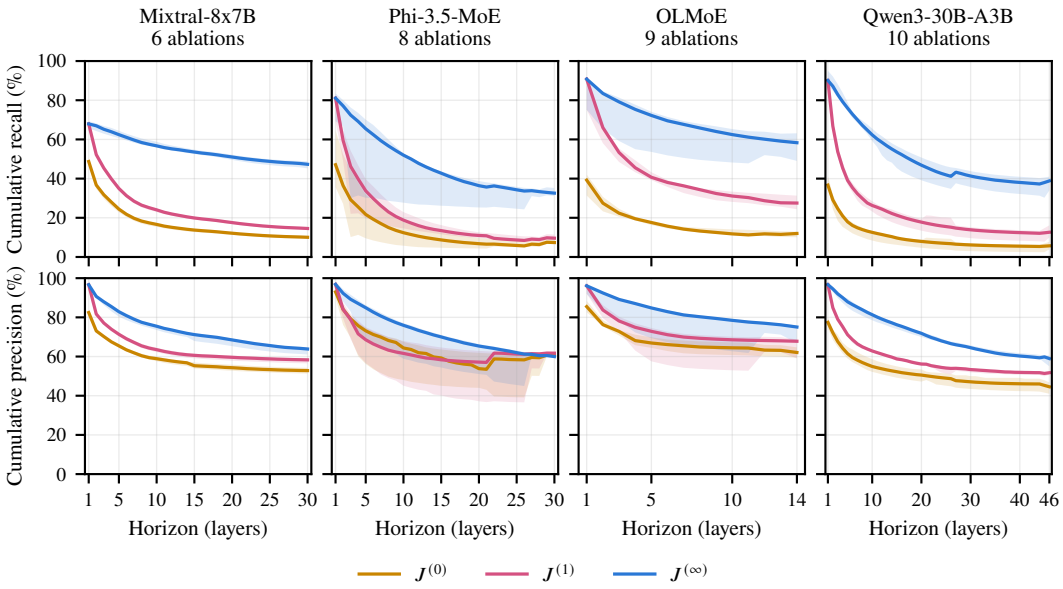


Figure 9: Same as Figure 3, over the same ablations, but with each expert ablated at every position except $j^\star$, the position later tokens attend to most on the clean forward pass.

The ablations whose predictions the write at $j^\star$ breaks (Section 6.2) are Mixtral-8x7B L1E3*, Phi-3.5-MoE L3E7*, OLMoE L1E9 and L2E30, and Qwen3-30B-A3B L1E68* and L2E92*. Sparing $j^\star$ raises recall on eight of the nine ablations that include one of these experts; Qwen3 L2E92* is the exception, gaining 37.4 points of precision while losing 10.5 of recall. Among the remaining ablations, Phi L5E10*, Phi L10E0* and Qwen3 L21E69* lose 3.4 to 4.9 points of precision, and these three together with Qwen3 L3E82* lose 2.3 to 22.6 points of recall, for reasons the write at $j^\star$ does not explain.

Table 10: First-flip prediction by $J^{(0)}$ across hop distances. F, P, and R denote flips, precision, and recall, respectively; all entries are percentages, and the *All* column sums counts over every hop.

Model	Ablation	H1			H2			H3			H4–8			H9–18			H19+			All		
		F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
Mixtral-8x7B	L1E2	6.5	85.8	49.1	4.9	58.7	22.0	3.1	48.9	13.5	14.5	32.2	5.0	14.8	23.7	2.0	13.4	15.0	0.5	57.2	55.6	10.1
	L1E3*	42.3	74.8	2.2	31.9	41.3	0.9	11.3	14.1	0.4	13.1	3.0	0.4	1.3	0.0	0.0	0.1	0.0	0.0	100.0	23.5	1.3
	L1E0	6.2	82.4	48.4	4.7	52.6	19.6	2.8	45.2	12.5	14.0	31.9	5.1	14.0	21.8	2.2	11.8	14.8	0.6	53.5	52.2	10.0
	L1E4	6.1	83.9	50.4	4.8	53.2	19.0	2.7	45.0	13.4	13.9	33.9	4.8	14.4	21.1	1.7	15.3	14.8	0.4	57.1	54.4	9.3
	L1E6	6.1	82.5	46.2	5.0	54.8	20.1	3.0	47.9	9.7	14.8	32.5	4.8	14.9	24.1	1.9	13.5	17.0	0.7	57.4	53.5	9.1
	L1E0+L1E4+L1E6	15.7	80.9	54.2	10.2	51.1	24.7	5.6	43.0	15.5	21.7	29.2	8.1	15.7	17.4	3.8	12.2	10.6	1.2	81.1	51.0	17.7
Phi-3.5-MoE	L3E7*	30.1	72.2	1.3	30.8	59.1	1.3	17.8	23.9	1.2	17.2	6.8	0.7	3.0	1.0	0.4	0.6	0.0	0.0	99.6	19.1	1.2
	L5E10*	43.6	69.6	2.0	9.5	23.6	3.8	5.6	33.0	5.3	11.5	18.2	3.1	10.7	25.1	4.4	4.9	9.6	0.7	85.9	30.3	2.8
	L10E0*	2.1	90.1	13.5	1.4	66.7	5.2	1.7	66.2	4.2	9.6	88.1	9.7	16.0	14.4	0.5	4.5	12.7	0.9	35.3	59.9	4.2
	L3E7*+L5E10*+L10E0*	30.1	72.2	1.3	30.8	59.1	1.3	25.4	24.3	1.7	13.0	5.5	2.4	0.6	0.1	0.5	0.0	0.0	0.0	100.0	11.7	1.6
	L1E1	3.3	96.4	65.9	2.5	69.8	36.4	2.0	55.4	19.5	10.6	41.6	5.9	22.1	30.5	1.5	17.3	14.3	0.2	57.8	63.1	7.7
	L1E9	1.5	95.5	52.8	1.2	69.9	27.4	1.1	54.1	14.6	6.8	48.0	8.1	16.7	35.7	1.7	16.1	24.8	0.3	43.4	58.2	5.0
	L1E13	2.1	95.8	64.3	2.2	74.3	40.4	1.8	62.6	26.7	8.6	43.9	7.3	20.0	28.9	1.5	17.3	16.2	0.2	51.9	60.8	7.1
	L1E1+L1E9+L1E13	5.7	95.3	68.3	4.5	71.1	43.8	3.5	58.2	27.5	14.7	44.8	11.3	24.2	29.2	3.2	16.8	16.8	0.6	69.3	60.6	13.5
OLMoE	L1E7	13.9	85.5	42.1	11.7	50.3	10.4	7.3	35.9	4.2	19.7	19.8	1.1	10.8	14.9	0.2	-	-	-	63.4	66.9	12.0
	L1E9	33.8	87.5	15.5	14.9	43.9	7.1	9.0	29.8	3.9	20.5	15.1	1.5	7.1	11.5	0.3	-	-	-	85.2	59.2	8.2
	L2E30	18.6	87.7	14.5	12.5	58.9	3.7	9.4	33.9	3.4	27.4	33.3	0.9	13.5	3.0	0.0	-	-	-	81.5	64.8	4.6
	L4E14	15.7	91.7	31.3	12.2	65.7	8.8	9.0	40.9	4.6	21.3	29.7	1.7	6.8	55.7	1.9	-	-	-	64.9	73.0	10.6
	L1E9+L2E30+L4E14	32.3	87.5	15.5	19.8	49.1	12.1	11.9	32.4	4.8	26.2	18.9	5.1	6.3	4.0	0.5	-	-	-	96.4	46.2	9.7
	L1E20	10.3	84.5	46.1	7.6	56.0	12.8	6.2	43.0	7.3	18.1	32.0	3.0	11.4	21.4	0.5	-	-	-	53.5	65.3	12.7
	L1E33	3.6	80.0	41.3	2.8	56.8	16.9	2.7	46.7	9.8	10.2	35.5	3.2	8.0	36.2	0.7	-	-	-	27.3	60.3	9.5
	L1E51	5.9	79.7	39.2	5.1	52.2	13.8	4.4	44.6	8.2	15.1	33.6	2.9	10.9	26.0	0.7	-	-	-	41.4	58.5	9.4
	L1E20+L1E33+L1E51	15.0	82.6	51.3	10.2	55.9	18.8	7.8	42.6	11.5	21.5	31.5	4.9	11.8	22.5	1.2	-	-	-	66.3	62.1	17.7
Qwen3-30B-A3B	L1E15	2.3	72.8	21.3	2.0	43.1	9.8	2.2	47.1	7.9	10.1	39.5	3.9	18.2	21.6	0.8	31.5	12.6	0.2	66.3	40.0	2.2
	L2E92*	18.9	99.3	3.6	12.9	6.9	0.0	10.8	4.8	0.0	32.9	7.9	0.0	19.0	0.0	0.0	5.2	0.0	0.0	99.7	71.7	0.7
	L1E68*	63.9	85.7	2.8	27.3	45.7	1.9	4.4	8.9	1.9	2.7	3.6	2.7	0.7	0.9	2.0	0.6	0.6	1.2	99.7	27.4	2.5
	L3E82*	7.5	93.5	14.7	7.2	55.4	1.4	9.4	49.4	0.7	29.2	34.4	0.3	29.7	14.0	0.1	15.3	4.0	0.0	98.2	69.5	1.4
	L21E69*	8.6	87.2	36.2	8.5	60.1	14.6	8.0	54.1	11.2	23.9	42.4	6.4	22.3	27.3	3.6	6.7	24.3	1.5	77.9	53.9	9.8
	L1E68*+L2E92*+L3E82*+L21E69*	63.9	85.7	2.8	29.0	46.2	1.9	4.5	7.9	2.3	2.4	2.1	2.4	0.2	0.1	1.0	0.0	0.0	0.0	100.0	10.9	2.5
	L1E5	2.3	93.1	35.8	2.1	63.2	10.5	2.2	52.6	2.9	9.9	38.5	1.7	14.7	23.9	0.6	26.2	6.5	0.0	57.4	58.7	2.4
	L1E23	5.4	76.1	40.2	4.1	57.8	20.4	3.7	42.7	11.8	14.0	30.2	4.7	20.4	20.4	1.6	28.4	13.7	0.6	75.9	44.6	6.1
	L1E77	4.9	79.2	41.8	4.3	62.0	20.4	4.2	45.6	10.0	16.0	32.9	4.3	21.2	21.8	1.4	27.6	12.5	0.5	78.2	47.7	5.7
	L1E5+L1E23+L1E77	9.2	78.4	45.5	7.0	60.8	23.9	6.3	44.2	12.5	20.7	29.2	5.6	21.9	19.7	2.5	22.3	9.6	0.9	87.4	46.0	9.8

Table 11: First-flip prediction by $J^{(1)}$ across hop distances. Columns as in Table 10.

Model	Ablation	H1			H2			H3			H4–8			H9–18			H19+			All		
		F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
Mixtral-8x7B	L1E2	6.5	96.5	68.0	4.9	59.6	32.5	3.1	49.2	17.7	14.5	35.0	7.4	14.8	25.3	3.1	13.4	17.1	0.7	57.2	59.5	14.3
	L1E3*	42.3	72.0	3.1	31.9	41.2	2.1	11.3	20.4	1.1	13.1	2.8	0.8	1.3	0.3	1.8	0.1	0.0	0.0	100.0	8.1	2.2
	L1E0	6.2	97.1	69.8	4.7	55.4	30.4	2.8	46.2	16.9	14.0	36.6	8.1	14.0	24.0	3.3	11.8	18.9	0.8	53.5	58.2	14.8
	L1E4	6.1	97.3	72.3	4.8	58.7	32.9	2.7	47.3	20.1	13.9	39.1	8.8	14.4	29.4	4.3	15.3	30.5	2.7	57.1	58.7	15.4
	L1E6	6.1	95.8	67.8	5.0	56.2	29.9	3.0	49.8	16.1	14.8	34.6	7.4	14.9	28.0	3.1	13.5	18.0	0.8	57.4	58.4	13.6
	L1E0+L1E4+L1E6	15.7	96.9	65.1	10.2	54.9	31.6	5.6	44.4	19.5	21.7	31.6	10.6	15.7	17.9	4.8	12.2	15.0	2.2	81.1	55.5	22.0
Phi-3.5-MoE	L3E7*	30.1	78.7	2.9	30.8	48.7	2.5	17.8	23.0	3.3	17.2	5.3	7.1	3.0	1.0	8.0	0.6	0.4	11.4	99.6	5.5	3.8
	L5E10*	43.6	86.2	6.4	9.5	27.7	6.9	5.6	25.8	11.8	11.5	12.3	9.6	10.7	12.0	10.8	4.9	6.2	9.7	85.9	19.9	8.0
	L10E0*	2.1	88.1	54.0	1.4	65.9	34.0	1.7	52.4	26.0	9.6	51.6	33.7	16.0	35.5	16.4	4.5	30.3	19.8	35.3	45.3	25.0
	L3E7*+L5E10*+L10E0*	30.1	78.7	2.9	30.8	48.7	2.5	25.4	28.0	3.6	13.0	3.3	7.0	0.6	0.1	3.9	0.0	0.0	16.7	100.0	4.8	3.5
	L1E1	3.3	98.3	81.0	2.5	69.4	42.2	2.0	55.4	24.4	10.6	42.3	9.1	22.1	33.8	2.3	17.3	17.4	0.4	57.8	61.9	9.9
	L1E9	1.5	98.4	86.5	1.2	64.6	36.3	1.1	53.1	21.5	6.8	49.6	10.8	16.7	37.2	2.1	16.1	25.9	0.4	43.4	61.2	7.3
	L1E13	2.1	97.9	86.2	2.2	73.4	47.8	1.8	60.7	30.9	8.6	46.9	10.6	20.0	31.9	2.1	17.3	24.2	0.4	51.9	62.0	9.3
	L1E1+L1E9+L1E13	5.7	98.4	82.9	4.5	70.9	47.7	3.5	59.6	32.4	14.7	44.5	13.7	24.2	30.4	3.8	16.8	18.7	0.8	69.3	61.6	16.0
OLMoE	L1E7	13.9	96.1	90.7	11.7	47.6	36.2	7.3	45.1	18.7	19.7	32.3	7.7	10.8	25.3	1.2	-	-	-	63.4	65.6	31.3
	L1E9	33.8	87.8	31.9	14.9	43.5	17.6	9.0	28.5	15.9	20.5	15.0	21.1	7.1	9.2	1.9	-	-	-	85.2	35.9	22.6
	L2E30	18.6	91.1	35.5	12.5	62.0	16.6	9.4	45.5	12.9	27.4	34.7	8.9	13.5	20.3	16.4	-	-	-	81.5	46.6	17.9
	L4E14	15.7	55.3	75.2	12.2	36.3	36.2	9.0	41.9	18.7	21.3	33.7	7.9	6.8	36.9	3.3	-	-	-	64.9	45.9	30.5
	L1E9+L2E30+L4E14	32.3	87.8	31.9	19.8	53.1	26.1	11.9	32.1	16.9	26.2	20.6	13.4	6.3	6.6	9.5	-	-	-	96.4	40.1	22.3
	L1E20	10.3	97.1	90.9	7.6	62.2	32.5	6.2	49.8	21.3	18.1	36.3	8.0	11.4	24.9	1.1	-	-	-	53.5	71.0	27.5
	L1E33	3.6	96.1	91.0	2.8	60.9	34.2	2.7	51.9	22.2	10.2	39.6	7.7	8.0	36.6	1.5	-	-	-	27.3	68.0	21.0
	L1E51	5.9	96.5	92.6	5.1	59.6	35.1	4.4	51.1	20.7	15.1	38.8	7.8	10.9	29.1	1.4	-	-	-	41.4	67.8	22.9
	L1E20+L1E33+L1E51	15.0	96.3	91.0	10.2	62.6	37.0	7.8	49.1	24.0	21.5	34.4	9.9	11.8	22.5	1.9	-	-	-	66.3	69.4	32.6
Qwen3-30B-A3B	L1E15	2.3	97.2	94.4	2.0	52.7	30.8	2.2	52.0	23.6	10.1	44.8	11.5	18.2	29.4	2.7	31.5	19.9	0.9	66.3	51.7	7.9
	L2E92*	18.9	99.6	16.7	12.9	62.2	18.6	10.8	54.1	12.5	32.9	37.2	12.9	19.0	11.9	8.6	5.2	1.6	7.3	99.7	22.8	13.2
	L1E68*	63.9	87.0	8.5	27.3	34.6	3.4	4.4	6.5	3.1	2.7	2.3	4.9	0.7	0.5	4.0	0.6	0.2	1.9	99.7	21.7	6.7
	L3E82*	7.5	98.6	71.1	7.2	73.9	27.9	9.4	72.8	38.0	29.2	52.2	16.5	29.7	24.5	11.1	15.3	6.2	8.8	98.2	35.5	20.7
	L21E69*	8.6	91.2	55.0	8.5	67.1	26.1	8.0	58.1	20.0	23.9	29.9	23.9	22.3	20.3	18.0	6.7	17.7	5.6	77.9	35.7	23.9
	L1E68*+L2E92*+L3E82*+L21E69*	63.9	87.0	8.5	29.0	37.2	4.3	4.5	6.8	4.6	2.4	1.2	4.1	0.2	0.0	3.0	0.0	0.0	0.0	100.0	9.2	7.0
	L1E5	2.3	93.8	80.0	2.1	67.2	30.6	2.2	54.4	19.7	9.9	40.4	8.9	14.7	23.8	2.3	26.2	14.1	0.7	57.4	50.5	7.5
	L1E23	5.4	97.0	89.2	4.1	65.8	40.8	3.7	48.0	24.2	14.0	36.7	10.2	20.4	24.9	3.9	28.4	15.6	1.4	75.9	52.6	13.2
	L1E77	4.9	95.4	87.3	4.3	68.8	37.1	4.2	56.4	26.6	16.0	39.0	11.2	21.2	24.0	3.4	27.6	14.4	1.2	78.2	52.6	12.6
	L1E5+L1E23+L1E77	9.2	96.0	87.5	7.0	67.2	41.7	6.3	53.1	26.7	20.7	35.8	12.6	21.9	21.3	4.9	22.3	11.0	2.0	87.4	52.0	19.1

Table 12: First-flip prediction by $J^{(\infty)}$ across hop distances. Columns as in Table 10.

Model	Ablation	H1			H2			H3			H4–8			H9–18			H19+			All		
		F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
Mixtral-8x7B	L1E2	6.5	96.5	68.0	4.9	84.5	65.9	3.1	77.5	59.7	14.5	66.9	52.7	14.8	54.3	38.3	13.4	40.5	26.0	57.2	63.7	45.9
	L1E3*	42.3	72.0	3.1	31.9	41.0	2.4	11.3	20.6	2.0	13.1	3.0	1.8	1.3	0.2	2.8	0.1	0.0	4.9	100.0	4.3	2.6
	L1E0	6.2	97.1	69.8	4.7	84.6	66.9	2.8	77.4	62.5	14.0	68.5	54.9	14.0	56.3	40.6	11.8	43.9	31.1	53.5	65.4	49.1
	L1E4	6.1	97.3	72.3	4.8	83.8	67.6	2.7	77.6	64.9	13.9	68.9	55.3	14.4	58.2	43.2	15.3	51.3	36.7	57.1	66.6	50.5
	L1E6	6.1	95.8	67.8	5.0	84.7	65.5	3.0	79.3	57.5	14.8	68.0	51.0	14.9	53.0	36.8	13.5	42.5	26.7	57.4	64.0	45.0
	L1E0+L1E4+L1E6	15.7	96.9	65.1	10.2	79.5	61.7	5.6	71.1	57.3	21.7	58.2	50.4	15.7	41.6	36.9	12.2	36.4	33.1	81.1	60.6	49.9
Phi-3.5-MoE	L3E7*	30.1	78.7	2.9	30.8	52.4	3.0	17.8	25.8	4.3	17.2	5.6	6.0	3.0	1.0	6.9	0.6	0.3	6.2	99.6	6.6	3.9
	L5E10*	43.6	86.2	6.4	9.5	35.7	11.3	5.6	37.8	19.5	11.5	24.4	32.1	10.7	21.2	34.5	4.9	14.6	27.7	85.9	26.9	16.0
	L10E0*	2.1	88.1	54.0	1.4	78.7	62.0	1.7	75.2	52.4	9.6	70.1	46.1	16.0	53.3	31.9	4.5	49.8	30.5	35.3	62.1	39.1
	L3E7*+L5E10*+L10E0*	30.1	78.7	2.9	30.8	52.4	3.0	25.4	29.0	4.2	13.0	3.4	6.6	0.6	0.1	5.2	0.0	0.0	0.0	100.0	5.4	3.8
	L1E1	3.3	98.3	81.0	2.5	88.1	72.8	2.0	80.5	62.9	10.6	69.9	45.8	22.1	52.1	26.0	17.3	34.7	16.8	57.8	58.9	33.3
	L1E9	1.5	98.4	86.5	1.2	91.2	78.5	1.1	85.8	67.3	6.8	75.4	49.5	16.7	55.6	22.8	16.1	40.5	14.1	43.4	61.8	28.7
	L1E13	2.1	97.9	86.2	2.2	90.8	78.2	1.8	86.7	68.8	8.6	69.8	47.4	20.0	53.5	24.4	17.3	37.6	16.4	51.9	59.8	31.8
	L1E1+L1E9+L1E13	5.7	98.4	82.9	4.5	88.9	75.1	3.5	81.9	65.0	14.7	69.0	50.4	24.2	49.4	29.7	16.8	32.0	20.1	69.3	60.1	40.8
OLMoE	L1E7	13.9	96.1	90.7	11.7	88.1	77.2	7.3	78.1	71.0	19.7	65.6	53.2	10.8	51.6	37.1	-	-	-	63.4	76.8	65.2
	L1E9	33.8	87.8	31.9	14.9	51.6	31.0	9.0	40.4	28.7	20.5	26.6	25.0	7.1	17.4	22.9	-	-	-	85.2	43.9	29.0
	L2E30	18.6	91.1	35.5	12.5	72.8	27.7	9.4	59.8	28.4	27.4	48.3	25.6	13.5	31.8	34.0	-	-	-	81.5	53.6	29.9
	L4E14	15.7	55.3	75.2	12.2	53.0	52.7	9.0	43.7	40.1	21.3	35.9	27.3	6.8	30.9	25.6	-	-	-	64.9	46.3	45.2
	L1E9+L2E30+L4E14	32.3	87.8	31.9	19.8	55.7	33.3	11.9	42.8	27.4	26.2	23.9	33.0	6.3	11.3	22.0	-	-	-	96.4	37.9	31.3
	L1E20	10.3	97.1	90.9	7.6	86.8	75.6	6.2	80.6	69.2	18.1	66.9	51.0	11.4	52.2	35.9	-	-	-	53.5	75.7	61.1
	L1E33	3.6	96.1	91.0	2.8	85.2	73.8	2.7	80.4	63.0	10.2	68.9	44.1	8.0	55.6	30.8	-	-	-	27.3	74.1	51.3
	L1E51	5.9	96.5	92.6	5.1	87.6	79.0	4.4	80.2	69.1	15.1	69.1	50.3	10.9	56.8	37.2	-	-	-	41.4	75.1	58.4
	L1E20+L1E33+L1E51	15.0	96.3	91.0	10.2	87.3	77.1	7.8	78.6	69.6	21.5	66.3	54.0	11.8	50.7	39.8	-	-	-	66.3	76.0	65.2
Qwen3-30B-A3B	L1E15	2.3	97.2	94.4	2.0	91.8	86.2	2.2	86.9	76.1	10.1	78.5	54.3	18.2	58.0	29.1	31.5	32.3	13.3	66.3	58.5	31.0
	L2E92*	18.9	99.6	16.7	12.9	74.0	24.3	10.8	65.0	28.3	32.9	46.8	32.2	19.0	21.5	32.3	5.2	5.9	30.1	99.7	30.8	27.7
	L1E68*	63.9	87.0	8.5	27.3	32.9	5.8	4.4	7.4	7.3	2.7	2.9	15.3	0.7	1.3	29.6	0.6	0.5	14.3	99.7	12.1	8.1
	L3E82*	7.5	98.6	71.1	7.2	94.8	63.1	9.4	89.3	62.9	29.2	77.2	58.0	29.7	53.1	48.4	15.3	24.1	39.4	98.2	58.5	54.0
	L21E69*	8.6	91.2	55.0	8.5	79.1	37.3	8.0	69.6	34.8	23.9	48.0	37.0	22.3	31.5	39.8	6.7	26.3	31.5	77.9	45.0	39.1
	L1E68*+L2E92*+L3E82*+L21E69*	63.9	87.0	8.5	29.0	33.2	6.1	4.5	6.3	7.2	2.4	1.1	7.8	0.2	0.1	17.2	0.0	0.0	23.1	100.0	8.3	7.7
	L1E5	2.3	93.8	80.0	2.1	83.2	66.8	2.2	76.8	55.5	9.9	63.5	43.1	14.7	47.0	24.6	26.2	31.8	13.6	57.4	51.6	27.7
	L1E23	5.4	97.0	89.2	4.1	91.4	83.5	3.7	84.2	74.2	14.0	73.6	55.7	20.4	55.2	31.5	28.4	29.4	15.1	75.9	60.4	38.8
	L1E77	4.9	95.4	87.3	4.3	90.5	82.3	4.2	82.7	73.5	16.0	72.2	56.1	21.2	53.2	32.0	27.6	29.3	16.7	78.2	59.0	40.0
	L1E5+L1E23+L1E77	9.2	96.0	87.5	7.0	88.9	81.6	6.3	82.1	71.8	20.7	69.7	57.5	21.9	50.1	36.1	22.3	26.6	20.1	87.4	60.8	48.7

Table 13: First-flip prediction by $J^{(0)}$ across hop distances. Each expert is ablated at every position except $j^\star$, the position later tokens attend to most on the clean forward pass. F, P, and R denote flips, precision, and recall, respectively; all entries are percentages, and the *All* column sums counts over every hop.

Model	Ablation	H1			H2			H3			H4–8			H9–18			H19+			All		
		F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
Mixtral-8x7B	L1E2	6.5	85.8	49.1	4.9	58.7	22.0	3.1	48.9	13.5	14.5	32.2	5.0	14.8	23.7	2.0	13.4	15.0	0.5	57.2	55.6	10.1
	L1E3*	4.1	82.1	24.8	2.9	57.8	14.6	1.8	51.0	9.2	9.6	41.2	4.7	10.6	17.4	2.1	9.7	23.7	0.5	38.7	47.7	6.0
	L1E0	6.2	82.4	48.4	4.7	52.6	19.6	2.8	45.2	12.5	14.0	31.9	5.1	14.0	21.8	2.2	11.8	14.8	0.6	53.5	52.2	10.0
	L1E4	6.0	84.0	51.3	4.6	53.9	20.0	2.6	44.4	13.5	13.5	34.0	5.0	13.1	20.3	1.8	11.4	16.5	0.7	51.1	54.5	10.4
	L1E6	6.1	82.5	46.2	5.0	54.8	20.1	3.0	47.9	9.7	14.8	32.5	4.8	14.9	24.1	1.9	13.5	17.0	0.7	57.4	53.5	9.1
	L1E0+L1E4+L1E6	15.6	80.8	54.2	10.2	51.4	24.9	5.6	42.5	15.4	21.5	29.1	8.1	15.6	17.7	3.9	10.8	12.7	1.6	79.3	51.0	18.1
Phi-3.5-MoE	L3E7*	1.2	90.7	41.5	1.8	64.2	24.5	2.4	50.3	16.6	10.9	39.3	6.7	19.2	32.5	2.0	13.4	28.2	0.5	48.9	47.4	5.1
	L5E10*	43.0	70.4	2.0	9.2	23.6	4.0	5.4	34.0	5.8	10.5	19.6	3.6	9.2	10.7	1.4	4.3	4.0	0.2	81.6	29.0	2.5
	L10E0*	1.5	88.1	18.0	0.9	66.7	8.1	1.0	61.8	6.3	6.2	70.9	4.8	11.8	14.2	0.7	3.4	14.0	1.1	24.8	46.8	3.3
	L3E7*+L5E10*+L10E0*	1.2	90.7	41.5	1.8	64.2	24.5	42.1	67.3	2.9	23.7	22.6	5.4	12.1	11.8	2.9	5.6	4.5	0.6	86.4	30.6	4.4
	L1E1	3.3	96.4	65.9	2.5	69.8	36.4	2.0	55.4	19.5	10.6	41.6	5.9	22.1	30.5	1.5	17.3	14.3	0.2	57.8	63.1	7.7
	L1E9	1.5	95.5	52.7	1.2	69.6	27.4	1.1	54.6	14.9	6.8	48.0	8.1	16.6	34.5	1.7	15.9	23.1	0.2	43.1	57.9	5.0
	L1E13	2.1	95.8	64.3	2.2	74.3	40.4	1.8	62.6	26.7	8.6	43.9	7.3	20.0	28.9	1.5	17.3	16.2	0.2	51.9	60.8	7.1
	L1E1+L1E9+L1E13	5.7	95.3	68.3	4.5	71.2	43.8	3.5	58.3	27.6	14.7	44.7	11.3	24.1	29.0	3.2	16.8	17.2	0.6	69.2	60.6	13.5
OLMoE	L1E7	13.9	85.5	42.1	11.7	50.3	10.4	7.3	35.9	4.2	19.7	19.8	1.1	10.8	14.9	0.2	-	-	-	63.4	66.9	12.0
	L1E9	21.5	88.4	24.6	10.9	52.1	10.4	7.2	35.3	5.4	19.2	20.8	2.0	9.5	21.5	0.5	-	-	-	68.2	64.1	10.6
	L2E30	7.2	87.8	35.4	4.2	59.3	11.0	3.5	68.9	9.5	12.6	39.7	2.3	7.8	22.9	0.2	-	-	-	35.4	73.3	10.3
	L4E14	15.7	91.8	31.4	12.2	66.0	8.8	9.0	40.7	4.5	21.1	25.7	1.4	6.5	15.1	0.3	-	-	-	64.5	72.6	10.4
	L1E9+L2E30+L4E14	20.5	88.4	24.6	13.5	61.5	21.3	8.0	40.8	8.9	32.5	36.6	7.5	9.6	12.1	0.8	-	-	-	84.2	57.5	13.2
	L1E20	10.3	84.5	46.1	7.6	56.0	12.8	6.2	43.0	7.3	18.1	32.0	3.0	11.4	21.4	0.5	-	-	-	53.5	65.3	12.7
	L1E33	3.6	80.0	41.4	2.8	56.8	16.9	2.6	46.4	9.8	10.2	35.5	3.2	8.0	36.2	0.7	-	-	-	27.3	60.3	9.6
	L1E51	5.9	79.6	39.2	5.0	52.2	13.8	4.4	44.3	8.2	15.1	33.7	2.9	10.9	26.1	0.7	-	-	-	41.3	58.5	9.4
	L1E20+L1E33+L1E51	15.0	82.6	51.3	10.2	55.9	18.9	7.8	42.6	11.5	21.5	31.4	4.9	11.8	22.5	1.2	-	-	-	66.3	62.1	17.7
Qwen3-30B-A3B	L1E15	2.3	72.8	21.3	2.0	43.1	9.8	2.2	47.1	7.9	10.1	39.5	3.9	18.2	21.6	0.8	31.5	12.6	0.2	66.3	40.0	2.2
	L2E92*	0.0	76.5	61.9	0.0	66.7	34.5	0.0	75.0	30.0	0.2	48.5	10.5	0.7	36.4	1.8	2.5	15.0	0.2	3.6	48.7	2.5
	L1E68*	3.1	74.0	36.7	2.9	59.0	20.9	2.9	50.9	14.7	12.0	36.5	6.1	18.1	23.8	2.1	26.6	15.3	0.7	65.6	42.0	5.3
	L3E82*	1.0	80.2	32.3	0.9	60.2	12.1	0.9	44.4	6.3	3.8	30.6	2.2	7.5	16.1	0.3	12.2	8.0	0.1	26.4	50.1	2.3
	L21E69*	8.1	86.6	36.5	8.0	61.3	15.3	7.8	54.6	11.6	22.3	40.5	6.2	20.6	23.9	3.1	6.1	19.8	1.2	73.0	53.1	9.9
	L1E68*+L2E92*+L3E82*+L21E69*	3.1	74.0	36.7	3.0	59.4	21.4	3.6	59.4	21.1	14.0	37.8	6.9	19.9	24.1	2.2	43.5	30.6	9.8	87.1	37.0	9.4
	L1E5	2.3	93.1	35.9	2.1	63.2	10.5	2.2	52.6	2.9	9.8	38.5	1.7	14.7	23.1	0.6	26.2	6.5	0.0	57.4	58.7	2.4
	L1E23	5.4	76.1	40.2	4.1	57.8	20.4	3.7	42.7	11.8	14.0	30.2	4.7	20.4	20.4	1.6	28.4	13.5	0.6	75.9	44.6	6.1
	L1E77	4.9	79.2	41.8	4.3	62.0	20.3	4.2	45.6	10.0	16.0	32.9	4.3	21.2	21.8	1.4	27.6	12.1	0.4	78.2	47.6	5.7
	L1E5+L1E23+L1E77	9.2	78.4	45.5	7.0	60.8	23.9	6.3	44.2	12.5	20.7	29.2	5.6	21.9	19.7	2.4	22.3	9.4	0.8	87.4	46.0	9.7

Table 14: First-flip prediction by $J^{(1)}$ across hop distances. Columns and intervention as in Table 13.

Model	Ablation	H1			H2			H3			H4–8			H9–18			H19+			All		
		F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
Mixtral-8x7B	L1E2	6.5	96.5	68.0	4.9	59.6	32.5	3.1	49.2	17.7	14.5	35.0	7.4	14.8	25.3	3.1	13.4	17.1	0.7	57.2	59.5	14.3
	L1E3*	4.1	87.2	34.4	2.9	58.9	19.1	1.8	53.4	11.8	9.6	41.4	5.9	10.6	22.6	4.1	9.7	15.2	4.9	38.7	39.1	9.5
	L1E0	6.2	97.1	69.8	4.7	55.4	30.4	2.8	46.2	16.9	14.0	36.6	8.1	14.0	24.0	3.3	11.8	18.9	0.8	53.5	58.2	14.8
	L1E4	6.0	96.9	71.5	4.6	56.7	31.0	2.6	46.8	18.9	13.5	37.2	8.1	13.1	23.8	3.1	11.4	17.8	0.8	51.1	59.3	15.2
	L1E6	6.1	95.8	67.8	5.0	56.2	29.9	3.0	49.8	16.1	14.8	34.6	7.4	14.9	28.0	3.1	13.5	18.0	0.8	57.4	58.4	13.6
	L1E0+L1E4+L1E6	15.6	96.8	65.0	10.2	54.6	31.4	5.6	43.9	19.1	21.5	31.0	10.4	15.6	17.9	4.7	10.8	12.4	1.5	79.3	56.0	22.1
Phi-3.5-MoE	L3E7*	1.2	95.7	81.1	1.8	66.6	37.9	2.4	49.5	23.5	10.9	42.7	10.2	19.2	34.1	2.8	13.4	27.0	0.6	48.9	51.5	8.1
	L5E10*	43.0	81.0	4.9	9.2	21.3	5.7	5.4	21.0	10.8	10.5	9.5	8.4	9.2	3.4	3.2	4.3	1.1	0.8	81.6	15.3	5.4
	L10E0*	1.5	83.4	34.1	0.9	57.8	22.6	1.0	48.0	18.4	6.2	45.0	13.7	11.8	30.1	7.2	3.4	31.3	9.1	24.8	41.4	11.8
	L3E7*+L5E10*+L10E0*	1.2	95.7	81.1	1.8	66.6	37.9	42.1	74.3	5.5	23.7	16.2	8.9	12.1	6.2	5.2	5.6	2.3	1.8	86.4	20.7	7.9
	L1E1	3.3	98.3	81.0	2.5	69.4	42.2	2.0	55.4	24.4	10.6	42.3	9.1	22.1	33.8	2.3	17.3	17.4	0.4	57.8	61.9	9.9
	L1E9	1.5	98.4	85.6	1.2	64.3	35.8	1.1	53.2	21.8	6.8	49.3	10.8	16.6	35.8	2.1	15.9	25.5	0.4	43.1	60.8	7.2
	L1E13	2.1	97.9	86.2	2.2	73.4	47.8	1.8	60.7	30.9	8.6	46.9	10.6	20.0	31.9	2.1	17.3	24.2	0.4	51.9	62.0	9.3
	L1E1+L1E9+L1E13	5.7	98.4	82.9	4.5	70.8	47.6	3.5	59.7	32.5	14.7	44.5	13.7	24.1	30.2	3.8	16.8	19.2	0.8	69.2	61.5	16.0
OLMoE	L1E7	13.9	96.1	90.8	11.7	47.6	36.2	7.3	45.1	18.7	19.7	32.2	7.7	10.8	25.2	1.2	-	-	-	63.4	65.6	31.3
	L1E9	21.5	92.0	48.8	10.9	53.4	26.3	7.2	36.3	18.9	19.2	18.3	14.3	9.5	16.6	2.0	-	-	-	68.2	48.1	25.8
	L2E30	7.2	97.5	86.5	4.2	60.4	31.4	3.5	60.2	26.0	12.6	39.8	8.1	7.8	26.7	0.9	-	-	-	35.4	73.9	26.9
	L4E14	15.7	55.2	75.1	12.2	36.3	36.2	9.0	41.6	18.6	21.1	32.8	7.5	6.5	18.4	1.3	-	-	-	64.5	45.6	30.3
	L1E9+L2E30+L4E14	20.5	92.0	48.8	13.5	66.5	44.3	8.0	42.1	24.1	32.5	33.9	25.2	9.6	14.3	2.4	-	-	-	84.2	52.4	31.3
	L1E20	10.3	97.1	90.9	7.6	62.1	32.4	6.2	49.7	21.3	18.1	36.3	8.0	11.4	24.9	1.1	-	-	-	53.5	71.0	27.5
	L1E33	3.6	96.1	90.9	2.8	61.1	34.2	2.6	51.9	22.3	10.2	39.7	7.7	8.0	36.8	1.5	-	-	-	27.3	68.1	21.1
	L1E51	5.9	96.4	92.6	5.0	59.7	35.2	4.4	50.8	20.6	15.1	38.9	7.8	10.9	29.4	1.4	-	-	-	41.3	67.8	22.9
	L1E20+L1E33+L1E51	15.0	96.3	90.9	10.2	62.6	37.0	7.8	49.1	24.1	21.5	34.4	9.9	11.8	22.7	1.9	-	-	-	66.3	69.4	32.6
Qwen3-30B-A3B	L1E15	2.3	97.2	94.4	2.0	52.7	30.8	2.2	52.0	23.6	10.1	44.8	11.5	18.2	29.4	2.7	31.5	19.9	0.9	66.3	51.7	7.9
	L2E92*	0.0	100.0	95.2	0.0	68.2	51.7	0.0	62.5	25.0	0.2	39.5	11.1	0.7	36.4	2.6	2.5	21.9	0.4	3.6	48.1	3.3
	L1E68*	3.1	96.7	95.9	2.9	63.5	38.9	2.9	56.5	27.9	12.0	43.0	14.0	18.1	29.9	4.7	26.6	16.5	1.4	65.6	51.0	12.0
	L3E82*	1.0	96.9	90.9	0.9	68.0	38.3	0.9	58.1	21.9	3.8	44.5	13.2	7.5	31.3	3.9	12.2	19.8	1.2	26.4	52.3	9.1
	L21E69*	8.1	93.0	51.6	8.0	67.7	23.0	7.8	59.1	17.8	22.3	27.3	20.6	20.6	18.2	17.0	6.1	16.3	4.3	73.0	33.3	21.6
	L1E68*+L2E92*+L3E82*+L21E69*	3.1	96.7	95.9	3.0	64.0	39.4	3.6	67.5	42.3	14.0	45.8	17.0	19.9	29.6	5.5	43.5	20.0	18.7	87.1	30.4	19.8
	L1E5	2.3	93.8	79.8	2.1	67.5	30.9	2.2	54.4	19.8	9.8	40.4	8.9	14.7	23.6	2.3	26.2	14.2	0.7	57.4	50.5	7.5
	L1E23	5.4	97.0	89.3	4.1	65.7	40.8	3.7	48.0	24.2	14.0	36.7	10.2	20.4	24.8	3.9	28.4	15.5	1.4	75.9	52.6	13.2
	L1E77	4.9	95.4	87.4	4.3	68.7	37.1	4.2	56.3	26.6	16.0	39.0	11.2	21.2	24.2	3.4	27.6	14.1	1.2	78.2	52.6	12.6
	L1E5+L1E23+L1E77	9.2	96.0	87.4	7.0	67.2	41.6	6.3	53.0	26.7	20.7	35.9	12.6	21.9	21.2	4.9	22.3	10.9	1.9	87.4	52.0	19.1

Table 15: First-flip prediction by $J^{(\infty)}$ across hop distances. Columns and intervention as in Table 13.

Model	Ablation	H1			H2			H3			H4–8			H9–18			H19+			All		
		F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
Mixtral-8x7B	L1E2	6.5	96.5	68.0	4.9	84.5	65.9	3.1	77.5	59.7	14.5	66.9	52.7	14.8	54.3	38.3	13.4	40.5	26.0	57.2	63.7	45.9
	L1E3*	4.1	87.2	34.4	2.9	77.8	35.1	1.8	78.1	32.3	9.6	69.1	29.0	10.6	45.7	21.3	9.7	35.3	20.7	38.7	54.9	26.0
	L1E0	6.2	97.1	69.8	4.7	84.6	66.9	2.8	77.4	62.5	14.0	68.5	54.9	14.0	56.3	40.6	11.8	43.9	31.1	53.5	65.4	49.1
	L1E4	6.0	96.9	71.5	4.6	83.4	66.0	2.6	76.8	64.7	13.5	68.3	53.8	13.1	54.1	38.7	11.4	43.7	31.4	51.1	64.8	48.7
	L1E6	6.1	95.8	67.8	5.0	84.7	65.5	3.0	79.3	57.5	14.8	68.0	51.0	14.9	53.0	36.8	13.5	42.5	26.7	57.4	64.0	45.0
	L1E0+L1E4+L1E6	15.6	96.8	65.0	10.2	79.6	61.0	5.6	70.7	56.9	21.5	58.0	49.8	15.6	41.1	35.4	10.8	30.5	27.4	79.3	60.1	48.9
Phi-3.5-MoE	L3E7*	1.2	95.7	81.1	1.8	87.3	73.9	2.4	84.1	64.4	10.9	72.1	48.1	19.2	53.3	27.4	13.4	38.2	19.5	48.9	59.8	34.7
	L5E10*	43.0	81.0	4.9	9.2	30.6	9.9	5.4	35.1	19.2	10.5	22.5	30.2	9.2	17.5	30.6	4.3	12.8	26.2	81.6	23.5	13.7
	L10E0*	1.5	83.4	34.1	0.9	75.2	40.7	1.0	69.3	35.5	6.2	64.7	31.5	11.8	50.4	24.5	3.4	47.7	23.7	24.8	57.2	27.8
	L3E7*+L5E10*+L10E0*	1.2	95.7	81.1	1.8	87.3	73.9	42.1	81.0	6.9	23.7	29.9	21.5	12.1	18.6	30.4	5.6	12.3	24.7	86.4	28.3	17.7
	L1E1	3.3	98.3	81.0	2.5	88.1	72.8	2.0	80.5	62.9	10.6	69.9	45.8	22.1	52.1	26.0	17.3	34.7	16.8	57.8	58.9	33.3
	L1E9	1.5	98.4	85.6	1.2	91.1	78.2	1.1	85.8	67.4	6.8	75.4	49.6	16.6	55.9	23.0	15.9	40.8	14.3	43.1	61.9	28.9
	L1E13	2.1	97.9	86.2	2.2	90.8	78.2	1.8	86.7	68.8	8.6	69.8	47.4	20.0	53.5	24.4	17.3	37.6	16.4	51.9	59.8	31.8
	L1E1+L1E9+L1E13	5.7	98.4	82.9	4.5	89.0	74.8	3.5	82.0	65.2	14.7	69.1	50.5	24.1	49.4	29.7	16.8	32.3	20.2	69.2	60.2	40.9
OLMoE	L1E7	13.9	96.1	90.8	11.7	88.0	77.2	7.3	78.1	70.9	19.7	65.5	53.2	10.8	51.6	37.1	-	-	-	63.4	76.8	65.2
	L1E9	21.5	92.0	48.8	10.9	68.2	49.5	7.2	58.2	46.4	19.2	45.7	38.3	9.5	34.5	30.6	-	-	-	68.2	59.5	43.2
	L2E30	7.2	97.5	86.5	4.2	87.6	74.3	3.5	83.9	70.0	12.6	68.3	50.1	7.8	54.7	34.8	-	-	-	35.4	76.9	59.0
	L4E14	15.7	55.2	75.1	12.2	53.1	52.7	9.0	43.5	40.1	21.1	35.5	27.0	6.5	27.1	22.3	-	-	-	64.5	45.9	44.9
	L1E9+L2E30+L4E14	20.5	92.0	48.8	13.5	74.0	59.7	8.0	62.9	53.5	32.5	39.0	44.8	9.6	22.9	24.2	-	-	-	84.2	51.5	46.6
	L1E20	10.3	97.1	90.9	7.6	86.8	75.6	6.2	80.6	69.3	18.1	67.0	51.0	11.4	52.2	35.9	-	-	-	53.5	75.7	61.1
	L1E33	3.6	96.1	90.9	2.8	85.4	73.9	2.6	80.2	63.1	10.2	69.0	44.4	8.0	55.6	30.9	-	-	-	27.3	74.1	51.4
	L1E51	5.9	96.4	92.6	5.0	87.5	78.9	4.4	80.1	69.0	15.1	69.1	50.4	10.9	56.7	37.0	-	-	-	41.3	75.1	58.3
	L1E20+L1E33+L1E51	15.0	96.3	90.9	10.2	87.3	77.1	7.8	78.7	69.7	21.5	66.3	54.0	11.8	50.6	39.7	-	-	-	66.3	76.0	65.2
Qwen3-30B-A3B	L1E15	2.3	97.2	94.4	2.0	91.8	86.2	2.2	86.9	76.1	10.1	78.5	54.3	18.2	58.0	29.1	31.5	32.3	13.3	66.3	58.5	31.0
	L2E92*	0.0	100.0	95.2	0.0	96.4	93.1	0.0	85.0	85.0	0.2	81.7	55.6	0.7	70.7	23.0	2.5	54.3	8.6	3.6	68.2	17.2
	L1E68*	3.1	96.7	95.9	2.9	95.3	89.4	2.9	88.9	81.4	12.0	76.0	56.2	18.1	55.9	30.6	26.6	31.4	15.2	65.6	60.5	37.0
	L3E82*	1.0	96.9	90.9	0.9	91.8	83.8	0.9	88.2	71.9	3.8	76.4	52.2	7.5	58.4	29.4	12.2	34.6	14.5	26.4	59.2	31.4
	L21E69*	8.1	93.0	51.6	8.0	81.7	31.8	7.8	75.4	28.0	22.3	43.5	29.2	20.6	25.7	31.8	6.1	20.0	23.6	73.0	40.2	32.1
	L1E68*+L2E92*+L3E82*+L21E69*	3.1	96.7	95.9	3.0	95.3	89.4	3.6	90.5	83.2	14.0	77.9	59.7	19.9	56.7	33.7	43.5	30.5	29.6	87.1	49.4	42.0
	L1E5	2.3	93.8	79.8	2.1	82.9	66.4	2.2	76.6	55.6	9.8	63.4	43.0	14.7	46.9	24.5	26.2	31.9	13.6	57.4	51.5	27.6
	L1E23	5.4	97.0	89.3	4.1	91.3	83.5	3.7	84.2	74.2	14.0	73.7	55.6	20.4	55.2	31.5	28.4	29.4	15.1	75.9	60.4	38.8
	L1E77	4.9	95.4	87.4	4.3	90.5	82.2	4.2	82.6	73.4	16.0	72.1	56.1	21.2	53.2	32.0	27.6	29.3	16.7	78.2	59.0	40.0
	L1E5+L1E23+L1E77	9.2	96.0	87.4	7.0	88.9	81.5	6.3	82.1	71.8	20.7	69.7	57.4	21.9	49.9	36.1	22.3	26.6	20.1	87.4	60.7	48.6

L ROUTING LEVERAGE AGAINST ITS COMPONENTS, BY HOP

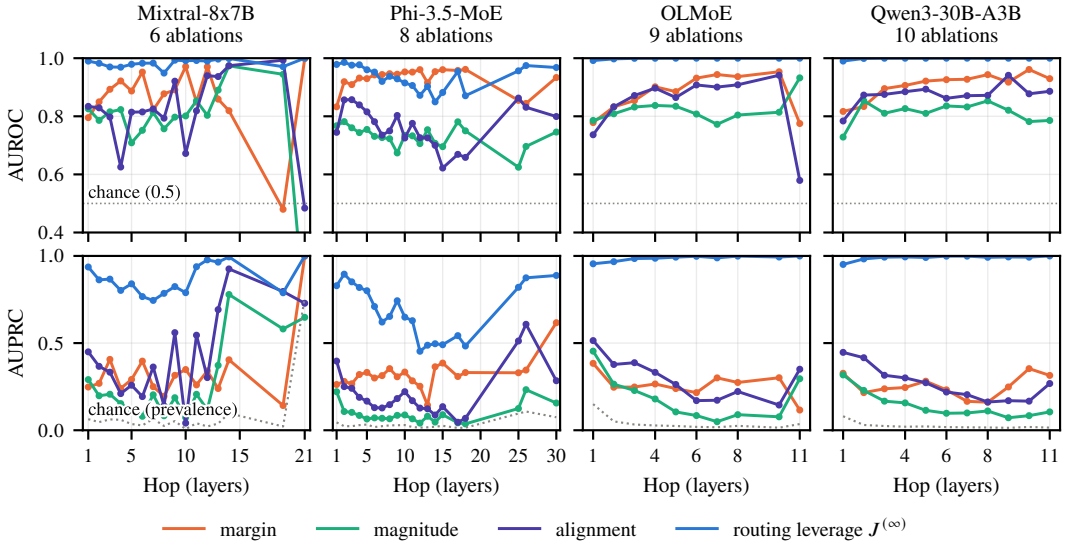


Figure 10: Same as Figure 4, inside the clean-path regime.

Table 16: AUROC by hop band.

Model	Band	Prev.	Margin	Magnitude	Alignment	Normalized alignment	$J^{(\infty)}$
Mixtral-8x7B	hop 1	6.4%	79.5%	82.6%	83.4%	88.9%	**99.0%**
	hops 2–4	4.0%	89.4%	74.3%	81.2%	87.6%	**96.8%**
	hops 5+	2.6%	**94.2%**	65.0%	72.8%	78.5%	89.4%
Phi-3.5-MoE	hop 1	4.5%	83.3%	76.7%	74.5%	82.1%	**97.8%**
	hops 2–4	3.2%	91.0%	69.2%	76.7%	86.7%	**94.6%**
	hops 5+	3.3%	**94.2%**	62.2%	64.3%	68.6%	84.1%
OLMoE	hop 1	15.0%	77.8%	78.6%	73.6%	78.2%	**99.2%**
	hops 2–4	10.6%	85.5%	68.9%	78.0%	83.5%	**95.8%**
	hops 5+	5.9%	**91.0%**	65.2%	72.1%	76.8%	89.5%
Qwen3-30B-A3B	hop 1	8.0%	81.7%	72.8%	78.4%	84.4%	**99.0%**
	hops 2–4	8.4%	87.1%	66.1%	79.5%	84.6%	**96.7%**
	hops 5+	5.0%	**92.0%**	60.6%	66.4%	70.0%	85.4%

Table 17: AUROC by hop band, inside the clean-path regime.

Model	Band	Prev.	Margin	Magnitude	Alignment	Normalized alignment	$J^{(\infty)}$
Mixtral-8x7B	hop 1	6.4%	79.5%	82.6%	83.4%	88.9%	**99.0%**
	hops 2–4	5.3%	87.2%	79.9%	78.5%	83.5%	**97.8%**
	hops 5+	5.5%	89.2%	75.7%	81.3%	91.9%	**98.3%**
Phi-3.5-MoE	hop 1	4.5%	83.3%	76.7%	74.5%	82.1%	**97.8%**
	hops 2–4	2.6%	91.9%	76.8%	85.2%	92.0%	**98.1%**
	hops 5+	2.6%	**94.3%**	73.1%	77.1%	82.9%	93.7%
OLMoE	hop 1	15.0%	77.8%	78.6%	73.6%	78.2%	**99.2%**
	hops 2–4	4.3%	84.9%	81.8%	85.2%	90.1%	**99.7%**
	hops 5+	2.2%	91.5%	81.1%	89.4%	95.4%	**100.0%**
Qwen3-30B-A3B	hop 1	8.0%	81.7%	72.8%	78.4%	84.4%	**99.0%**
	hops 2–4	2.8%	86.2%	84.1%	87.6%	92.8%	**99.9%**
	hops 5+	1.8%	93.3%	81.2%	88.8%	96.6%	**100.0%**

Section 6.3 compares the routing leverage against its individual components. Figure 10 repeats that comparison inside the clean-path regime of Section 6.1: a token is scored only while no earlier token of its sequence has diverged. The margin uses no predicted perturbation and so does not depend on

that condition, whereas the three estimators do. Fewer tokens remain at the deeper hops, where the curves fluctuate accordingly.

Tables 16 and 17 report the comparison numerically, without and with that restriction, grouped into the hop distances past the ablation. Entries are AUROC, averaged within an ablation over the hops of a band with weights given by the number of divergences at each hop, then taken as the median across that model's ablations. *Prev.* is the share of tokens diverging in that band. *Normalized alignment* divides alignment by the margin of the same boundary, so it differs from alignment only by the normalization and from $J^{(\infty)}$ only by the search for the most susceptible boundary.

M SCOPE

All interventions and routing leverage measurements in this paper are performed on C4. We study an internal routing mechanism rather than downstream task performance: how an expert's contribution propagates to downstream routers and whether it crosses a Top-K boundary. We evaluate this mechanism across five architecturally diverse MoE models that differ in normalizer, attention variant, routing mechanism, expert count and Top-K (Appendix B). While evaluating additional domains would further assess the generality of the reported quantities, we do not expect it to alter the underlying routing mechanism studied here.