Path-Space: Estimating Transformer Component Effects and Their Interactions from Shared Backward Passes

Anonymous authors
Paper under double-blind review

Abstract

Which components of a transformer matter for a prediction, and which matter only in combination? Ablation answers this at one forward pass per component and one per pair, and the cheap readings used instead, the logit lens and direct-logit attribution, ignore everything downstream of the component. We present Path-Space, a method that estimates component effects and their pairwise interactions from shared backward passes. It reads each block's write through the exact per-context transport to the output, as attribution patching does, at every block and position from one backward pass; it obtains a removal's exact effect as the line integral of the gradient along the removal, evaluated by Gauss–Legendre quadrature; and it introduces second-order attribution patching, which predicts a component's interaction with every other component from one backward pass on the run with that component removed, with an exact error identity and a symmetrised form. On five models from 0.36B to 70B parameters, scored against measured removals on the model's own next-token log-probability, the per-context reading predicts removal effects at Spearman 0.69–0.82, against 0.20–0.29 for the identity reading behind the logit lens; six quadrature nodes reproduce the effects to a median error below 2×10^{-4} at every block but the first; and the symmetrised interaction estimate reaches 0.65–0.76 up to 8B and 0.56 at 70B in bf16, recovering strong interactions between individually weak components that rankings by single effects miss. Its setup costs about $4L$ forward passes against $L^2/2$ for all pairs: at 80 blocks it beats measuring pairs in the order of their single effects at equal cost from a fifth of all pairs on, a comparison we did not pre-register, while at 24–32 blocks it does not yet pay for itself.

1 Introduction

When a transformer produces a token, which of its blocks built the prediction, and at which positions? Tools that read the residual stream disagree about how a block's write reaches the output. The logit lens (nostalgebraist, 2020) and direct-logit attribution (Elhage et al., 2021) read a write Δ_k as if the remaining blocks were the identity, $\tilde{g} \cdot \Delta_k$ with $\tilde{g}$ the answer's unembedding direction; the tuned lens (Belrose et al., 2023) and the Jacobian lens (Gurnee et al., 2026) read it through a transport learned or averaged over a corpus; attribution patching (Nanda, 2023; Syed et al., 2024; Kramár et al., 2024) reads it through the gradient at that block, per context. These approximate one object, the map from block k's output to the final logit, and can be checked against one fact: what the logit does when the write is removed.

We present Path-Space, a method that estimates component effects and their pairwise interactions from shared backward passes, and check it and these readings against removal on the model's own output. The transport from block k is a product of block Jacobians; a product of $(I + A)$ factors is a sum over paths through depth (Elhage et al., 2021; Veit et al., 2016) whose first-order classes are differences of consecutive gradients, which one backward pass yields at every block and position; these are Path-Space's coordinates. What they make cheap is not a single component's effect, which attribution patching already estimates, but

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.

the interaction of two, $I(k, j) = \delta_{j|k} - \delta_j$: the change in j's effect once k is removed, of which redundancy and self-repair are made (Gong et al., 2026).

We report five things. (1) Scored on the model's own log-probability with each target's own gradient, the per-context transport predicts what removing a write does at Spearman 0.69–0.82 against 0.20–0.29 for the identity in five models from 0.36B to 70B; on the frozen-norm logit the gap is +0.37 to +0.59, widest at 70B (§4.1). (2) The identity's sign for early writes is within two to six points of its shuffle null in every model (§4.2). (3) The removal effect is exactly a line integral; six Gauss–Legendre nodes reproduce it on the log-probability, normalisation included, at every block but block 0 (§4.4). (4) One backward pass on the run with component k removed predicts k's interaction with every other component; averaged over its two directions it orders measured interactions at Spearman 0.65–0.76 up to 8B and 0.56 at 70B. We give its exact error identity, the condition under which its errors cancel, its cost at a fixed budget (at 80 blocks it beats single-effect ordering at equal cost) and the interactions only it finds (§4.5, §4.6); it transfers to head granularity and to AtP*'s setting (§4.7). (5) Nine pre-registered attempts to turn the object into capability close negative or as trades, for the methods, tasks and budgets tried (§5). Every experiment was pre-registered with pass/fail criteria and null controls; Appendices L and M report what failed and which artefacts the nulls caught.

2 Setting

Models and data. Five models in four families: Qwen3-0.6B (Qwen Team, 2025) (28 blocks, $d = 1024$, RMSNorm, tied embeddings), pythia-410m (Biderman et al., 2023) (24 blocks, parallel attention/MLP, LayerNorm, untied head, rotary positions) and SmolLM2-360M (Ben Allal et al., 2025) (32 blocks, Llama-style) in fp32, and Meta-Llama-3-8B and -70B (Grattafiori et al., 2024) (32 and 80 blocks, $d = 4096$, 8192) in fp32 on one A800-80GB and bf16 on two. Llama-3 enters the headline, sign, lattice, mediation and second-order analyses; lens baselines, head-level IOI and capability experiments use the small models. Evaluation positions come from FineWeb (Penedo et al., 2024) sequences of up to 128 tokens of one corpus (seed 1), 150 per model and 120 at 70B unless stated; averaged transports are fitted on disjoint corpora (seeds 0, 2).

The logit with a frozen norm. Let $h_k[q] \in \mathbb{R}^d$ be the residual stream after block k at position q, h_{-1} the embedding output, and $\Delta_k[q] = h_k[q] - h_{k-1}[q]$ block k's write. For a target position p and token t define

$$\tilde{z} \;=\; \tilde{g} \cdot h_{L-1}[p], \qquad \tilde{g} = \frac{W_U[t] \odot \gamma}{\mathrm{rms}(h_{L-1}[p])}, \tag{1}$$

with the final RMSNorm's scale rms(·) frozen at its clean-run value (for LayerNorm, the standard deviation is frozen and the mean subtraction kept). $\tilde{z}$ is then linear in h_{L-1}, which turns the response decomposition and a sum rule with a computable remainder into identities. It is a choice of target, not a requirement: under the model's own norm, attributions sum to zero by zero-homogeneity (Appendix M) yet can each be informative.

What the frozen norm leaves out. Removing a write also changes the norm's own scale, which rescales every logit; this common-mode term is as large as the write effects, which is why $\Delta\tilde{z}$ and the model's own Δz correlate at only 0.59–0.62 (Appendix F).

Targets. We score four targets at the position p: the frozen-norm logit $\tilde{z}$, for which the decompositions of §3 are identities; the model's own logit z_t; its log-probability $\log p_t$, the primary target; and the logit difference $z_t - z_c$ to the clean run's top competitor c. Each target m has its own per-context predictor $\hat{\delta}_k^m = -\nabla_{h_k[p]} m \cdot \Delta_k[p]$ and its own identity reading, the same product with the gradient taken at the last residual.

Three readings of a write. With $\nabla_k[q] = \partial\tilde{z}/\partial h_k[q]$, removing $\Delta_k[q]$ is predicted as

$$\text{identity (DLA, logit lens):} \quad -\,\tilde{g}\cdot\Delta_k[q], \tag{2}$$

$$\text{averaged transport (lens):} \quad -\,\big(\bar{J}_k^{\top}\tilde{g}\big)\cdot\Delta_k[q], \tag{3}$$

$$\text{per-context transport:} \quad -\,\nabla_k[q]\cdot\Delta_k[q], \tag{4}$$

where $\bar{J}_k$ is the Jacobian lens of Gurnee et al. (2026) fitted on 32 FineWeb sequences (the corpus-averaged, future-position-summed transport), and $\nabla_k[q] = J_k^{(\mathrm{ctx})\top}\tilde{g}$ is the exact per-context transport applied to the output direction. Eq. (4) is attribution patching at block-write granularity (Nanda, 2023); we also fit a least-squares affine transport $h_{L-1} \approx W_k h_k + b_k$ on 25 600 positions of the independent corpus. The tuned lens, trained on the output distribution, is reported under its own objective in §4.1.

The arbiter. The effect of removing block k's write at position q is

$$\delta_k[q] \;=\; \tilde{z}\big(h_k[q] \leftarrow h_{k-1}[q]\big) - \tilde{z}, \tag{5}$$

measured by a forward pass in which block k's output at position q is replaced by its input, later positions seeing the ablated stream. Predictors are scored by Spearman correlation with δ over (position, block) cells, always against a within-position block-shuffle null and with bootstrap intervals over positions.

3 Path-Space

Path-Space is the coordinate system over routes that the depth chain induces: for a target (position p, token t), the first-order path classes and attributions of every block at every position, computed exactly per context, together with the interference, curvature and interaction structure built on them (Fig. 1). This section defines each piece and states what is literal in the borrowed vocabulary and what is not.

3.1 The depth chain as a matrix-product operator, contracted by one backward pass

Write $B_k = \partial h_k/\partial h_{k-1} \in \mathbb{R}^{(Td)\times(Td)}$ for the Jacobian of block k on the (position, channel) space of the whole residual stream, at the clean activations. The transport from block k's output to the last residual is the ordered product

$$\begin{aligned} J_k \;&=\; B_{L-1}\,B_{L-2}\cdots B_{k+1}, \\ [J_k]_{(p,b),(q,a)} \;&=\; \sum_{(q_1,a_1),\ldots} [B_{L-1}]_{(p,b),(q_1,a_1)}[B_{L-2}]_{(q_1,a_1),(q_2,a_2)}\cdots[B_{k+1}]_{(\cdot),(q,a)}, \end{aligned} \tag{6}$$

a matrix-product operator in the tensor-network sense (Orús, 2014): one site per block, physical index (position, channel) of dimension Td, and the bond between consecutive sites summed over the intermediate residual stream. Forming any B_k costs $(Td)^2$ and is never needed. What every result in this paper uses is the contraction of (6) with a single covector at the output end, $\tilde{g}$ (nonzero at position p only):

$$\nabla_k[q] \;=\; \big(J_k^{\top}\tilde{g}\big)[q] \;=\; \frac{\partial\tilde{z}}{\partial h_k[q]} \qquad \text{for all } k \text{ and } q, \tag{7}$$

and this is exactly what reverse-mode automatic differentiation computes: the backward pass applies $B_{L-1}^{\top}$, then $B_{L-2}^{\top}$, and so on to the covector, site by site from the output end, at the cost of one forward pass. The identification is literal, not an analogy: the backward pass is the MPO contraction. One backward pass rooted at the embedding output therefore returns the entire family $\{\nabla_k[q]\}$, $L\times T$ vectors of dimension d, which is the raw material of everything below. With the final norm's scale frozen (§2) $\tilde{z}$ is linear in h_{L-1}, so the covector is exact rather than a linearisation of the readout.

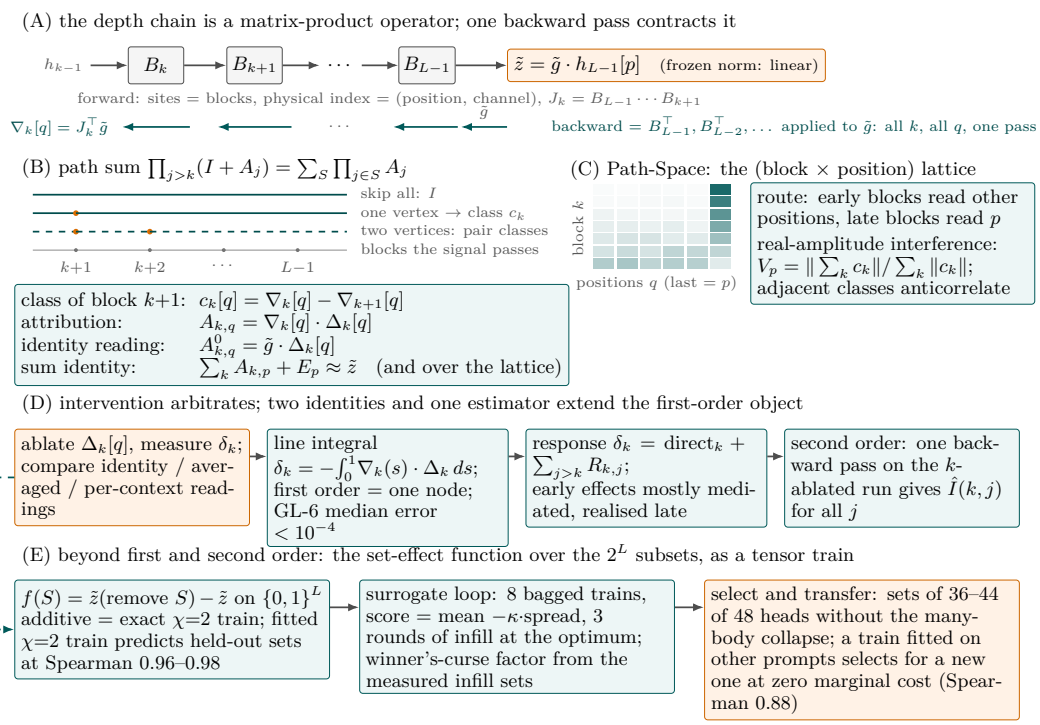


Figure 1: Path-Space. (A) The transport from block k to the output is a matrix-product operator over blocks; contracting it with the output covector $\tilde{g}$ is one backward pass, which returns $\nabla_k[q] = J_k^\top \tilde{g}$ at every block and position. (B) Expanding $\prod(I + A_j)$ gives a sum over paths; grouping them by their lowest vertex gives the first-order classes c_k, differences of consecutive gradients, and two-vertex classes of the clean chain motivate, but are not, the ablation interaction that the second-order estimator $\hat{I}$ targets (§3.2). (C) On the (block × position) lattice the attributions are the route of the prediction. (D) Ablation arbitrates between readings; two identities and one estimator extend the first-order object. (E) The effect of removing a set of components is a function on the 2^L subsets, modelled by a tensor train.

3.2 The path sum and its classes

Write $B_j = I + A_j$, the skip connection plus the block's own action; expanding the product,

$$J_k \;=\; \prod_{j>k} (I + A_j) \;=\; \sum_{S \subseteq \{k+1,\ldots,L-1\}} \;\prod_{j \in S} A_j, \tag{8}$$

a sum over 2^{L-k-1} terms, each a path: the signal leaving block k skips every block not in S and interacts with every block in S. In the diagrammatic reading (Fig. 1B) a path is a line from block k to the output with a vertex at each element of S; the order of a term is its number of vertices, $|S|$ (Elhage et al., 2021; Veit et al., 2016; Feynman, 1948). The sum is exact for the linearisation: combinatorics, not a path integral with a measure.

Two resummations of (8) are computable from (7) alone. Grouping paths by their lowest vertex gives the first-order classes: all paths whose first interaction after leaving block k is at block $k+1$ sum to $J_{k+1}A_{k+1} = J_k - J_{k+1}$, and applied to the output covector,

$$c_k[q] \;=\; \nabla_k[q] - \nabla_{k+1}[q], \qquad \sum_k c_k[q] \;=\; \nabla_{-1}[q] - \nabla_{L-1}[q]. \tag{9}$$

$c_k[q]$ is the sensitivity of the logit to what block $k{+}1$ writes at position q, transported by everything after it; it needs no per-block Jacobian and no inversion. The second-order estimator of §4.5 is suggested by this picture but is not a class of it: ablating block k changes the later blocks' activations, so its definition is an intervention.

3.3 The lattice

Multiplying each class by the write it acts on gives the first-order attribution of block k's write at position q to the logit at p,

$$A_{k,q} \;=\; \nabla_k[q] \cdot \Delta_k[q], \qquad E_q \;=\; \nabla_{\text{emb}}[q] \cdot h_{-1}[q], \tag{10}$$

and the collection $\{A_{k,q}\}$ over blocks and positions is the route of the prediction on the (block $\times$ position) lattice (Fig. 1C). Its identity-reading counterpart replaces the transport by I: $A^0_{k,q} = \tilde{g} \cdot \Delta_k[q]$, nonzero only at $q = p$.

The identity reading sums exactly, $\sum_k A^0_{k,p} + \tilde{g} \cdot h_{-1}[p] + b = \tilde{z}$, by telescoping. The transported sum does not: it differs by a remainder computable from the same backward pass, which vanishes only when every transport acts as the identity on the write it carries (Appendix B gives a two-block counterexample). On Qwen3-0.6B the remainder formula is exact to 3×10^{-7} and the sum rule holds in aggregate but not pointwise: median ratio 0.999 at the target position and 1.017 over the lattice, while $|R/\tilde{z}|$ has median 0.057 and maximum 1.04. We use the sum only as an aggregate check.

Interference. The classes are real vectors whose sum is the transport, so their coherent and incoherent sums differ. On Qwen3-0.6B over 400 contexts, adjacent classes anticorrelate in every context (cosine -0.10 to -0.12 against a layer-permutation null of -0.02, $z = -152$ to -176) and the coherence ratio is 0.14 against the incoherent 0.19: the residual stream overshoots and corrects at adjacent blocks (Appendix G; two attempts to exploit it failed).

3.4 The arbiter, two identities and one estimator

Every reading of a write is scored against the same intervention: remove the write and measure the logit (§2). Two identities and one estimator extend the first-order object and are measured in §4.4–4.5: the line integral holds for any differentiable target and the response decomposition for $\tilde{z}$; second-order attribution patching is an estimator whose reliability is an empirical question. The line integral: along the ablation path $h_k[p] \to h_k[p] - s\Delta_k[p]$, $\delta_k = -\int_0^1 \nabla_k(s) \cdot \Delta_k \, ds$ exactly; the first-order attribution is this integral with one node at $s = 0$, and the quadrature ladder (secant, trapezoid, Gauss–Legendre) measures the curvature the first order omits (Sundararajan et al., 2017). The response decomposition: $\delta_k = -\tilde{g}\cdot\Delta_k + \sum_{j>k} \tilde{g}\cdot(\Delta'_j - \Delta_j)$, the removed write's own output projection plus every later block's changed write read at the output, exact because $\tilde{z}$ is linear in h_{L-1} (Vig et al., 2020). Second-order attribution patching (§4.5): one backward pass on the run with block k's write removed predicts k's interaction with every other block. The set-effect tensor: the effect of removing a set S of components, $f(S) = \tilde{z}(\text{remove } S) - \tilde{z}$, is a function on $\{0,1\}^L$, a tensor with L binary sites; a tensor train fitted to sampled ablations (Oseledets, 2011) measures how compressible it is along the chosen ordering (rank is not interaction order: $\prod_i s_i$ has rank one) and turns subset selection into an optimisation on the hypercube (Appendix K). Algorithm 1 (Appendix Q) collects the first-order and second-order computation.

The tensor-network and path-integral vocabulary is used in a specific, checkable sense; Appendix N states which parts are literal and which are analogy.

4 Results

4.1 Which reading predicts what a removal does to the model's output

Table 1 scores the readings on the model's own next-token log-probability. With the gradient of that target, the per-context transport predicts the change caused by removing a write at 0.69–0.82 against 0.20–0.29 for the identity in all five models, both where the model's top prediction is the target (38–53% of positions) and where it is not, and on in-context number comparison, a task on which the models' top prediction is right on only 20–76% of prompts (0.73–0.76 against 0.18–0.27, Appendix D). Predicting the same target from the frozen-norm gradient drops the per-context score to 0.25–0.42: an attribution scored against

Table 1: Spearman between prediction and measured change over all (position, block) cells, per-context / identity. Left: the model's own next-token log-probability, each prediction from the gradient of that target. Middle: the same target predicted from the frozen-norm gradient, a mismatched target. Right: the frozen-norm logit. Document-clustered 95% intervals on the left-column gap: $[+0.41, +0.46]$, $[+0.40, +0.49]$, $[+0.48, +0.54]$, $[+0.46, +0.53]$, $[+0.57, +0.63]$; 70B in bf16, readout in fp32 (Appendix E).

model	$\Delta \log p_t$, own gradient	$\Delta \log p_t$, $\tilde{z}$-gradient	$\Delta \tilde{z}$
Qwen3-0.6B	**0.724** / 0.290	0.422 / 0.151	0.808 / 0.437
pythia-410m	**0.724** / 0.277	0.361 / 0.214	0.787 / 0.385
SmolLM2-360M	**0.757** / 0.250	0.343 / 0.190	0.839 / 0.384
Llama-3-8B	**0.693** / 0.201	0.252 / 0.172	0.718 / 0.350
Llama-3-70B	**0.815** / 0.214	0.259 / 0.180	0.829 / 0.237

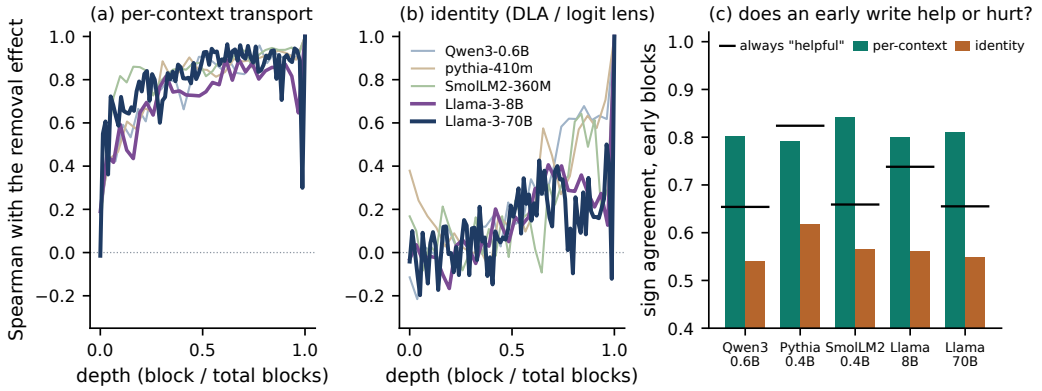


Figure 2: Five models from 0.36B to 70B parameters. (a, b) Spearman between each reading and the measured effect of removing block k's write, by relative depth; the per-context transport is informative from the second block in every model, while the identity is uninformative through the first half of the stack and, at 70B, until the last block. (c) Sign agreement of early writes with the removal effect; in every model the identity scores below always answering "helpful" (black ticks).

a target other than the one it differentiates looks weak because of the mismatch, not because of the transport. On $\tilde{z}$ the gap is +0.37 to +0.46 up to 8B and +0.59 at 70B, the widest among the five models tested (Fig. 2): the per-context transport is informative from the second block in every model, while the identity stays within ± 0.23 of zero through the first half of the stack and at 70B reaches at most 0.50 before the last block; an 8B run in bf16 matches fp32 to within 0.013, so the bf16 70B run is comparable (Table 15).

The averaged readings do not close the gap at this task. On the small models the Jacobian lens used as a transport scores 0.530, a tuned lens trained under its own objective (Belrose et al., 2023) 0.455 and 0.406, and a least-squares affine transport 0.486, all uninformative through the first third of the stack. Fitted on 8 to 128 disjoint contexts, an averaged transport stays at 0.50–0.55 with or without summing future positions (Appendix C): the loss is context averaging, not estimation. The comparison asks one question, the finite effect of removing a write in the current context; it does not test what lenses were built for.

4.2 The sign of early writes

A common use of the logit lens is a depth-wise story: early blocks push toward the wrong token, later blocks correct it. Figure 2c and Table 19 (Appendix Q) test the sign that story rests on. On the early blocks of all five models the identity's sign agrees with the measured effect 54–62% of the time, within two points of its within-block shuffle null in every model (+0.003 at 8B, +0.005 at 70B) and below the 65–82% that always answering "helpful" would score; the per-context transport's sign agrees 79–84%. Of the early writes the identity calls harmful, 63–80% help, the base rate (permutation $p = 0.21$–0.42 at 0.6B, 8B and 70B). On

the model's own log-probability, with own gradients, the identity's early sign is 0.52–0.56 against nulls of 0.50 and the per-context transport's 0.78–0.84 in the small models, and 0.50 and 0.51 against 0.78 and 0.81 at 8B and 70B; pythia's identity is 5.9 points above its null, over the five we pre-registered. This is mediation (Appendix I): an early write's direct projection is a small part of its effect, and the rest need not share its sign.

4.3 The lattice, mediation, and what the five models share

In all five models 80–90% of the early-block attribution magnitude sits at positions other than the target and nets to 2–7% of the logit (Appendix H), and in all but pythia removing an early write changes the logit 72–100% through later blocks' writes, not directly (Appendix I); which positions the route reads and how much it defers are architectural (Appendix J).

4.4 Curvature: first order is a one-node quadrature

Why does the first-order transport order early effects correctly (Fig. 8a) yet miss their size? Along the ablation path $h_k[p] \to h_k[p] - s\Delta_k[p]$, $s \in [0,1]$, the effect is exactly

$$\delta_k \;=\; -\int_0^1 \nabla_k(s) \cdot \Delta_k \, ds, \tag{11}$$

and the first-order attribution is this integral evaluated with one node at $s = 0$. Fig. 9a (Appendix Q) compares four quadratures of (11) on 40 positions. Pooled over blocks 1–6, the median relative error is 0.69 for first order, 0.47 for a local second-order term formed from one extra gradient at $s = 0.1$, 0.38 for the trapezoid rule using the gradient at the ablated point, and below 10^{-4} for six Gauss–Legendre nodes over 1 003 cells at blocks 1–27 (99th percentile 0.015, worst case 0.18 at block 1); on Llama-3-8B in fp32 the ladder repeats (first order 0.40, trapezoid 0.12) and the six-node median error is at most 0.003 at each of the 30 blocks after block 0 (Fig. 9c). The line integral is an identity, so the first-order error is curvature; that six nodes suffice is an empirical finding about these paths, not a consequence of smoothness. This is integrated gradients (Sundararajan et al., 2017) along the ablation path, block by block. The identity holds for any differentiable target, and on the model's own log-probability, whose path passes through the norm and the softmax, six nodes reproduce the removal effect to a median error of 3×10^{-5} to 2×10^{-4} in the three small models (90th percentile at most 1.1×10^{-3}, worst cases 0.08–0.25 over 351–441 cells), so the normalisation effect a frozen norm leaves out is inside the integral.

Block 0 is the exception in every model (six-node error 1.25 at 8B): removing its write leaves the raw embedding in the stream and reorganises the downstream gradient 204–411×, so it is better read as the second half of the embedding than the first link of the chain. The same curvature governs a continuous edit dose (Fig. 9b).

4.5 Second-order attribution patching

The interaction of removing two writes, $I(k,j) = \delta_{k,j} - \delta_k - \delta_j = \delta_{j|k} - \delta_j$, is what the additive assumption behind attribution patching sets to zero; measuring it needs $L^2/2$ pair ablations. The first-order attribution of block j changes when block k's write is removed, and that change is a prediction of the interaction,

$$\hat{I}(k,j) \;=\; \big[-\nabla_j^{(k)} \cdot \Delta_j^{(k)}\big] - \big[-\nabla_j \cdot \Delta_j\big], \tag{12}$$

where (k) marks the run with block k's write removed; one backward pass on that run yields $\hat{I}(k,j)$ for every j. With $e_j = \delta_j + \nabla_j \cdot \Delta_j$ and $e_{j|k} = \delta_{j|k} + \nabla_j^{(k)} \cdot \Delta_j^{(k)}$ the first-order errors of the two marginal effects, $\hat{I} - I = e_j - e_{j|k}$ exactly: the estimate is as good as the two errors are alike. The run with the later block removed, read at the earlier one, gives a second estimate $\hat{I}(j \to k) = -(\nabla_k^{(j)} - \nabla_k) \cdot \Delta_k$; I is symmetric and we average the two.

On the model's own log-probability (Table 2) the forward estimate orders the measured interactions at Spearman 0.62–0.72 up to 8B, the reverse at 0.48–0.62 and their average at 0.65–0.76; the two directions disagree in sign on 26–32% of pairs, so the estimator is not

Table 2: Second-order attribution patching on the model's own log-probability, all $L(L-1)/2$ pairs at 20 positions per model (10 for Llama-3-8B; †70B: 400 sampled pairs at 8 positions, bf16, Appendix E). Spearman with the measured interaction for each direction and their average (shuffle nulls ≤ 0.012); median pair-effect error given the two measured singles, alone and with the symmetrised $\hat{I}$; correlation of the two first-order errors whose difference is the estimator's error, over the third of pairs with the smallest and the largest conditioning removal $|\delta_k|$; and, for strong interactions whose singles are both weak, the share found in the first 10% of pairs ranked by $|\hat{I}|$ and by $|\delta_k \delta_j|$.

model	pairs	Spearman fwd / rev / sym	pair error: singles → $+\hat{I}$	corr($e_{j\|k}, e_j$): small / large $\|\delta_k\|$	weak singles found: $\|\hat{I}\|$ / $\|\delta_k\delta_j\|$
Qwen3-0.6B	7 560	0.718 / 0.572 / **0.759**	0.088 → 0.042	0.94 / 0.33	0.42 / 0.00
pythia-410m	5 520	0.687 / 0.566 / **0.729**	0.103 → 0.050	0.92 / 0.26	0.55 / 0.00
SmolLM2-360M	9 920	0.722 / 0.618 / **0.760**	0.040 → 0.016	0.92 / 0.59	0.43 / 0.00
Llama-3-8B	4 960	0.619 / 0.480 / **0.651**	0.021 → 0.010	0.60 / 0.38	0.12 / 0.00
Llama-3-70B†	3 200	0.532 / 0.480 / **0.557**	0.009 → 0.007	1.00 / 0.97	0.53 / 0.00

symmetric and the average is the better use of the same passes. Given the two measured singles, the symmetrised estimate cuts the median pair-effect error by 50–60%. The error identity says when to trust it: in the small models, over the third of pairs with the smallest conditioning removal the two first-order errors correlate at 0.92–0.94 and nearly cancel, and over the third with the largest at 0.26–0.59, where the estimate degrades (Spearman 0.55–0.73) without failing; at 8B the largest third is again the worst (0.38) but the smallest cancels less than the middle one (0.60 against 0.94). On the frozen-norm logit the forward estimate scores 0.695 at 8B and 0.641 at 70B (9 920 and 4 800 pairs), cutting pair error by 57% and 30% (Table 16). At 70B, in bf16, the test is limited by precision: the measured interactions, not the estimates, carry the rounding (at 8B the bf16 estimate ranks fp32-measured ones at 0.64); with an fp32 readout the 70B estimate scores 0.557 (Appendix E).

4.6 Finding interactions at a fixed budget

The estimator's setup is the clean pass and one forward and backward pass per component, about $4L$ forward equivalents against $L^2/2$ for measuring every pair, a share of about $8/L$: 26–36% at 24–32 blocks, 10% at 80 and 1–2% among hundreds of heads. Per pair measured, ranking by $|\hat{I}|$ recovers 62–70% of the top-decile interactions after 10% of pairs, against 7–13% at random and 39–56% in the order of the measured singles $|\delta_k \delta_j|$. At equal total cost, setup included, at 24–32 blocks it only ties that ordering at half the pairs and trails it below; our pre-registered criterion (10% of pairs, setup counted) fails at every depth, the setup alone being 10.2% of the pairs at 80 blocks. From a budget of 20% of pairs on, at 70B with the fp32 readout, it beats that ordering at equal cost (0.65–0.93 against 0.60–0.82), a comparison we did not pre-register. What only the estimator does is find strong interactions between components whose single effects are both below the median: they are 4–10% of the top decile, and the $|\hat{I}|$ ranking finds 42–55% of them in its first 10% of pairs in the small models, 12% at 8B and 53% at 70B, where rankings by single effects find at most 4%.

4.7 Head granularity and AtP*'s footing

At head granularity on IOI (Wang et al., 2023), over 23 760–29 760 (prompt, component) cells in the small models, the per-context transport predicts the effect of zero-ablating a component at Spearman 0.883–0.953 against 0.12–0.15 for the identity, recovers 58–84% of each prompt's twenty most impactful components against 28–47%, and the second-order estimator predicts 1 800 sampled interactions among the top forty at 0.70–0.91 (Table 18, Fig. 11). In AtP*'s own activation-patching setting (Kramár et al., 2024) on pythia-410m, AtP ranks node effects at 0.950 and AtP* at 0.945, while AtP* cuts the nodes to verify for the true top ten from 58 to 26; path patching's direct effect ranks the 408 end-position nodes at 0.765 against 0.987 for AtP, a gap of 0.22, below the 0.3 we pre-registered; and the second-order estimator predicts interactions among the top forty nodes at 0.88 (Table 20).

5 Toward capability, and what did not work

Nine interventions built on the object were pre-registered and run (Appendices K and L); the negatives concern the methods, tasks and budgets tried, not the class. Pruning by per-component or pairwise scores collapses the task, because redundancy among many components is not pairwise; route-localised editing fails because the most sensitive cells are the most curved; a task-specialised head set raises the task margin at twice the general loss of random removal; and no probe we tried reads the per-token route from the residual. A learned model of set-removal effects (a tensor train, boosting or a small network; one model and task) transfers across prompts and avoids the many-body collapse of per-component rules.

6 Related work

The readings compared in §4.1 are the logit lens and direct-logit attribution (nostalgebraist, 2020; Elhage et al., 2021; Geva et al., 2022), the tuned and Jacobian lenses (Belrose et al., 2023; Gurnee et al., 2026) and attribution patching (Nanda, 2023; Syed et al., 2024; Kramár et al., 2024); that indirect effects dominate early is folklore, which we quantify on one footing by intervention. The path expansion (Elhage et al., 2021; Veit et al., 2016), integrated gradients (Sundararajan et al., 2017) and causal mediation (Vig et al., 2020) are the sources of (8), (11) and (16); path patching (Goldowsky-Dill et al., 2023; Wang et al., 2023) measures direct edge effects, and the hydra effect (McGrath et al., 2023) reports adjacent self-repair where we find the response realised at the readout end. Edge attribution patching and EAP-IG (Syed et al., 2024; Hanna et al., 2024) and the Hessian-vector-product correction of Zhang & Wang (2026) estimate single components' effects; conditional co-ablation (Gong et al., 2026) measures how much each remaining unit's effect grows once a set is removed, the conditional quantity that (12) predicts for every unit from one backward pass (Table 3). Tensor trains (Oseledets, 2011; Oseledets & Tyrtyshnikov, 2010) and surrogate optimisation (Jones et al., 1998; Breiman, 1996) are the tools of Appendix K; tensor-network (Orús, 2014) and path-sum (Feynman, 1948) readings guided the search; what survived is classical.

7 Limitations

Five models in four families cannot establish a trend with scale; "shared" means the same in the models tested. At 70B, run in bf16, measured interactions on log p sit near the rounding floor (Appendix E). Independent samples are fewer than cells: FineWeb positions come from 40 documents of 128 tokens (intervals are clustered by document), the head-level results from 30 IOI prompts per model and the AtP* comparison from ten prompt pairs. The first-order transport is right in order and wrong in size below block 18; the six-node quadrature is an empirical finding with a worst case of 0.25; the interaction estimator needs measured singles to improve pair predictions and does not pay for itself at equal cost at 24–32 blocks. Gradient×write (attribution patching) and its path integral (integrated gradients) are not claimed (Nanda, 2023; Sundararajan et al., 2017); Appendix O has the full list.

8 Conclusion

We presented Path-Space, which estimates transformer component effects and their pairwise interactions from shared backward passes. Read through the exact per-context transport, a block's write predicts the effect of its removal on the model's own log-probability at Spearman 0.69–0.82 in five models up to 70B, against 0.20–0.29 for the identity reading behind the logit lens; six quadrature nodes recover the effect itself. Second-order attribution patching predicts each component's interaction with every other from one backward pass per ablated run, and recovers strong interactions between individually weak components that single-effect rankings miss. Its setup does not yet pay for itself at 24–32 blocks but does at 80 in an exploratory comparison; because its share of the cost falls as $8/L$, we expect it to matter most among the hundreds of attention heads of large models.

AI use statement

In this work we used generative AI tools, a large-language-model coding and writing assistant, to implement methods and experiments; to design experiments and draft their pre-registered pass/fail criteria; to formulate and check the identities of §3 and §4.5; to propose and refine hypotheses; to interpret results; to generate the templated prompt sets and to tokenise and chunk the FineWeb corpus; and to translate between Chinese and English. We have not used generative AI tools to generate synthetic training data or for qualitative or thematic data analysis, which are not applicable to this work. Additionally, we used generative AI tools to write and edit code, create the figures, draft and edit the text, search for and summarise related literature, format references, and suggest the structure and title. We have reviewed all AI-assisted work: every number in the paper is computed by a script in the released repository from saved result files, every experiment's criteria were committed before it was run, and results that failed a criterion are reported as failures. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work measures publicly released language models on public web text and templated prompts; it involves no human subjects and no personal data. Its methods estimate what model components contribute to an output and release no modified model.

Reproducibility statement

The repository, an anonymised copy of which accompanies the submission, contains the pathspace package, one script per experiment named for the identifier in the ledger of Appendix R, the design document that fixed each experiment's criteria before it was run, and the raw result files from which every figure and table is generated. The three small models were run in float32 on one Apple M5 (PyTorch MPS); Llama-3-8B in float32 and Llama-3-70B in bfloat16 on two NVIDIA A800-80GB, with the final norm and head evaluated in float32 for the model's own targets (Appendix E); batching all single-block ablations of a position into one forward pass fits the 70B model in memory.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, et al. SmolLM2: When smol goes big — data-centric training of a small language model. arXiv preprint arXiv:2502.02737, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning, 2023.

Leo Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, https://transformer-circuits.pub/2021/framework/index.html, 2021.

R. P. Feynman. Space-time approach to non-relativistic quantum mechanics. Reviews of Modern Physics, 20:367–387, 1948.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Proceedings of EMNLP, 2022.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. arXiv preprint arXiv:2304.05969, 2023.

Zhiren Gong, He Lu, Tiantong Wang, Yichi Zhang, Yixin Wang, Zihao Zeng, Ming Xiao, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. arXiv preprint arXiv:2607.01940, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. arXiv preprint arXiv:2607.15495, 2026. Transformer Circuits Thread, https://transformer-circuits.pub/2026/workspace/index.html.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In Conference on Language Modeling (COLM), 2024. arXiv:2403.17806.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. Journal of Global Optimization, 13(4):455–492, 1998.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. arXiv preprint arXiv:2403.00745, 2024.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. arXiv preprint arXiv:2307.15771, 2023.

Neel Nanda. Attribution patching: Activation patching at industrial scale. https://www.neelnanda.io/mechanistic-interpretability/attribution-patching, 2023.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

Román Orús. A practical introduction to tensor networks: Matrix product states and projected entangled pair states. Annals of Physics, 349:117–158, 2014.

Ivan Oseledets and Eugene Tyrtyshnikov. TT-cross approximation for multidimensional arrays. Linear Algebra and its Applications, 432(1):70–88, 2010.

Ivan V. Oseledets. Tensor-train decomposition. SIAM Journal on Scientific Computing, 33 (5):2295–2317, 2011.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. arXiv preprint arXiv:2406.17557, 2024.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. Annals of Physics, 326(1):96–192, 2011.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International Conference on Machine Learning, 2017.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pp. 407–416, 2024. arXiv:2310.10348.

Andreas Veit, Michael Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In Advances in Neural Information Processing Systems, 2016.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In Advances in Neural Information Processing Systems, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In International Conference on Learning Representations, 2023.

Luyang Zhang and Jialu Wang. When attribution patching lies: Diagnosis and a second-order correction. arXiv preprint arXiv:2606.09899, 2026.

A Methods compared

Table 3: What each method estimates, from what, with which error, at what cost (L components, one position). "Fwd" is a forward pass; a backward pass costs about two. The second-order estimator targets the same conditional quantity as conditional co-ablation, $\delta_{j|S} - \delta_j$ for $S = \{k\}$, but predicts it for every j from one backward pass instead of measuring it; the Hessian-vector-product correction targets a single component's error at the clean point.

method	intervention and object	error source	cost for all L components
direct-logit attribution, logit lens	none; a write read through the unembedding	all downstream processing ignored	1 fwd
tuned lens, Jacobian lens	none; a learned or corpus-averaged transport	context averaging (and position summation)	a fit, then 1 fwd
attribution patching (Nanda, 2023)	first-order Taylor of a removal, single effects	curvature along the removal	1 fwd + 1 bwd
AtP* (Kramár et al., 2024)	AtP with the QK fix and GradDrop	residual curvature	$\approx L$ bwd
integrated gradients, EAP-IG (Sundararajan et al., 2017; Hanna et al., 2024)	path integral, single effects	quadrature	n bwd per path
HVP correction (Zhang & Wang, 2026)	second-order Taylor at the clean point, single effects	cubic remainder	1 fwd + 1 double bwd
activation patching, ablation	measured single effects	none	L fwd
measured pair ablations	measured interactions $I(k, j)$	none	$L^2/2$ fwd
conditional co-ablation (Gong et al., 2026)	measured $\delta_{j\|S} - \delta_j$ after removing a set S	none	1 fwd per candidate j, per S
second-order attribution patching (ours)	$\hat{I}(k, j)$ for all j from the k-removed run	$e_{j\|k} - e_j$, cancels when both marginals curve alike	L fwd + L bwd for all $L^2/2$ pairs

B The sum rule and its remainder

With the frozen norm $\tilde{z} = \tilde{g} \cdot h_{L-1}[p] + b$ and $h_{L-1}[p] = h_{-1}[p] + \sum_k \Delta_k[p]$, so the identity reading sums exactly, $\sum_k A^0_{k,p} + \tilde{g} \cdot h_{-1}[p] + b = \tilde{z}$. The transported sum does not: it differs from $\tilde{z}$ by a remainder that is exactly computable from the same backward pass,

$$\Big(\sum_k A_{k,p} + E_p\Big) - \tilde{z} + b \;=\; R, \qquad R \;=\; \sum_k \big(\nabla_k[p] - \tilde{g}\big) \cdot \Delta_k[p] \;+\; \big(\nabla_{\text{emb}}[p] - \tilde{g}\big) \cdot h_{-1}[p]. \quad (13)$$

R vanishes only when every transport acts as the identity on the write it carries, so the sum rule is an empirical property of residual stacks and not a theorem. It fails badly outside them: for two blocks $h \mapsto 2h$ with a linear readout, the writes are h_0 and $2h_0$, the gradients are $2\tilde{g}$ and $\tilde{g}$, and the transported sum is 8 against $\tilde{z} = 4$ — the whole of which Eq. 13 accounts for ($R = 4$, dominated by the embedding term). On Qwen3-0.6B over 120 positions Eq. 13 is exact to 3×10^{-7} relative, and the sum rule holds in aggregate but not pointwise: the median ratio is 0.999 at the target position (1.017 over the lattice), while $|R/\tilde{z}|$ has median 0.057, ninetieth percentile 0.158 and maximum 1.04. The route accounts for the logit on average; it does not account for it at every position, and we use the sum only as an aggregate check.

C Averaged transports: convergence and position summation

The Jacobian lens of §4.1 was fitted on 32 sequences and sums the transport over future positions. To separate estimation from averaging, we fitted two averaged transports on up to 128 contexts from a FineWeb corpus disjoint from the test cells (Qwen3-0.6B, 64 tokens, one random position per context) and applied each test cell's own output covector: one that averages the same-position transport from $h_k[p']$ to $h_{L-1}[p']$ over contexts, and one that also sums over future positions, as the Jacobian lens does.

Table 4: Spearman with the measured removal effect on 75 test cells × 28 blocks of Qwen3-0.6B (frozen-norm logit), pooled and median over the first third of the stack, as the number of fitting contexts grows. On the same cells the per-context transport scores 0.819 and the identity 0.438.

fitting contexts	8	16	32	64	128
same position, pooled / first third	0.529 / 0.09	0.517 / 0.11	0.549 / 0.15	0.521 / 0.14	0.501 / 0.11
future positions summed	0.505 / 0.08	0.481 / 0.02	0.503 / 0.08	0.529 / 0.11	0.483 / 0.09

Neither transport improves with sixteen times more fitting data, and summing future positions changes the score by at most 0.05 in either direction: the averaged transport's failure in the early stack is the averaging over contexts, not an under-fitted estimate and not the position sum.

D A second task: in-context number comparison

To test a computation of a different structure from next-token prediction on web text, we asked the three small models and Llama-3-8B which of two two-digit numbers is larger (four templates, 150 prompts; the answer token is the larger number, or its tens digit for tokenizers that split numbers into digits, and the competitor is the smaller) and scored every reading on the log-probability of the answer at the answer position, with each target's own gradient. The models are often wrong: the answer is the top prediction on 61%, 20%, 48% and 76% of prompts (Qwen3, pythia, SmolLM2, Llama-3-8B).

Table 5: Number comparison, the model's own log-probability of the answer, all blocks at the answer position of 150 prompts. Per-context / identity with each target's own gradient; document-clustered intervals are per prompt here. Second order: all pairs at 10 prompts.

model	per-context / identity	gap [95% CI]	identity early sign / null	second order fwd / sym
Qwen3-0.6B	0.759 / 0.268	+0.49 [+0.46, +0.52]	0.508 / 0.493	0.672 / 0.663
pythia-410m	0.751 / 0.198	+0.55 [+0.53, +0.57]	0.563 / 0.519	0.672 / 0.735
SmolLM2-360M	0.749 / 0.212	+0.54 [+0.51, +0.56]	0.508 / 0.499	0.671 / 0.741
Llama-3-8B	0.729 / 0.180	+0.55 [+0.52, +0.57]	0.526 / 0.512	0.613 / 0.663

The FineWeb picture carries over: the per-context transport leads by about half a rank correlation, the identity's early sign is at its null, and the second-order estimator orders interactions at 0.61–0.67. Symmetrising helps in three of the four models and slightly hurts in Qwen3 (0.672 → 0.663).

E Precision: what bf16 does to a measured interaction

Llama-3-70B runs in bf16 on two 80 GB cards; every other run is fp32. A bf16 run rounds the residual stream and, if the final norm and head run in the model's dtype, the readout: each logit is then rounded to bf16 before the log-softmax, a step of 0.125 for logits in $[16, 32)$. The frozen-norm target was always read out in fp32. On Llama-3-8B we ran the same positions and pairs in fp32, in bf16, and in bf16 with the norm and head evaluated in fp32 (Table 6; pre-registered as Q76).

bf16 leaves the estimator nearly untouched: the bf16 estimate ranks the fp32-measured interactions at 0.638 against 0.651 for the fp32 estimate, and the two estimates agree at 0.969. It does not leave the measurement untouched: bf16-measured interactions agree with the fp32

Table 6: The model's own log-probability in three precisions at 8B (150 positions; second order over all 4 960 pairs at 10 positions) and two at 70B (120 positions; 3 200 sampled pairs at 8 positions). "Scored against fp32" ranks the estimate against the fp32 run's measured interactions on the same pairs; "agreement" is the Spearman between a run's measured interactions and the fp32 ones.

run	per-context / identity	second order sym	scored against fp32	agreement with fp32	median $\lvert I\rvert$
Llama-3-8B, fp32	0.693 / 0.201	0.651	0.651	1	0.021
Llama-3-8B, bf16	0.666 / 0.203	0.495	0.638	0.780	0.044
Llama-3-8B, bf16, fp32 readout	0.689 / 0.200	0.593	0.646	0.882	0.028
Llama-3-70B, bf16	0.720 / 0.200	0.307	—	—	0.015
Llama-3-70B, bf16, fp32 readout	0.815 / 0.214	0.557	—	—	0.009

ones at 0.780, their median deviation (0.027) exceeds the median fp32 interaction (0.021), and scored against them the same estimate falls to 0.495. Evaluating the norm and head in fp32 restores the main phase (0.689 against 0.693) and halves the measurement's deviation from fp32 (0.014; agreement 0.882); scored against its own measurement the estimate recovers to 0.593, short of the 0.60 we pre-registered, and against the fp32 measurement it is 0.646. What remains is the bf16 residual stream, and it is in the measurement, not in the estimate. At 70B the fp32 readout moves the main phase on log p from 0.720 to 0.815 and the median measured interaction from 0.015 to 0.009; the estimate scores 0.307 against the bf16-read measurement and 0.557 against the fp32-read one, in the band we pre-registered as degraded, and, unlike in the fp32 models, worst where the conditioning removal is smallest (0.33 against 0.69 for the largest third), consistent with the smallest interactions sitting below the bf16 residual stream's rounding; on the frozen-norm logit, whose readout was always fp32, the forward estimate scores 0.641 (Table 16). The 70B rows of Tables 1 and 2 are from the fp32-readout run.

F The frozen norm and the model's own logit

What the frozen norm leaves out, exactly. $\tilde{z}$ is not the model's logit, and the difference is not an approximation error but one identified term. Writing $\tilde{z}(h) = L(h)/s + c$ with L linear and c the norm and readout bias, removing a write changes the residual and hence the final norm's own scale $s \to s'$, so the model's logit changes by

$$\Delta z \;=\; \Delta\tilde{z} \;+\; \Big(\tfrac{s}{s'} - 1\Big)\big(\tilde{z} + \Delta\tilde{z} - c\big). \tag{14}$$

The second term multiplies every logit by s/s': it changes the temperature of the distribution, not the target token's standing against its competitors. It is not small. Over 150 positions and every block of the three models its median magnitude is 0.43–0.59 against 0.58–0.86 for $\Delta\tilde{z}$ itself (median s/s' 1.011–1.029), which is why $\Delta\tilde{z}$ and Δz correlate at only 0.59–0.62; subtract the common-mode term and the agreement is exact (Spearman 1.000 in all three models). Everything we decompose is $\Delta\tilde{z}$, the token-specific part. §4.1 reports every headline comparison against Δz and against the change in $\log p_t$ as well, and the ordering of the readings is the same under all three.

G The real-amplitude interference observable

The classes $c_k[p]$ are real vectors whose sum is the transport, so their coherent and incoherent sums can differ:

$$V_p \;=\; \frac{\big\|\sum_k c_k[p]\big\|}{\sum_k \|c_k[p]\|} \;\in\; \Big[\tfrac{1}{\sqrt{L}},\, 1\Big] \text{ for uncorrelated classes,} \tag{15}$$

with values below $1/\sqrt{L}$ only when classes cancel. This is interference in the real-amplitude sense — amplitudes that add before the norm is taken — and nothing more; the residual stream has no complex structure to carry a phase (we tested for one and found none), so the observable is the cosine structure between consecutive classes and the coherence ratio V_p. On Qwen3-0.6B, over 400 contexts, adjacent classes anticorrelate in every context: cosine -0.10 to -0.12 against a layer-permutation null of -0.02 ($z = -152$ to -176), and

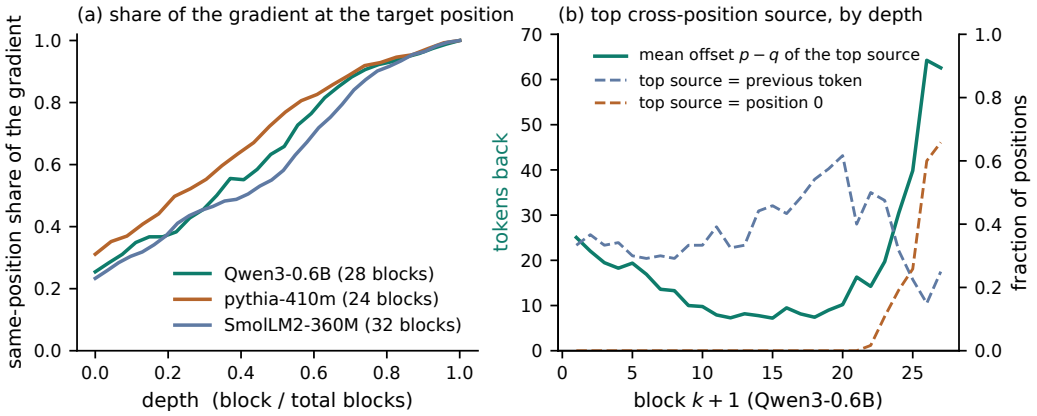


Figure 3: (a) The share of the target logit's gradient that sits at the target's own position, by block, for three architectures: a quarter to a third at block 0, half at 35–45% depth, nine tenths at 70–80% depth. (b) Qwen3-0.6B: which other position is the top cross-position source of the path class at each block. Early blocks read far back in the context (17–25 tokens), middle blocks read the previous token (the top source at up to 62% of positions), the last two blocks read position 0, the attention sink, at 60–66% of positions.

$V_p = 0.14$ against the incoherent 0.19. The residual stream overshoots and corrects at adjacent blocks; the same structure appears across positions in Appendix H, where the cross-position attributions are 90% of the early magnitude and net to 2% of the logit. We use interference as a descriptive fact about the depth chain; two attempts to exploit it (adjacent-pair pruning, route-guided steering) closed negative (Appendix L).

H The lattice: where the route goes

The route is not confined to the target's position. Fig. 3a shows the same-position share of the gradient rising monotonically from 0.22–0.31 at block 0 to 1 at the last block in all three architectures, with the same S shape. For the first ten blocks of Qwen3-0.6B, 90% of the first-order attribution magnitude lies at positions other than the target while its net sum is +2% of the logit: the cross-position terms are large individually and cancel collectively, the same overshoot-and-correct structure as the adjacent anticorrelation of §3, seen across positions. The lattice identity holds (median 1.017), and the off-diagonal cells are causally real: ablating a block's write at a source position $q \neq p$ changes the logit at p by a median 0.06 logits, predicted by the lattice attribution at Spearman 0.69 (320 cells; shuffle null 0.08 ± 0.06, $z = +10.8$; sign agreement 0.84).

Which positions the route reads is architectural (Fig. 3b, Table 7). Qwen3's early blocks draw from far back in the context, its middle blocks from the previous token, and its final two blocks from position 0, the attention sink, which is the top source at 60–66% of positions and carries 31–52% of those blocks' cross-position mass. SmolLM2 shares the previous-token regime (0.53 at block 20) and shows only a trace of the sink (0.13); pythia has neither.

I Mediation: where a write's effect is realised

With the frozen norm the effect of a removal decomposes exactly into the removed write's own output projection and the change in every later block's write, read at the output:

$$\delta_k = \underbrace{-\tilde{g} \cdot \Delta_k}_{\text{direct}_k} + \sum_{j>k} \underbrace{\tilde{g} \cdot (\Delta'_j - \Delta_j)}_{R_{k,j}}, \tag{16}$$

where Δ'_j is block j's write on the ablated run (maximum error over 1 120 cells: 1.8×10^{-5}). On Qwen3-0.6B the mediated share is 0.84–0.99 for blocks 0–9 and falls to 0.2–0.4 at blocks

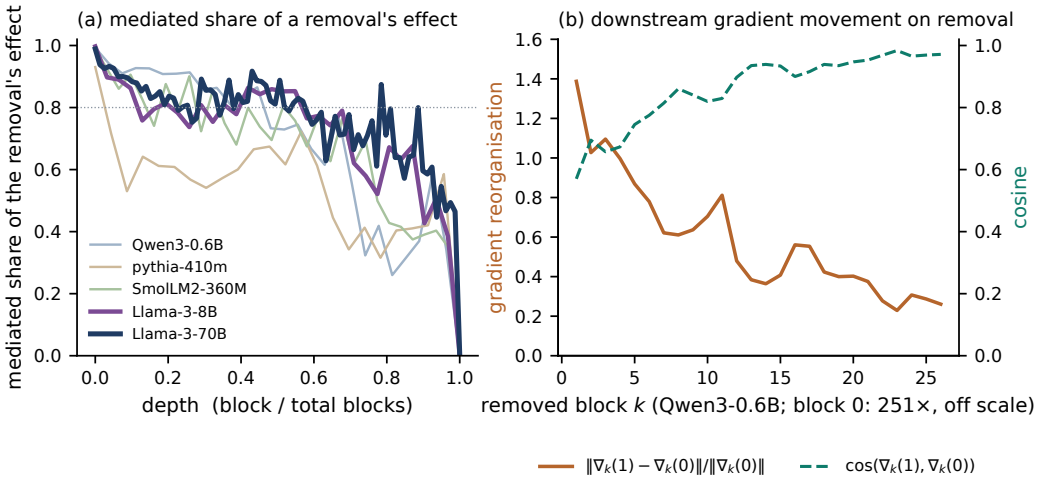


Figure 4: (a) The mediated share of a removal's effect, $|\sum_j R_{k,j}|/(|\text{direct}_k| + |\sum_j R_{k,j}|)$, by depth in five models (Llama-3 in bold): near total at block 0, zero at the last block, and architectural in between (0.9 early in Qwen3, 0.75–0.9 in both Llama-3 models and SmolLM2, 0.55–0.7 in pythia's parallel-residual stack). (b) Qwen3-0.6B: how far the downstream gradient moves when a write is removed; early removals rotate it by 45–55° and change its norm by 100%.

Table 7: Depth-chain structures in three architectures at the 0.4–0.6B scale (60 positions per model for the lattice statistics, 30 for the response decomposition) and, where the scale runs measured them, in Llama-3-8B and 70B (—: not measured at scale). At 8B the block-0 boundary shows in the quadrature: six nodes err by 1.25 at block 0 and by 0.000 at every other block.

	Qwen3-0.6B	pythia-410m	SmolLM2-360M	Llama-3-8B	Llama-3-70B	
same-position gradient share at block 0	0.25	0.31	0.23	—	—	shared
block-0 gradient reorganisation on removal	234×	204×	411×	—	—	shared
per-context vs identity gap (Table 17)	+0.37	+0.44	+0.45	+0.37	+0.59	shared
early mediated share (min over first third)	0.86	0.53	0.72	0.74	0.75	architectural
late-half share of early removals' response	0.82	0.50	0.69	—	—	architectural
final blocks' top source is position 0	0.61	0.06	0.11	—	—	Qwen3-specific
previous token as top source (max, block)	0.68 (20)	0.53 (1)	0.53 (20)	—	—	modern two

21–26 (Fig. 4a); the direct term, which is what the identity reading measures, is 1–8% of an early write's effect. The linearised response has the right sign 79% of the time; its magnitude error is the curvature of §4.4.

Where is the response realised? For an early removal, 75–85% of $\sum_j |R_{k,j}|$ lies in blocks 16–27 and 2–8% in the three blocks that follow the removed one; the responses of the next five blocks are zero at the median. Adjacent self-repair (McGrath et al., 2023) does not appear at the logit level here; early blocks write features that are not yet output-aligned, and the last twelve blocks turn them into the prediction. In pythia the response is spread through the stack (late-half share 0.50) and early writes are a third output-aligned at once; in SmolLM2 (0.69) and Qwen3 (0.82) it is deferred. The degree of mediation tracks the architecture: sequential RMSNorm stacks defer the readout most, and it is there that direct-logit attribution is blindest.

J Shared across the models, and architectural

Table 7 separates what is a property of the depth chain at this scale from what is one model's. Three facts hold in all three small architectures and, as far as the scale runs measured them, in Llama-3-8B and 70B: the route leaves its position early and returns monotonically (80 and 84% of early attribution off-target at scale), block 0 is a regime boundary (at 8B six quadrature nodes err by 1.25 there and by 0.000 elsewhere), and the per-context transport predicts interventions where the identity cannot (+0.37 and +0.59) (Fig. 5). Three are

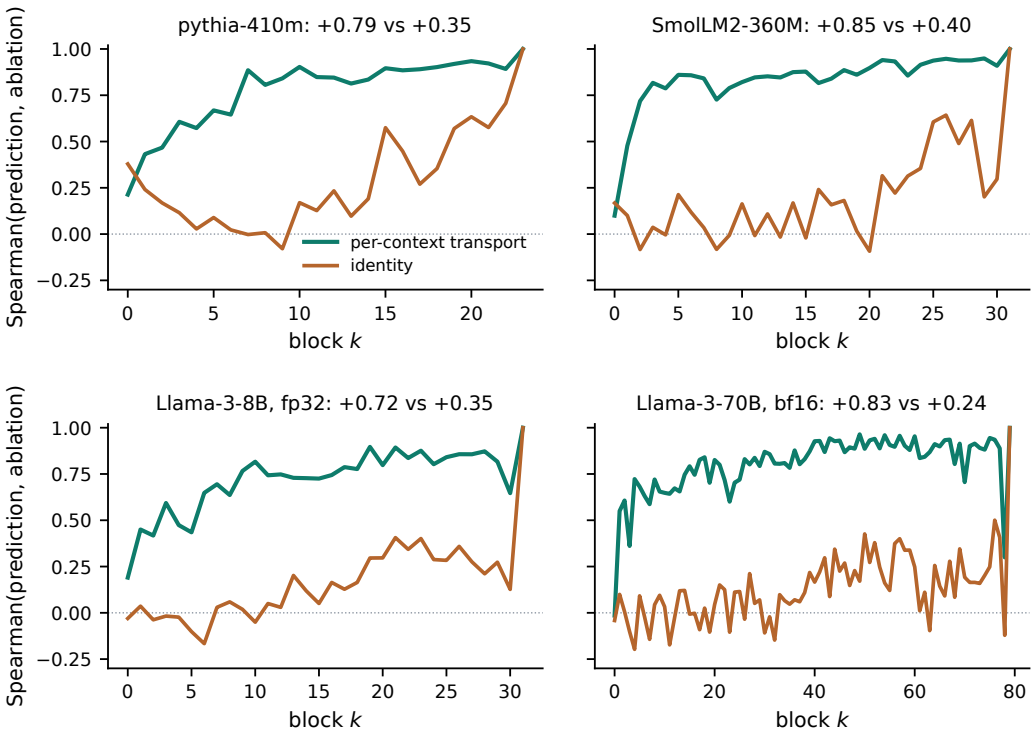


Figure 5: Per-block Spearman between predictor and ablation effect in pythia-410m and SmolLM2-360M (100 positions each), Llama-3-8B in fp32 (150) and Llama-3-70B in bf16 (120). The identity is uninformative through block 14 of 24 and 24 of 32 in the small models; in Llama-3-8B its largest per-block value before the last block is 0.41, in Llama-3-70B 0.50.

architectural: how much early effects are mediated, where the response is realised, and whether the final blocks read a position-0 sink. The parallel-residual, LayerNorm, untied-head model (pythia) makes early writes partly output-aligned at once and spreads the response through the stack; the two sequential RMSNorm models defer both.

K Toward capability: what Path-Space does and does not buy

Everything so far reads a model. Several routes lead from the same object to changing what a model does at a given cost. We tested five by intervention, in the order of their expected promise: interaction-aware pruning, route-localised editing, task-specialised pruning with its collateral, diagnosis-driven targeted fine-tuning, and per-token layer skipping (Fig. 6). One object built along the way, the set-effect tensor train, is a positive result as a model of joint removals and as a zero-cost selector across prompts; the five interventions close negative or as trades, for reasons that recur: many-body redundancy, sensitivity is not editability, and the information that makes a per-token route exploitable lives in the backward pass.

Interaction-aware pruning (tested, negative). Structured pruning removes the components whose joint removal costs least; additive importance scores ignore interactions, and Q58–Q60 showed that one backward pass per component predicts pairwise interactions among impactful components with the right sign 80–90% of the time. We pre-registered the natural test: on pythia-410m and IOI, from the pool of the 96 heads (of 384) whose individual removal is least harmful on 15 selection prompts, remove K heads at all positions chosen by first-order importance, by measured single effects, or greedily by the pairwise-corrected predicted joint effect, and measure the damage on 15 held-out prompts (Table 8). The result is a clean negative that bounds the method. Removing 24 individually harmless heads raises the metric; removing 48 collapses it under every rule (held-out accuracy 0–20%),

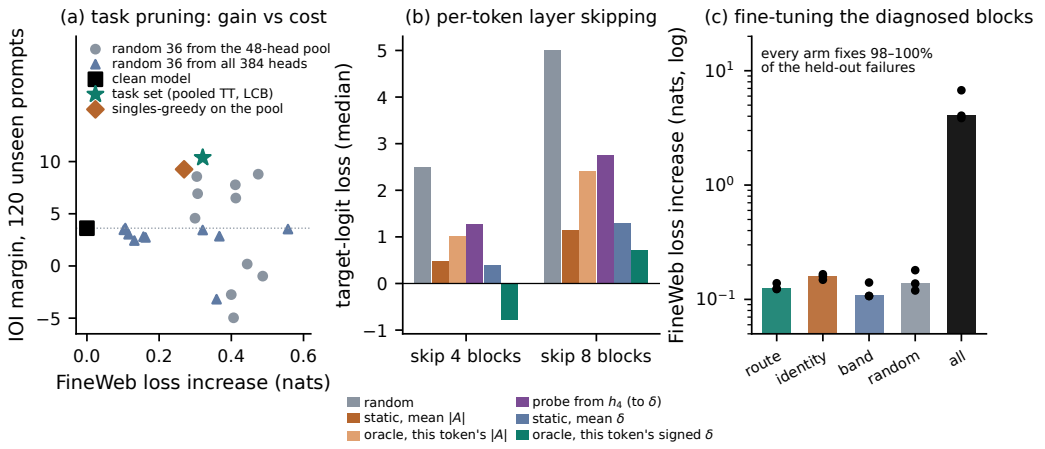


Figure 6: Capability routes. (a) Task-specialised pruning on pythia-410m: IOI margin on 120 unseen prompts against FineWeb loss increase, for the pooled tensor-train set, singles-greedy, and random 36-head sets from the pool and from all heads; the task gain sits on the same cost frontier as random removal. (b) Per-token layer skipping on Qwen3-0.6B: median target-logit loss when skipping 4 or 8 blocks at the token by each rule; negative means the target logit rises. The oracle that knows this token's signed costs beats every static schedule; the probe from the block-4 residual is worse than static. (c) Targeted fine-tuning of four blocks on a failure set (in-context number comparison, pythia-410m): forgetting (FineWeb loss increase, log scale; bars medians, dots seeds) by the rule that chose the blocks (route $\{7, 10, 11, 17\}$, identity $\{0, 13, 17, 18\}$, band 10–13, random four, all 24), Every arm fixes 98–100% of the 200 held-out failures, so the choice of blocks changes only the forgetting, and the blocks the route diagnoses forget about as much as a random four (median 0.12 against 0.14 nats).

Table 8: Interaction-aware head pruning on pythia-410m, IOI. Pool: the 96 heads with the least harmful measured single removal on 15 selection prompts. Held-out mean change of $\tilde{z}_{\mathrm{IO}} - \tilde{z}_{\mathrm{S}}$ (positive = better) and held-out accuracy after removing K heads at all positions; random draws from the same pool averaged over 10 draws.

K	first-order greedy	measured-singles greedy	pairwise-corrected greedy	random from pool
24	+6.06 (acc. 1.00)	+5.23 (1.00)	+4.95 (1.00)	+1.70
48	−7.31 (0.13)	−8.57 (0.00)	−6.50 (0.20)	**−3.42**
72	−6.55 (0.07)	−6.80 (0.13)	−7.42 (0.07)	**−5.80**

selection prompts, $K = 48$: sum of singles +9.76 vs measured joint −8.25; pairwise-corrected prediction +26.0 vs measured −5.59

while 48 random heads from the same pool cost less than half as much. On the selection prompts the sum of single effects of the 48-head set predicts a gain of +9.8 and the pairwise-corrected prediction +26.0; the measured joint effect is −8.2 and −5.6. Redundancy among the individually least harmful heads is a many-body effect: removing one redundant head makes removing a second less harmful, so pairwise interactions inside the pool are positive and point the wrong way, until the whole group is gone. The second-order estimator is validated for pairs and is not a set-level estimator; a pruning guide would need higher-order terms or a representation of shared function.

The set-effect function as a tensor train (tested: a model that avoids the collapse, not a cheaper optimiser). The pruning failure says that $f(S)$, the joint effect of removing a set of components, is many-body. Treating it as a tensor with one binary site per component and fitting a tensor train (Oseledets, 2011) to sampled ablations asks how compressible it is along the chosen ordering of components: the additive model is the exact rank-two tensor train, so any gain from $\chi > 2$ is structure that no per-component score can hold (Table 11). Rank is not interaction order — the fully many-body $\prod_i s_i$ has tensor-train rank one — so we read χ as separability and leave the many-body claim to the predictive comparisons. Warm-started from the additive model and fitted by alternating least squares with validation early stopping, a tensor train of bond dimension $\chi = 2$ — one dimension

beyond additivity — cuts the held-out error on block sets of up to 12 of 27 by 23% (0.166 → 0.128; the pairwise estimator adds nothing over additive at this set size) and on head sets of up to 24 of a 48-head pool by 36% (0.105 → 0.067) at Spearman 0.98; larger χ overfits 1 250 samples, and the shuffled-label null sits at 0.45–0.51. The hypercube of component subsets is the setting the matrix-product toolkit was built for: a product space with 2^{48} points and a function on it of low tensor-train rank, unlike the transport of §3.1, whose space is polynomially small and needs no compression. Used as a surrogate for selection, simulated annealing on the tensor train chooses 24-head sets whose measured effect beats singles-greedy and the random median in 6 of 8 prompts (medians +7.6 against +6.7 and +5.2 logits of improvement, since these heads help when removed); the two failures are the optimiser finding holes in the surrogate (a predicted +1177 measured +3.2), the standard failure of optimising a fitted model outside its sampled region. The collapse regime is where this matters, and a pre-registered amendment tested it: with training subsets of up to 40 heads and the surrogate clipped to the observed range of f before optimisation, a $\chi = 2$ tensor train predicts held-out sets of up to 40 heads at Spearman 0.96 (error 0.065 against 0.136 additive), every prediction at the chosen optimum is within 15% of its measured value, and the sets it selects beat singles-greedy and the random median in 8 of 8 prompts at $K = 36$ and 7 of 8 at $K = 44$ (Table 10). On the three prompts where singles-greedy and random selection collapse, as every rule did in the pruning test above, the tensor-train sets stay at +8 to +12 logits: the surrogate has learned which heads are redundant with each other and removes at most one of each group. The fair competitor is not a per-component score but a search that spends the same budget on measured joint effects: a sequential greedy that adds, at each step, the head whose measured joint removal with the current set is least harmful uses 1 166 forwards to reach $K = 44$, against $\approx 1\,650$ for fitting the tensor train. It selects better sets than the tensor train at $K = 36$ in 8 of 8 prompts (medians +11.6 against +9.6) and at $K = 44$ in 7 of 8 (+10.9 against +10.1), and it collapses once itself (−7.1 at $K = 44$ on a prompt where the tensor-train set gives +10.1). As an optimiser at matched cost the tensor train is therefore slightly below measured greedy; what it is, and what greedy is not, is a model of the whole set-effect function, 2^{48} subsets predicted at Spearman 0.96 from 1 648 samples with no collapse at any tested set. Whether that model has to be a tensor train is the next paragraph.

Does it have to be a tensor train? (tested: no.) The comparison so far is against per-component scores; the stronger baseline is another regressor given the same data. We ran five alternatives on the identical measurements — the same 1 249 subsets, the same eight prompts, the same five folds with the forty-eight measured singles always in the training fold, the same clipping and the same swap-search optimiser — and let each select its own $K = 36$ set, which we then measured (Table 9). Two of them beat the tensor train. Gradient boosting has 19% lower held-out error (0.223 against 0.275) and the highest rank correlation (0.827 against 0.770); a two-layer perceptron selects the best sets (+9.1 median measured improvement against +8.0) and, like every learned surrogate except the size-only control, never collapses, where the tensor train still collapses once in eight. What survives is the claim one level up, and it is the one the pruning failure motivated: the set-effect function is learnable from of order a thousand sampled ablations, and any learned surrogate avoids the many-body collapse that every per-component rule suffers — the size-only control, which knows only $|S|$, collapses on all eight prompts and is the null that shows the surrogates are reading structure and not set size. The tensor train is one such surrogate, the most compact (two numbers per component) and the only one whose form states what it assumes, but it is not the most accurate, and we withdraw the claim that the tensor-network structure is what does the work here.

Deploying the tensor train (tested: transfer is the gain). Three refinements of the tensor-train route were pre-registered on one shared dataset (a common pool of the 48 heads least harmful on average over 8 prompts, 1 249 subsets measured on every prompt). A better optimiser buys nothing: single-start annealing finds the exact surrogate optimum on every prompt (verified by exhaustive enumeration of the $\binom{48}{4}$ complements at $K = 44$), so the remaining gap between predicted and measured effect (+15.9 against +7.7) is surrogate fidelity, not search. Adaptive sampling by maxvol-pivoted TT-cross loses to

Table 9: Surrogates for the set-effect function at matched data and budget (pythia-410m, IOI, 48-head pool). Five-fold cross-validation on the same 1 249 measured subsets of the deployment dataset, medians over 40 folds; the measured column is the true effect of the $K = 36$ set each surrogate selects, and a collapse is a set whose measured effect is negative. Subset sizes here run to 40, so the errors are not comparable with Table 11.

surrogate	rel. error ($\lvert S\rvert \geq 4$)	Spearman	measured at $K = 36$	collapses
additive (measured singles)	0.568	+0.588	+7.7	0/8
quadratic ridge (all pairs)	0.438	+0.701	+8.0	0/8
MLP 2×256	0.294	+0.794	**+9.1**	0/8
gradient boosting	**0.223**	**+0.827**	+8.1	0/8
size only (control)	0.448	+0.364	−3.4	8/8
tensor train $\chi = 2$	0.275	+0.770	+8.0	1/8

random sampling with least squares at equal budget (held-out error 0.78 against 0.52 at rank 2). Transfer is where the gain is: a $\chi = 2$ tensor train fitted on the other seven prompts' 8 743 measured subsets, with targets normalised per prompt, predicts a held-out prompt's set-effect function at Spearman 0.83–0.90 (median 0.88; the candidate pool in this protocol was chosen using all eight prompts, so we re-ran it with the pool re-selected from the seven training prompts alone — the pools differ by 2–5 of 48 heads and the leak-free median is 0.90, with an MLP at 0.93) — better than the prompt's own tensor train fitted on its own 1 000 ablations (0.51–0.65), because the redundancy structure among heads is shared across prompts and seven prompts' data beat one prompt's — and the sets it selects for the unseen prompt, with no new ablations, have positive measured effect at $K = 36$ and $K = 44$ on all eight prompts (medians +7.2 and +7.4), match the prompt's own tensor train (+7.0, +7.9), and beat singles-greedy and random where those collapse (−1.6 and −3.3 at $K = 44$). A set-effect model learned once from ablations on a few prompts and applied to new prompts at zero marginal cost is the deployable form of the tensor network here, and it is a capability-relevant decision rule: which heads can be removed together without breaking the behaviour.

Uncertainty-aware surrogate (tested: removes the collapses, matches measured greedy at $K = 44$). The remaining failure of the surrogate route is the optimiser exploiting regions the data do not constrain: the single tensor train's optimum over-predicts its measured effect 4× at $K = 36$ and on two of sixteen selections collapses (−4.9 and −10.8 logits). Two pre-registered remedies were tested on the shared dataset: bagging eight tensor trains on bootstrap resamples (Breiman, 1996) and optimising the lower confidence bound (ensemble mean minus κ standard deviations), and infilling three rounds of the ten best sets on the optimiser's trajectory, measured and added to the training set (30 forwards), the infill step of surrogate-assisted optimisation (Jones et al., 1998). The lower bound moves the optimum to where the ensemble spread is below that of random sets and removes every collapse; infill lifts the level and calibrates the destination. Together they beat the single tensor train on 8 of 8 prompts at both K (medians +8.5 and +8.6 against +7.2 and +7.5), bring the prediction at the optimum from 1.8× to 1.10× the measured value at $K = 44$, and match the measured sequential greedy at $K = 44$ (+8.63 against +8.55) with 31 forwards against 1 116; at $K = 36$ greedy remains ahead (+10.2 against +8.5) and the optimum is still over-predicted 2×, the honest limit. In the deployable leave-one-prompt-out form, 30 ablations on the new prompt spent at the transferred surrogate's optimum improve the selection on 8 of 8 held-out prompts (medians +8.5 against +7.4 at $K = 44$). The residual 2× over-prediction at $K = 36$ is not a bias of the model class: out-of-bag residuals of the ensemble have near-zero mean at every subset size (+0.9 at $\lvert S\rvert = 34$–37 on effects of ~ 9) while their standard deviation grows 16× from small to large sets (0.28 → 4.4), so the optimum is a winner's curse on size-dependent noise. Accordingly, conditioning the surrogate on size — a size window, an out-of-bag size offset, a heteroscedastic penalty $\sqrt{\text{spread}^2 + \hat{\sigma}^2(\lvert S\rvert)}$ — leaves the optimum over-predicted 1.65–2.6× (pre-registered, negative), with a small selection gain from the heteroscedastic penalty (+9.0 against +8.6 at $K = 36$); what calibrates it is the measurement the procedure already makes: the median measured-to-predicted ratio over

Table 10: Removing K of the 48 least harmful heads (pythia-410m, IOI; positive = the metric improves). Sets chosen by simulated annealing on a clipped $\chi = 2$ tensor train fitted to 1 648 sampled ablations of the same prompt, by measured single effects (greedy), and at random (median of 10). Collapse cases in bold.

	$K = 36$			$K = 44$		
prompt	tensor train	singles-greedy	random	tensor train	singles-greedy	random
1	+12.2	+11.8	+9.6	+12.5	+12.2	+11.6
2	**+12.6**	**−8.0**	**−3.7**	**+11.5**	**−4.9**	**−4.4**
3	+7.4	+5.1	+4.2	+6.9	+5.5	+5.3
4	+9.3	+7.5	+6.9	+10.0	+8.6	+9.0
5	**+7.8**	**−6.1**	**−0.2**	**+8.7**	**−6.1**	**−4.4**
6	+10.9	+10.4	+7.3	+11.0	+11.0	+10.6
7	+9.8	+9.6	+7.8	+10.0	+10.3	+10.1
8	**+8.5**	**−2.2**	**−6.0**	**+10.1**	**−8.8**	**−7.1**
median	**+9.6**	+6.3	+5.5	**+10.1**	+7.0	+7.2

Table 11: The set-effect function $f(S)$ as a tensor train over components. Held-out median relative error on subsets of size ≥ 4 and Spearman with the measured f. Blocks: Qwen3-0.6B, 27 sites, sets of size 2–12, 12 positions; heads: pythia-410m IOI, the 48 least harmful heads, sets of size 2–24, 8 prompts. Training 1 250 sampled subsets per target.

	blocks		heads	
model	rel. error	Spearman	rel. error	Spearman
additive (measured singles)	0.166	0.922	0.105	0.960
pairwise (singles $+\hat{I}$)	0.169	0.917	—	—
tensor train $\chi = 1$	0.170	0.926	0.120	0.978
tensor train $\chi = 2$	0.128	0.947	0.067	0.982
tensor train $\chi = 4$	0.141	0.935	0.078	0.972
tensor train $\chi = 8$	0.161	0.908	0.090	0.944
shuffled-label null ($\chi = 4$)	0.445	0.062	0.511	−0.127

the last infill round's ten sets, 0.4–0.6 at $K = 36$, applied to the optimum's prediction gives predicted-to-measured 0.95–1.19 for every arm. Ten measured sets at the destination calibrate the eleventh; this is a reporting rule, and we state it as one.

Is the task gain free? (tested, negative.) The tensor-train set was selected for the task; what it costs elsewhere was never measured. Selecting one 36-head set for IOI with the bagged lower bound on a tensor train pooled over the eight prompts (no infill) and removing those heads at every position, IOI accuracy on 120 prompts with fresh names and templates stays at 100% and the margin triples (+3.6 → +10.4), the transfer the surrogate was built to deliver. General next-token loss on FineWeb rises from 3.287 to 3.608 nats, +0.32 or 10%, more than twice the cost of removing 36 heads at random from the whole model (+0.16) and only somewhat less than removing 36 at random from the pool (+0.41); singles-greedy on the same pool costs +0.27. The pool of heads least harmful to the task is biased toward heads that carry general functions — which is why their removal sharpens the task — so the specialisation is a trade of general ability for task margin, and Path-Space does not find structure that is free to remove.

Route-localised editing (tested, negative). If the per-context gradient tells where a target's route is most sensitive, editing there should flip the answer with the smallest dose. We pre-registered one-cell activation edits αu at the last position on 42 factual swaps (Qwen3-0.6B) and 30 IOI flips (pythia-410m), with the minimal flipping dose found by bisection and collateral measured as the KL over the vocabulary with the two answers removed (Table 12). The cell with the largest gradient norm is block 0 in 41 of 42 factual trials and never flips the answer at four times the residual norm; on IOI it is blocks 10–13 and fails in 22 of 30. A pre-registered amendment that walks blocks in gradient order and accepts the first whose check

Table 12: One-cell small-dose editing. Fail = no flip at four times the residual norm; dose = edit norm over residual norm at flip; KL = collateral over the vocabulary with the two answers removed. Gradient direction unless stated.

	Qwen3-0.6B, 42 factual swaps			pythia-410m, 30 IOI flips		
arm	fail	dose	KL	fail	dose	KL
largest gradient norm at p (block 0 / blocks 10–13)	42	—	—	22	1.15	1.89
largest gradient norm over the lattice	40	0.26	0.22	24	1.24	1.89
curvature-verified route (block 18 / 14–17)	39	0.85	3.70	10	0.18	0.16
band, block 16	41	1.75	7.68	5	0.18	0.08
logit-lens layer (largest direct write; 20–27 / 11, 18)	20	0.13	0.19	7	0.10	0.018
identity direction at the verified cell	10	0.50	0.43	0	0.28	0.046
random direction at the verified cell	42	—	—	30	—	—

gradient at the predicted dose agrees with the base gradient chooses block 18 (factual) and blocks 14–17 (IOI) and still fails in 39 of 42 and 10 of 30, with the actual dose 2–7× the first-order prediction. The cheapest and cleanest edits are at the late blocks where the answer is written and the identity reading is valid: dose 0.10–0.13 of the residual norm, collateral KL 0.02–0.19, flipping 52–77% of trials; and at any cell the plain unembedding direction flips more reliably than the per-context gradient direction. Sensitivity is not editability: the cells the route ranks highest are exactly the curved ones (§4.4), where a finite edit is absorbed. Reading what a write did needs the per-context transport because early effects are mediated; writing a new signal is easiest where the readout is direct and linear. The route's one use in editing is as a negative selector.

Edit design in the linear regime (measured, modest). At small dose the per-context prediction of an edit's effect is exact (Spearman 0.996 across layers, error 4% at $\alpha = 0.05$; 0.96, 21% at $\alpha = 0.25$), so the effect of a family of small edits on a chosen logit difference can be screened by gradients without forward passes, and the layer at which a small edit is amplified least (the late layers stay linear to full dose) can be read off Fig. 9b. This is a productivity use of the object — cheap screening of small interventions — and not a capability gain in itself.

Diagnosis-driven targeted fine-tuning (tested, negative: any four blocks fix the failure). If the route says which blocks build a wrong answer, training only those blocks should fix more failures with less forgetting than training blocks chosen by the identity reading or by a fixed schedule. Finding a failure set took three tasks: pythia-410m fails 0.4% of IOI prompts and 0.9% of subject–verb agreement prompts with an attractor, and 62% of in-context number comparisons ("Which is bigger, 58 or 70? Answer:", four templates; 73% when the larger number comes second). On 200 failing prompts the four blocks with the most negative transported attribution of the margin are $\{7, 10, 11, 17\}$; the identity picks $\{0, 13, 17, 18\}$. Fine-tuning four blocks for 150 steps (Adam 2×10^{-5}, batch 8, three seeds) fixes 99% of 200 held-out failures whichever four are trained — route 0.990, identity 0.995, fixed band 10–13 0.995, random 0.995 — with general accuracy 0.95–0.98 and FineWeb loss +0.11 to +0.16 nats, while training all 24 blocks at the same rate fixes as many and forgets catastrophically (+4.0 nats; Fig. 6c). The route's blocks forget less than the identity's (+0.12 vs +0.16; the identity's choice includes block 0) and more than the band's (+0.11), so both pre-registered criteria fail. The restriction to four blocks is the regulariser; which four barely matters, because a failure that any four blocks can absorb has no location to find.

Per-token layer skipping (tested: the ceiling exists, a cheap router does not reach it). The route is per token; inference could skip, at each position, the blocks whose writes that position does not need, if the route could be read without a backward pass. On 400 FineWeb positions of Qwen3-0.6B we measured the single-block skip cost $\delta_k(p) = \tilde{z}(p) - \tilde{z}_{-k}(p)$ for all 25 skippable blocks (blocks 0 and the last two excluded) and evaluated skipping rules on 100 held-out positions (Fig. 6b). An oracle that knows this token's costs and skips the four blocks whose writes oppose the target raises the target logit by 0.79, where the best static schedule (the four blocks with the smallest mean cost over training positions) costs 0.41 and random costs 2.50; the first-order profile from one backward pass reaches most of that ceiling

(−0.47; per-position Spearman between A_k and δ_k is 0.87). A ridge probe from the block-4 residual predicts the profile at Spearman 0.36 against 0.57 for the static mean profile, its per-token part at 0.10, and skips worse than the static schedule at both $m = 4$ and $m = 8$ (+1.27 vs +0.41; +2.75 vs +1.30): the information that makes per-token skipping pay lives in the backward pass, not in the early forward state, and the pre-registered criteria (R1, R2 and their measured-cost forms R1b, R2b) all fail. Two further facts bound the design space. Skipping the blocks with the smallest first-order magnitude $|A_k|$ costs twice the static schedule (1.02 vs 0.49): a rule that selects at the bottom of the first-order ranking selects where the estimate is least reliable in relative terms (§4.4), where a small $\nabla \cdot \Delta$ can hide a large write orthogonal to the current gradient. And the gains are not additive: at $m = 8$ the eight single costs sum to a predicted rise of 1.49 and the joint skip costs 0.72, the many-body effect of Appendix K again, so a per-token rule cannot skip many blocks even with oracle knowledge.

Non-linear and later-layer routers (tested: neither linearity nor read position is the bottleneck). The router above was linear and read block 4. On 3 000 positions of a disjoint FineWeb corpus we trained ridge and two-layer MLP routers (512 units, early stopping on held-out chunks, three seeds averaged) from the residual after block $k \in \{4, 8, 12\}$ to the signed first-order profile of the blocks after k, and tested them on the 400 measured positions (Table 13). The MLP beats the linear probe everywhere — every ridge fit is worse than the constant predictor on held-out chunks (standardised validation MSE 1.24–1.50) while the MLP reaches 1.05, 1.03 and 0.92 at $k = 4, 8, 12$ — and reading later helps a little: the per-token part of the measured profile is predicted at Spearman 0.20, 0.24 and 0.28. A router given only the token embedding matches the block-4 MLP (0.573 vs 0.586 on the full profile; +0.26 vs +0.28 when skipping four blocks), so what the early residual predicts is token type, not context. Skipping four blocks, no router beats the static schedule on its own allowed set (gap to the oracle closed −0.08, −0.01, +0.16 at $k = 4, 8, 12$, against oracles of −0.80, −0.63, −0.11 and static schedules of +0.20, +0.30, +1.12); skipping eight, the MLP closes a third of the gap at $k = 4$ and 8 (+0.32, +0.40) and none at $k = 12$, where eight of thirteen late blocks cannot be skipped by anyone. The training-size curve (500, 1 000, 3 000 positions) rises slowly without saturating (per-token Spearman 0.14, 0.15, 0.20 at $k = 4$; 0.18, 0.26, 0.28 at $k = 12$), so more data would help at a rate that puts a useful router orders of magnitude away. The pre-registered predictability criterion passes on the strength of the static component (0.615 against 0.566 for the static profile alone) and the skipping criterion fails: at this scale the per-token route is carried by the backward pass, not by any cheap forward read we tried.

More data and bigger routers (tested: data helps slowly, capacity does not). Twelve times more labels — 36 000 positions of a third, disjoint FineWeb corpus, one backward pass each — lift the per-token part of the profile a two-layer MLP predicts from 0.22 to 0.28 at block 4 and from 0.22 to 0.34 at block 8, slopes of +0.05 and +0.11 per decade of data (Table 13, lower rows). The block-8 router becomes the first to beat the static schedule when skipping four blocks (+0.18 against +0.30, closing 13% of the gap to the oracle at −0.63) and closes 57% of the gap when skipping eight (+1.49 against +2.29; oracle +0.89). Capacity does not help: a six-block residual MLP (0.27, 0.33) and a one-layer transformer reading sixteen context residuals (0.24 at block 4, no better than a two-layer MLP trained on the same 13 500 labels, 0.23) do not improve on the two-layer MLP at either depth. Extrapolating the block-8 slope, a per-token Spearman of 0.5 would need of the order of 10^6 labelled positions, a million backward passes: data is the lever, and it is a long one.

Writing the route into the forward state (tested, negative: the gain is data, not the objective). If the route is not in the residual, a model can be trained to put it there. We fine-tuned blocks 1–4 of Qwen3-0.6B on the same corpus with the language-modelling loss plus λ times the squared error of a linear head from $h_4[p]$ to the standardised profile of blocks 5–25, at $\lambda \in \{0.3, 3\}$, and re-measured the fine-tuned model's own single-block skip costs on the 400 test positions. The head reaches a per-token Spearman of 0.257, against 0.141 for the linear probe of Appendix K on the untouched model. The gain is not the auxiliary loss. A ridge fitted on the untouched model with the same 36 000 labels reaches 0.260; the same fine-tuning run with $\lambda = 0$, with routers fitted afterwards, reaches 0.243

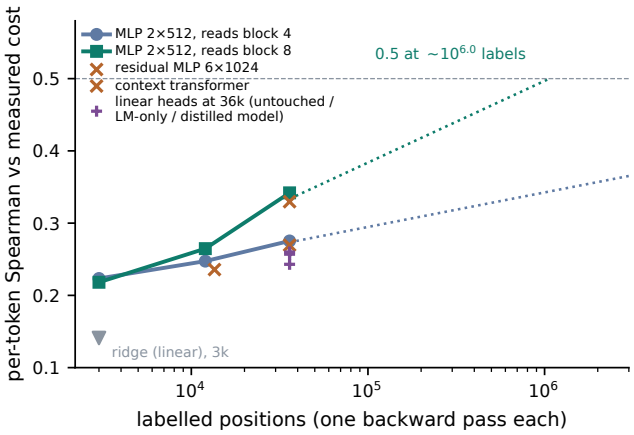


Figure 7: Three remedies for a cheap per-token router on Qwen3-0.6B, on one axis. The y axis is the median over 400 held-out positions of the Spearman correlation between the predicted and the measured single-block skip-cost profile after the static mean profile is removed from both, the part of the profile that is specific to the token. Training data moves it (+0.05 per decade reading block 4, +0.11 reading block 8); extra capacity does not (a six-block residual MLP and a transformer over sixteen context residuals land on the same curve); and neither does training the model to expose its route (three linear heads at 36 000 labels: fitted on the untouched model, fitted after language-model-only fine-tuning, and trained jointly with the fine-tuning objective). The dotted lines extrapolate the two fitted slopes.

Table 13: Routers for per-token skipping on Qwen3-0.6B (Q71, Q72). A router reading the residual after block k may skip blocks $k{+}1$ to 25. Validation MSE is on standardised targets (1.0 = the constant predictor). Spearman is the median over 400 test positions against the measured single-block skip costs on the allowed set; "per-token" removes the static mean profile from both. Skipping is the median target-logit loss over 200 positions when the m most negative predicted (or measured) blocks are skipped; negative means the target logit rises. Gap closed = (static − router)/(static − oracle) for the MLP at $m = 8$. Lower rows: routers trained on 36 000 labels from a third corpus (Q72); the validation MSE of the context router is on its own subset.

	val. MSE		Spearman vs measured cost			skip $m = 4$				skip $m = 8$: MLP / static	
k (allowed)	linear	MLP	linear	MLP	static	linear	MLP	static	oracle	loss	gap closed
4 (21 blocks)	1.50	1.05	0.43	0.59 (per-token 0.20)	0.57	+0.58	+0.28	+0.20	−0.80	+0.92 / +1.18	+0.32
token embedding only	1.62	1.08	0.44	0.57 (0.13)	—	—	+0.26	—	—	+1.03 / —	+0.19
8 (17 blocks)	1.44	1.03	0.46	0.60 (0.24)	0.58	+0.88	+0.31	+0.30	−0.63	+1.73 / +2.29	+0.40
12 (13 blocks)	1.24	0.92	0.49	0.62 (0.28)	0.57	+1.33	+0.93	+1.12	−0.11	+3.72 / +3.08	−4.05
12× more labels: 36 000 positions of a third FineWeb corpus (Q72)											
4 (21 blocks), MLP 2×512	—	0.89	—	0.60 (0.28)	0.57	—	+0.21	+0.20	−0.80	+0.87 / +1.18	+0.37
residual MLP 6×1024	—	0.90	—	0.60 (0.27)	—	—	—	—	—	— / —	—
transformer over 16 context residuals (13.5k labels)	—	0.88	—	0.57 (0.24)	—	—	+0.46	—	—	+1.18 / —	+0.01
8 (17 blocks), MLP 2×512	—	0.88	—	0.65 (0.34)	0.58	—	**+0.18**	+0.30	−0.63	+1.49 / +2.29	**+0.57**
residual MLP 6×1024	—	0.89	—	0.64 (0.33)	—	—	—	—	—	— / —	—

(ridge) and 0.269 (two-layer MLP), and closes 6–13% of the oracle gap at $m = 4$ against the head's 15% — inside the pre-registered attribution margin of 0.10 in Spearman and 0.2 in gap closed, so the objective contributes nothing measurable and the whole improvement over Appendix K is the twelvefold increase in labels. Both fine-tuning runs also lowered held-out loss by 0.22 nats (a post-trained model drifting toward web text in two epochs) and flattened the route: the static schedule costs +0.37 instead of +0.20 and the oracle falls from −0.80 to −0.52, so the adapted model is less skippable than the one we started with. Of the three obvious remedies for a cheap router — more data, more capacity, training the model to cooperate — only the first moves the number, at +0.11 Spearman per decade of labels (Fig. 7).

L What did not work

Four pre-registered directions closed negative and are reported because they bound the claims. Token-level Path-Space. The same construction applied to the averaged lens (a per-token profile $\pi_\ell(t) \propto \|D_\ell^\top g_t\|^2$ over depth) is reproducible across independent half-lenses

($r = 0.89$), class-structured (content nouns are built 0.36 blocks later than function words, $d \approx 1$) and carries information beyond unembedding geometry, but its token-specific part correlates with the per-context route at $r = 0.04$–0.09: it is a property of the lens, not of the model's computation on a context. Steering guidance. Choosing the eight layers of a concept-swap edit by the route's predicted effect does not change whether the swap lands (six layer-selection arms all flip 50–58% of trials with overlapping intervals); at that dose the edit is outside the linear regime (Fig. 9b) and breakage decides. The chain as a codec. Fitting the averaged lens as a low-rank matrix-product chain, with alternating least squares and environment-whitened truncation transplanted from tensor-network optimisation (Schollwöck, 2011; Orús, 2014), beats independent per-layer SVD at low rank but is Pareto-dominated as a compression of the lens by scalar quantisation with alternate-layer storage; Magnus resummation of the averaged chain is non-perturbative. The answer's emergence layer. On sixteen factual prompts the identity and the per-context transport agree on the block where a one-token answer peaks (blocks 20–24 in 9 of 16), and ablation confirms it: the "answer emerges late" reading of the logit lens is correct, because it concerns the late, direct part of the route where all predictors coincide; the disagreement of §4.2 lives in the sign and ranking of early and middle writes.

M Statistical pitfalls the nulls caught

Table 14: Artefacts that produced spurious positives during this work, each detected by a null control and fixed at the source.

artefact	symptom	fix
rank correlation with index-order tie breaking on a binary indicator	label-permutation null at +0.288 instead of 0	average ranks, or OLS on raw values
float32 least squares at $n \approx 2800$	shuffled data gave a partial correlation of +0.30	float64 / QR
RMSNorm zero-homogeneity, $\nabla \cdot h = 0$	full-gradient DLA sums identically zero	freeze the final scale
differencing shared estimates (D_ℓ, $D_{\ell+1}$ share $J_{\ell+1}$)	spurious adjacent anticorrelation under noise	cross-half estimator
a single permutation	a $+4\sigma$ effect misjudged as marginal	repeat the permutation
shared monotone trend in "X against distance"	$\rho = 0.74$–0.98 from co-monotonicity	detrend first
band-limited baseline	a +0.06 "gain" from weakening the baseline	full-depth rerun
absolute sign-agreement thresholds	fail when the base rate is 0.82	state criteria relative to the null

N What is literal and what is borrowed

The backward pass is the MPO contraction (§3.1); the path sum and its classes are exact combinatorics of the linearisation (§3.2); interference is a statement about real vector sums (Appendix G). The tensor-network optimisers (alternating least squares, environment-whitened truncation) and the resummation of the path sum (Magnus) that the vocabulary suggests were tried as ways to compress the averaged lens and closed negative (Appendix L); what survived is the contraction, the classes and the measurements built on them.

O Limitations in full

The five models span 0.36B to 70B parameters in four families (Qwen3, Pythia, SmolLM2, Llama-3); "shared" throughout means "the same in the models we tested", never universal. The scale runs cover the predictor comparison, the early sign, the route, mediation, second order and, at 8B in fp32, the quadrature; the 70B run is bf16, which moves the 8B main-phase numbers by at most 0.013 on the frozen-norm logit and 0.004 on log p with an fp32 readout, breaks the quadrature's exactness, so the quadrature is not tested at 70B,

and adds rounding noise to measured pairwise interactions comparable to their size (Appendix E). The first-order transport is right in order and wrong in size below block 18; the quadratures that fix the size cost forward+backward passes per block. The second-order estimator needs the single effects to improve pair predictions; it has been validated at block-write granularity on FineWeb positions, at head granularity on IOI in three architectures against the additive baseline, on same-layer head pairs (Spearman 0.95 between estimated and measured interaction, pair error cut 79%) as well as cross-layer ones, and on AtP*'s activation-patching footing (§4.7); that comparison is on one model and ten prompt pairs, a placement rather than a replication of their distribution-level study. The interference statistics are descriptive. The decompositions of §3 are identities for the frozen-norm logit; the model's own logit also moves by the common-mode term of Eq. 14, which an attribution taken on $\tilde{z}$ does not carry, and the true targets are scored with their own gradients (Table 1); the own-gradient comparison covers all five models and the interaction reliability analyses the three small models and Llama-3-8B. The second-order estimator is a pair-level object: as a guide to removing sets of components it fails (Appendix K), because redundancy among many components is not pairwise. The per-context gradient is an attribution instrument and not a steering one: as a guide to where to edit it fails (Appendix K), because the most sensitive cells are the most curved. The set-effect tensor train is validated on one model and one task (pythia-410m, IOI, 8 prompts, a pool of 48 heads); its optimum remains over-predicted 2× at $K = 36$ before the infill correction, and a cost-matched measured greedy still selects better sets at $K = 36$. The capability routes of Appendix K (task-specialised pruning, targeted fine-tuning, per-token skipping) were each tested on one model and one task at this scale; they bound what the object buys here and do not rule the routes out at scale; small non-linear routers at three depths, deeper and context architectures, a twelvefold increase in training labels and an auxiliary distillation objective were tested (Appendix K); the best closes 57% of the gap to the oracle at eight skipped blocks and 13% at four, and the measured slope in training data is +0.11 Spearman per decade. Two pre-registered thresholds in §4.2 were stated in absolute terms and failed in replication on base-rate calibration; they are reported relative to their nulls, and the substance holds. Gradient×write is attribution patching (Nanda, 2023; Syed et al., 2024); integrating it along a path is integrated gradients (Sundararajan et al., 2017) and, in this literature, EAP-IG (Hanna et al., 2024); the direct/indirect decomposition is standard causal mediation. None of those is claimed here. What is new is the shared-cost second-order estimator of §4.5 — one backward pass per component for that component's interaction with every other, within a layer and across layers — and, as measurement rather than method, the quantification of where each reading of a write fails, on five models in four families, against removal, with the readings' own objectives implemented (§4.1) and with the target of the measurement decomposed (Eq. 14).

P Practical guidance

For attributing a prediction to blocks or heads in the first two thirds of a stack, use the per-context gradient (attribution patching) and not the direct projection: the projection's ranking and sign are at chance there in every model we tested, and no averaged transport repairs it. Take the gradient where the write enters the stream (the attention output for a head in a sequential block), and take the gradient of the target you want to explain: freezing the final norm's scale makes the frozen-norm logit linear in the last residual, but the transported attributions then sum to it only up to the remainder of Eq. (13), and for the model's own log-probability that target's own gradient is the right one. Trust first-order magnitudes only from the last third of the stack or for small perturbations; for early blocks, a secant at $s = 0.1$ recovers a third of the missing magnitude and six Gauss–Legendre nodes nearly all of it (median error below 10^{-4}, worst case 0.18). Exclude block 0 from any linearised account; it is a regime boundary. For pairwise interactions among the components that matter, one backward pass on each single-component-ablated run predicts all interactions with that component; combine with the measured singles, not with first-order singles. Do not select components at the bottom of a first-order ranking (the least attributed blocks to skip, the least harmful heads to remove): that is where the estimate is least reliable in relative terms and where cancellation hides large writes; a static schedule

Table 15: Scale replication (Q73). Top: the headline comparison on five models. Bottom: the pre-registered criteria at scale. Llama-3-8B in fp32 (second order and quadrature on 20 and 10 positions), Llama-3-70B in bf16 (120, 10 and 8 positions); an 8B bf16 run matches the fp32 one to within 0.013 on every main-phase number.

model	blocks	d	per-context	identity	gap
Qwen3-0.6B	28	1024	0.808	0.437	+0.37
pythia-410m	24	1024	0.787	0.385	+0.40
SmolLM2-360M	32	960	0.839	0.384	+0.46
Llama-3-8B	32	4096	0.718	0.350	+0.37
Llama-3-70B	80	8192	0.829	0.237	+0.59

	criterion	Llama-3-8B	Llama-3-70B
S1	gap ≥ 0.25, CI excludes 0	+0.368 [+0.334, +0.398]	+0.592 [+0.562, +0.619]
S2	first third: identity ≤ 0.20, per-context ≥ 0.50	-0.020 / 0.534	0.015 / 0.692
S3	early sign: identity ≤ 0.58, per-context ≥ 0.70	0.561 / 0.800 (base rate 0.738)	0.549 / 0.811 (base rate 0.655)
S4	early attribution off-target ≥ 0.70	0.80	0.84
S5	mediated share ≥ 0.80 at every early block	fail: 0.74–1.00	fail: 0.75–0.99
S6	second order: Spearman ≥ 0.5, pair error cut $\geq 30\%$	0.695, 57%	0.641, 30.2%
S7	GL-6 relative error ≤ 0.05 at blocks ≥ 1 (fp32 only)	0.000 (block 0: 1.25)	not testable in bf16

Algorithm 1 Path-Space for one target (p, t): lattice attribution and second-order interactions.

clean forward; freeze $\mathrm{rms}(h_{L-1}[p])$; record $h_k[q]$ and the writes $\Delta_k[q] = h_k[q] - h_{k-1}[q]$ for all k, q
one backward pass of $\tilde{z} = \tilde{g} \cdot h_{L-1}[p]$ rooted at the embedding output $\Rightarrow \nabla_k[q]$ for all k, q ▷ contracts the MPO (6) with $\tilde{g}$
classes $c_k[q] \leftarrow \nabla_k[q] - \nabla_{k+1}[q]$; attributions $A_{k,q} \leftarrow \nabla_k[q] \cdot \Delta_k[q]$; check $\sum_{k,q} A_{k,q} + \sum_q E_q \approx \tilde{z}$
interference $V_p \leftarrow \| \sum_k c_k[p] \| / \sum_k \| c_k[p] \|$; identity reading $A^0_{k,q} \leftarrow \tilde{g} \cdot \Delta_k[q]$ for comparison
for each block k whose interactions are wanted do
    forward with $h_k[p] \leftarrow h_{k-1}[p]$; record writes $\Delta^{(k)}_j$; one backward pass $\Rightarrow \nabla^{(k)}_j[p]$ for $j > k$
    $\hat{I}(k, j) \leftarrow \big[-\nabla^{(k)}_j \cdot \Delta^{(k)}_j \big] - \big[-\nabla_j \cdot \Delta_j \big]$ for all $j > k$ ▷ L backward passes for $L^2/2$ interactions
end for
return $\{A_{k,q}\}$, $\{c_k[q]\}$, V_p, $\{\hat{I}(k, j)\}$; the single effects δ_k by L forwards when magnitudes are needed

by mean measured cost beats it, and a set-effect model with an uncertainty bound is needed once more than a few components are removed together. Any task-specialised removal should be reported with its collateral on a general corpus next to random removal of the same size; the task gain in Appendix K was twice as costly as random.

Q Supporting tables and figures

These are referenced from the main text and held here for space.

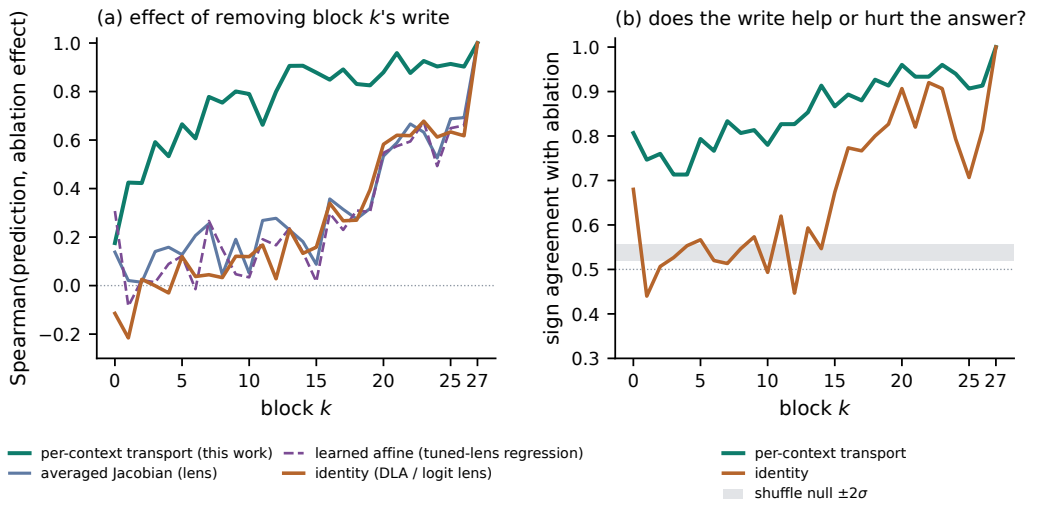


Figure 8: Qwen3-0.6B, 150 positions. (a) Per block, the Spearman correlation across positions between each predictor and the measured ablation effect. The identity, the averaged Jacobian and the learned affine transport carry no information about blocks 0–15 and become usable only from block 19–20; the per-context transport is informative from block 1 and above 0.8 from block 7. (b) Whether a write helps or hurts the answer: sign agreement with the ablation. The identity's sign sits inside its shuffle null through block 15.

Table 16: Pairwise interactions $I(k, j) = \delta_{k,j} - \delta_k - \delta_j$ from L backward passes. Spearman between predicted and measured interaction (within-position shuffle null in parentheses), sign agreement, and the median relative error of the pair effect $\delta_{k,j}$ from the two single effects alone versus with the predicted interaction added.

run	pairs	Spearman (null)	sign	error: singles $\to +\hat{I}$	additive AtP
Qwen3, pos. 0–40 (pre-reg.)	adjacent, 1 080	0.600 (0.02)	0.79	$0.389 \to 0.229^{\dagger}$	0.569
	non-adjacent, 600	0.855 (0.07)	0.88	$0.122 \to 0.041^{\dagger}$	0.321
Qwen3, pos. 40–80 (confirm.)	adjacent, 1 080	0.586 (0.01)	0.78	$0.351 \to 0.237$ (−33%)	0.556
	non-adjacent, 600	0.873 (0.10)	0.89	$0.119 \to 0.041$ (−66%)	0.350
pythia-410m, 40 positions	adjacent, 920	0.613 (0.04)	0.76	$0.257 \to 0.154$ (−40%)	0.497
	non-adjacent, 600	0.744 (0.06)	0.81	$0.114 \to 0.050$ (−56%)	0.375
Llama-3-8B, fp32, 20 positions	all $k < j$, 9 920	0.695 (0.02)	0.81	$0.142 \to 0.061$ (−57%)‡	—
Llama-3-70B, bf16, 10 positions	sampled, 4 800	0.641 (0.01)	0.73	$0.063 \to 0.044$ (−30%)‡	—

†post hoc on the first run; the estimator was then pre-registered and confirmed on the two runs below.
‡median absolute pair error in logits (Q73); the rows above report median relative error.

R Pre-registration ledger

Every experiment in this paper was specified with pass/fail criteria and null controls in a design document committed to the repository before it was run; results were written next to the criteria. The ledger, with the experiment identifiers used in the repository, is: transport-vs-identity ablation validation (Q50; replications Q50b, lens baseline Q50c, learned affine Q55); context routes and sum identities (Q49, Q49b); lattice route map and off-diagonal ablation (Q51); response decomposition and quadrature (Q52); route-guided steering and dose response (Q53, Q53b); cross-architecture structure (Q54); factual-prompt emergence (Q56); sign of early writes (Q57, Q57b); second-order attribution patching (Q58; confirmation Q58b; pythia Q58c); scale replication on Llama-3-8B and 70B (Q73, seven criteria, S5 negative at both scales); true behavioural targets with their own gradients, the log-probability line integral and the reliability of the interaction estimator (Q74, six criteria: T3 negative in pythia, T6 negative at 24–32 blocks); averaged-transport convergence (Q75); bf16 precision and the fp32 readout (Q76, four criteria: the control partial, R1 narrowly missed); interaction-aware pruning (Q61, negative); route-localised editing (Q62, with a pre-registered amendment, negative); the set-effect tensor train (Q63, amendment

Table 17: Spearman correlation between a predictor and the measured effect of removing a block's write at one position, over all (position, block) cells; gaps carry 95% bootstrap intervals over positions. On Qwen3-0.6B (seed 1) the averaged Jacobian lens scores 0.530 and the learned affine transport 0.486 (per-block $|\rho| \leq 0.31$ in the first third for both).

model (corpus)	cells	per-context	identity	gap [95% CI]	identity per-block $\lvert\rho\rvert$, first third
Qwen3-0.6B (seed 1)	150×28	0.808	0.437	$+0.371$ $[+0.34, +0.40]$	≤ 0.23
Qwen3-0.6B (seed 0)	60×28	0.827	0.437	$+0.390$ $[+0.35, +0.44]$	≤ 0.25
pythia-410m (seed 1)	100×24	0.790	0.354	$+0.436$ $[+0.40, +0.47]$	≤ 0.24 (block 0: 0.38)
SmolLM2-360M (seed 1)	100×32	0.848	0.397	$+0.452$ $[+0.42, +0.48]$	≤ 0.21
Llama-3-8B, fp32 (seed 1)	150×32	0.718	0.350	$+0.368$ $[+0.33, +0.40]$	≤ 0.17
Llama-3-8B, bf16 (seed 1)	150×32	0.717	0.349	$+0.368$ $[+0.33, +0.40]$	≤ 0.17
Llama-3-70B, bf16 (seed 1)	120×80	0.829	0.237	$+0.592$ $[+0.56, +0.62]$	≤ 0.20

Table 18: Indirect-object identification at head granularity: 30 prompts each model gets right, metric $\tilde{z}_{\mathrm{IO}} - \tilde{z}_{\mathrm{S}}$ with the final norm's scale frozen; components are every head at the END and S2 positions and every MLP at END (792 / 924 / 992 per prompt), each zero-ablated. Pairs are sampled among each prompt's 40 most impactful components (recorded before the run: random pairs are dominated by components with no effect). For the sequential blocks of Qwen3 and SmolLM2 the gradient is taken at the attention output.

	pythia-410m (24 L)	Qwen3-0.6B (28 L)	SmolLM2-360M (32 L)
singles: Spearman, AtP (per-context) / identity	0.949 / 0.144	0.883 / 0.145	0.953 / 0.123
top-20 recovery, AtP / identity	0.777 / 0.465	0.583 / 0.282	0.835 / 0.427
identity per-layer Spearman near zero through layer	13 (≤ 0.26)	17 (≤ 0.12)	24 (≤ 0.20)
AtP per-layer Spearman, heads at END	0.86–1.00	0.65–1.00	0.60–1.00
sign agreement, impactful components in the first two thirds, AtP / identity (identity null)	0.965 / 0.540 (0.506)	0.927 / 0.499 (0.501)	0.973 / 0.499 (0.502)
identity "opposes" calls that actually helped	56.9%	54.6%	66.4%
pairs among the top-40 components (1 800 each)			
Spearman($\hat{I}$, I) (shuffle null)	0.770 (0.036)	0.697 (0.066)	0.909 (0.022)
sign agreement	0.86	0.81	0.92
pair error: singles $\rightarrow$ singles $+\hat{I}$	0.093 $\rightarrow$ 0.032 (-66%)	0.115 $\rightarrow$ 0.052 (-55%)	0.077 $\rightarrow$ 0.015 (-80%)
same-position / cross-position pairs	0.774 / 0.757	0.683 / 0.730	0.932 / 0.858
additive first-order (AtP) pair error	0.484	0.543	0.427

Q63b, cost-matched greedy baseline Q63c, deployment Q64: optimiser, cross-prompt transfer, TT-cross; uncertainty-aware surrogate Q65; size-conditioned surrogate and winner's-curse calibration Q66; collateral of task-specialised pruning Q67, negative; diagnosis-driven targeted fine-tuning Q68, whose IOI and agreement failure sets were empty at this scale (Q68, Q68b) and which ran on in-context number comparison (Q68c, negative); per-token layer-skipping router Q70 with measured-cost diagnostics Q70b, negative; non-linear and later-layer routers Q71, predictability criterion passed on the static component, skipping criterion negative; router data scaling, deeper and context routers and distillation of the route into the forward state Q72, with an LM-only control and a matched-data clean-model control added by amendment, negative); head-level IOI benchmark (Q59 pythia, Q59b Qwen3, Q59c SmolLM2, with a pre-run amendment to the pair sampling); AtP* comparison in the activation-patching setting (Q60); review-response checks: the sum rule's exact remainder (R1), same-layer interactions (R2), the tuned lens under its own objective (R3), the true-logit arbiter and its common-mode decomposition (R4, R4b), the leak-free cross-prompt transfer (R5), and surrogate baselines at matched budget (R6); token-level Path-Space (Q48, Q48b); depth-chain codec (Q41–Q47). Criteria that failed and criteria that were mis-stated are reported as such in the text.

Table 19: Does a write help or hurt the answer? Sign agreement with the ablation for early blocks (blocks 1–15 of 28, 1–13 of 24, 1–17 of 32, 1–17 of 32 and 1–45 of 80), each predictor against its own within-block shuffle null, and the precision of each predictor's "opposes the answer" calls against the base rate of harmful writes (permutation p for informativeness).

	Qwen3-0.6B	pythia-410m	SmolLM2-360M	Llama-3-8B	Llama-3-70B
base rate of helpful early writes	0.654	0.824	0.659	0.738	0.655
sign agreement, per-context (− null)	0.801 (+0.267)	0.792 (+0.169)	0.841 (+0.306)	0.800 (+0.189)	0.811 (+0.241)
sign agreement, identity (− null)	0.541 (+**0.004**)	0.617 (+**0.018**)	0.566 (+**0.020**)	0.561 (+**0.003**)	0.549 (+**0.005**)
identity "opposes" calls	857	453	611	1 040	2 108
of which actually helped	64.8%	79.7%	63.0%	71.7%	63.6%
precision vs. base rate (p)	0.352 vs 0.346 (0.37)	0.203 vs 0.176 (0.046)	0.370 vs 0.341 (0.039)	0.283 vs 0.262 (0.42)	0.364 vs 0.345 (0.21)
per-context "opposes" precision (p)	0.672 (0.000)	0.450 (0.000)	0.718 (0.000)	0.602 (0.000)	0.700 (0.000)

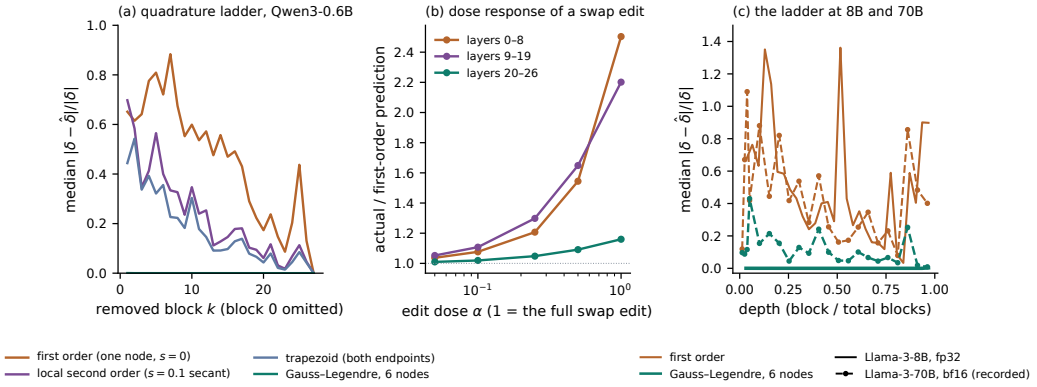


Figure 9: (a) Qwen3-0.6B, 40 positions: the median relative error of four estimates of the ablation effect, each a quadrature of the same line integral (11). First order (the gradient at the clean point) misses 55–90% of early effects; a secant at $s = 0.1$ removes a third of that; the two endpoints half; six Gauss–Legendre nodes nearly all of it (median below 10^{-4}, worst case 0.18). Block 0 (omitted) is a regime boundary at which every rule fails. (b) Dose response of a concept-swap edit applied at one layer: the ratio of actual to first-order-predicted change in the logit difference, by layer band, as the edit is scaled. Late layers respond linearly at full dose; early and middle edits are amplified 2–2.5× above the linear prediction. (c) The ladder at scale, by depth: first order and six Gauss–Legendre nodes on Llama-3-8B in fp32 (solid; 10 positions, every block) and Llama-3-70B in bf16 (dashed; 8 positions, 24 blocks). In fp32 the six-node median error is at most 0.003 at each block after block 0 (block 0: 1.25, omitted); in bf16, rounding holds them at 0.01–0.4, so the 70B curve is recorded, not tested.

Table 20: AtP*'s setting reproduced on pythia-410m: ten IOI prompt pairs from Kramár et al. (2024)'s template and names, metric $-\log p$(clean target), nodes = query, key, value and pre-projection output of every head plus every MLP at the five positions whose activations differ between the prompts (7 800 per pair), ground truth by activation patching (one forward per node). Medians over pairs. "Verify cost" is the number of nodes checked in estimator order before all true top-K nodes are found.

estimator	Spearman, all nodes	q	k	v	out	MLP	verify cost top-5 / 10 / 20
AtP	0.950	0.94	0.91	0.95	0.98	0.94	50 / 58 / 89
AtP + GradDrop	0.941	0.93	0.90	0.94	0.96	0.95	37 / 42 / 76
AtP + QK fix	0.956	0.95	0.92	0.95	0.98	0.94	26 / 27 / 58
AtP* (both)	0.945	0.94	0.91	0.94	0.96	0.95	20 / 26 / 54
direct effect (path patching to the logits), END out/MLP nodes	0.765	(AtP on the same 408 nodes per pair: 0.987)					
second-order $\hat{I}$, 600 pairs of top-40 nodes	Spearman($\hat{I}$, I) 0.875 (null 0.012), sign 0.91; pair error 0.134 → 0.032 given the singles (−76%)						

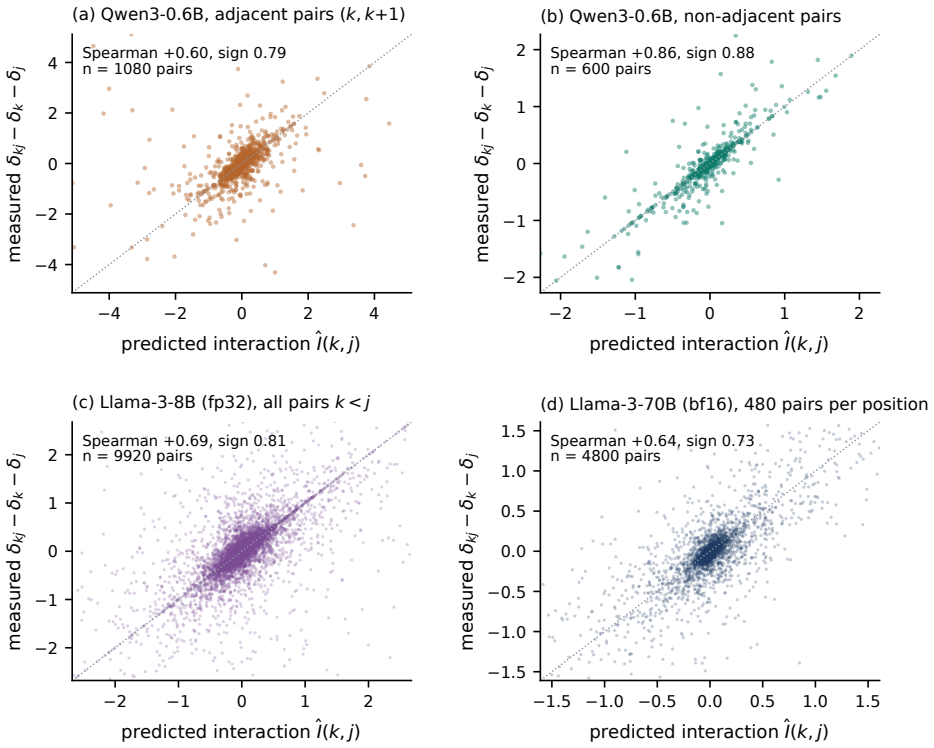


Figure 10: Predicted against measured pairwise interaction. (a, b) Qwen3-0.6B, positions 0–40, adjacent and non-adjacent block pairs; (c) Llama-3-8B in fp32, all 9 920 pairs $k < j$ at 20 positions; (d) Llama-3-70B in bf16, 4 800 sampled pairs at 10 positions. The prediction comes from one backward pass on the run with block k removed, L backward passes for all $L^2/2$ pairs.

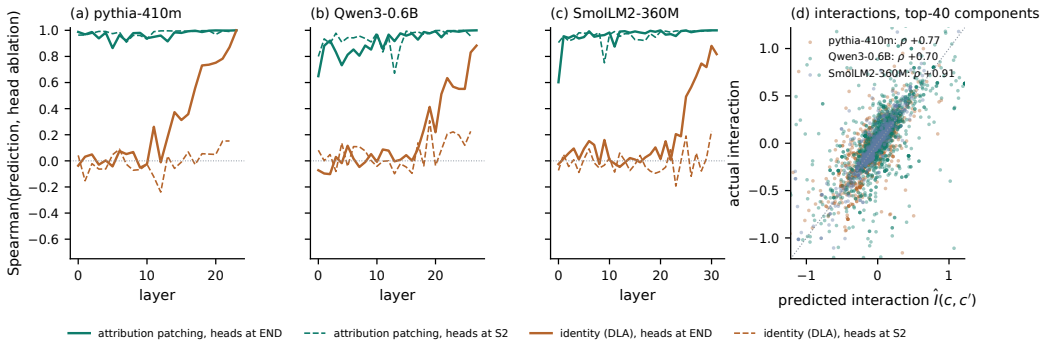


Figure 11: IOI at head granularity in (a) pythia-410m, (b) Qwen3-0.6B and (c) SmolLM2-360M: per layer, the Spearman correlation across 30 prompts × heads between each predictor and the measured effect of zero-ablating the head at END (solid) or at S2 (dashed). The identity reading is uninformative through layer 13 of 24, 17 of 28 and 24 of 32; attribution patching is above 0.6 at every layer. (d) Predicted against measured pairwise interaction for 1 800 cross-layer pairs of impactful components per model.