Reading Multiplicative Circuits in Gated Transformer MLPs

Anonymous authors
Paper under double-blind review

Abstract

Transformer MLPs are usually read through activations, sparse feature dictionaries, or logit-space projections. These views underuse the multiplicative structure of gated feed-forward blocks, where the native unit of computation is an interaction between two sources. Circuit methods built on activation patching estimate a patch's effect with a gradient, which is linear in the patch, so whatever two sources contribute jointly is assigned zero. We measure that joint contribution with a 2×2 ablation and find it runs from 0.12 to 1.19 times the size of the main effects those methods recover accurately. We then introduce behavior-conditioned integrated-Hessian tomography, a weight-native method that reads which residual-stream sources interact inside a pretrained gated MLP. Decomposing the normalized MLP input into named upstream sources decomposes the output into pairwise source interactions, and the decomposition is exact: the terms sum back to the MLP's true output to within one part in ten million. It predicts the measured interaction at Pearson correlation 0.95 to 0.97 over 1152 pairs at four layers. The same construction applies to attention scores, which are already exactly bilinear and need no path integral, so the method covers two of the three computational units of a transformer block. Patching a pair's write moves the next-token log-probability by the predicted amount, at slope 0.98 with $R^2 = 0.91$. The terms that compose the MLP output are about three orders of magnitude larger than the output itself, with output cancellation growing as the square of input cancellation, so ranking pairs by the magnitude of their write recovers which sources are large. Projecting each write onto the layer output gives an exact signed attribution that accounts for the cancellation and ranks pairs independently of their size.

1 Introduction

Mechanistic interpretability has made progress by tracing attention paths (Elhage et al., 2021; Olsson et al., 2022), decomposing activations into sparse features (Cunningham et al., 2023; Bricken et al., 2023), and building causal graphs over model components (Conmy et al., 2023; Ameisen et al., 2025). Feed-forward blocks remain harder to read in their native form. Modern decoder-only transformers inherit the residual-stream architecture (Vaswani et al., 2017), and many use gated feed-forward variants (Shazeer, 2020) whose activations multiply two learned projections. The native unit inside the MLP is therefore an interaction between sources.

This matters because many behaviors appear to require a gate-like computation. One source may retrieve candidate evidence while another supplies schema, phase, or local suffix information. A downstream write appears when those signals are multiplied. Linear feature readouts and vocabulary projections identify useful correlates, and they leave the source-pair structure of the computation implicit.

Current circuit tooling measures this structure indirectly. Activation patching and its first-order accelerations assign a number to a component or an edge (Conmy et al., 2023; Syed et al., 2024), and an estimate that is linear in the patch predicts that removing two sources together does what removing each of them separately does. Whatever the two sources

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.

contribute jointly is assigned zero. We measure that joint part and find it the same size as the separate effects when the sources are fully removed, which is the gap this paper addresses.

Bilinear MLPs offer a cleaner mathematical object (Sharkey, 2023; Pearce et al., 2024; 2025). In the static case a scalar output can be written as $y_c(x) = x^\top Q_c x$, so reading the weights becomes a matrix analysis problem. Pretrained language models usually use gated nonlinearities such as SiLU, which make the global map input dependent. We ask whether the same quadratic-form idea can be used locally inside an unmodified pretrained transformer.

It can. A bias-free Llama-style MLP satisfies $F(0) = 0$ and $J_F(0) = 0$. Along the path from the origin to a normalized residual input, its output can be written as an integrated Hessian, in the sense of path attribution (Sundararajan et al., 2017). Decomposing the input into named upstream sources decomposes the MLP output into pairwise source interactions. For a downstream behavior scalar these interactions become quadratic forms over source subspaces.

The method works with any additive split of the input. The identity needs only $x = \sum_i x_i$, so the parts can be attention and MLP writes, token positions, or dictionary features, and the same machinery reads all of them.

Our contributions are:

- A source-pair integrated-Hessian decomposition for gated transformer MLPs, exact to one part in ten million, with pure bilinear replacement as its constant-Hessian special case.
- The same decomposition for attention scores, which are exactly bilinear and need no path integral, extending the method to a second locus.
- A measurement of what first-order circuit methods leave out. The interaction of two sources, defined by a 2×2 ablation, is assigned zero by any estimate linear in the patch, and it runs from 0.12 to 1.19 times the size of the main effects those methods recover accurately. The decomposition predicts it at Pearson 0.95 to 0.97 over 1152 pairs at four layers.
- A calibration of the decomposition against intervention, together with an account of what such an intervention shows. Patching a pair's write moves the next-token log-probability by the predicted amount at slope 0.98 with $R^2 = 0.91$, against a corpus readout with no chosen target. Since the first-order response to any perturbation at the MLP output is its projection on the readout gradient, a matched random write reproduces the effect whenever the attribution is correct, and establishing that two sources were multiplied calls for an intervention on the sources themselves, which we run at both loci.
- A way to rank the terms. The pair writes are much larger than the output they sum to, by factors of 724 to 2868 for the component decomposition, and that ratio grows as the square of the corresponding ratio at the input. Sorting pairs by the size of their write reproduces the ranking of the sources by size. Projecting each write onto the layer output gives an exact signed attribution whose ranking is independent of that one.

2 Related work

Circuits found by patching

Activation patching replaces one activation with its value on a second input and reads the change in a metric (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024). Path patching restricts the substitution to a chosen set of paths (Goldowsky-Dill et al., 2023), and ACDC searches over edges for a subgraph that carries a behavior (Conmy et al., 2023). Attribution patching replaces the substituted forward pass with a first-order estimate, which reduces the cost to one backward pass and scales the search to whole models (Syed et al., 2024).

Each of these assigns a number to a component or an edge, and the estimate is linear in the patch, so the predicted joint effect of perturbing two sources is the sum of their separate effects and the interaction between them is zero for every pair. Section 5 measures that interaction and finds it comparable in size to the main effects the same methods recover accurately. Recent work improves the fidelity of the first-order estimate for a single patch with a Hessian-vector-product correction, and attributes the remaining error to non-linearity downstream of the patched site (Zhang & Wang, 2026), matching what we find in Section 6. The interaction is a property of a pair, and a score attached to one component leaves it out.

Interaction attribution

Integrated gradients attributes a prediction to input features by integrating along a path from a baseline (Sundararajan et al., 2017). Integrated Hessians extends this to pairs of features and supplies the identity we build on (Janizek et al., 2021). That work attributes a scalar output to pairs of input features for general architectures. We specialize it to the gated MLP, where the origin is the baseline the identity asks for, the attributed object is a vector in the residual stream, and the pair terms sum to the sublayer output. The specialization also yields a closed form in which context reaches each channel through two scalars, which is what makes the corpus-scale measurements in Section 8 affordable.

Reading MLP sublayers

Feed-forward layers admit a key-value reading in which the first matrix detects a pattern and the second writes an associated vector (Geva et al., 2021). Sparse autoencoders and transcoders supply a sparse basis whose units are easier to name (Bricken et al., 2023; Cunningham et al., 2023; Dunefsky et al., 2024; Ameisen et al., 2025). Our decomposition is defined for any additive split of the MLP input, so a basis from that line of work can serve as the sources; we report the component and token-position splits here.

Gated and bilinear structure

Gated linear units are standard in current language models (Shazeer, 2020). Replacing the gate with a bilinear form makes a sublayer analyzable from its weights alone (Sharkey, 2023; Pearce et al., 2024; 2025). We keep the trained non-linearity and recover a bilinear object at each input, so the analysis applies to pretrained models with no retraining or substitution.

3 Method

3.1 Source-pair tomography

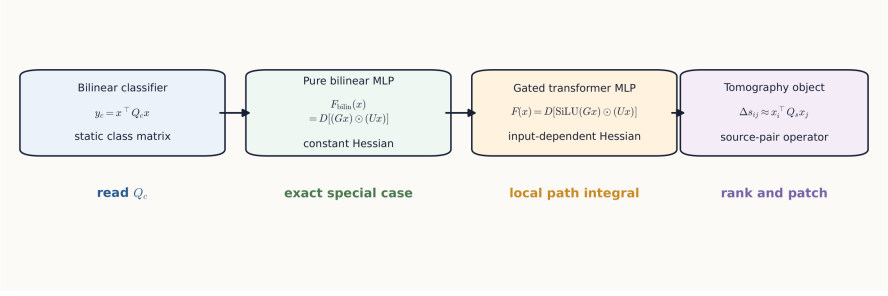

Figure 1: Bilinear MLPs expose a static quadratic form. Gated transformer MLPs expose a local quadratic form through an integrated Hessian. Behavior conditioning turns that form into a source-pair operator that can be ranked, decomposed, and patched.

Write the target MLP input as a sum of named residual writes, $x = \sum_i r_i$, obtained by decomposing the normalized residual stream through RMSNorm. The integrated-Hessian

identity then gives

$$F(x) = \sum_{i \le j} \Delta F(r_i, r_j), \tag{1}$$

with each term computed from the weights and the input alone. For a behavior scalar s with gradient $g = \partial s / \partial F$ at the target position, the contribution of a pair is $g \cdot \Delta F(r_i, r_j)$, and the pairs can be ranked by this quantity.

3.2 The channel form

Since $\mathrm{SiLU}(z) = z\,\sigma(z)$, a gated MLP satisfies

$$F(x) = W_{\text{down}}\big(\, g_x \odot u_x \odot \sigma(g_x) \,\big), \qquad g_x = W_{\text{gate}} x,\ u_x = W_{\text{up}} x. \tag{2}$$

Every input-dependent quantity in the operator reduces to two scalars per hidden channel. This makes all pairs at a token computable in one matrix product, which is what makes the corpus-scale measurements in Section 8 affordable.

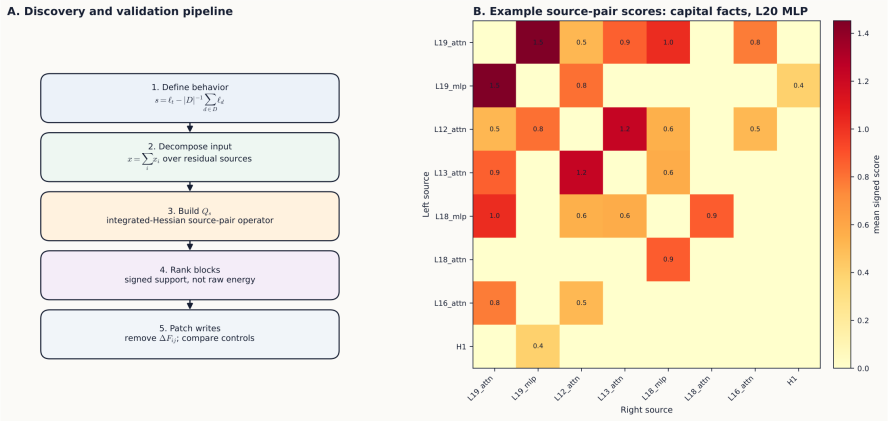

Figure 2: Behavior-conditioned tomography. A target margin defines the scalar objective, the normalized MLP input is decomposed into named sources, and the resulting source-pair operator is ranked and causally tested on held-out examples.

4 Exactness at both loci

Gated MLPs. Across five behaviors the mean relative closure error of Equation 1 is 1.29×10^{-7} to 1.56×10^{-7}. Pair rankings, top-five sets, and held-out patch effects are invariant to quadrature resolution over $K \in \{8, 16, 32, 64\}$: Spearman correlation against $K = 32$ is 1.000 and the largest change in held-out patch effect is 7.0×10^{-6}.

Attention scores. With RoPE, the rotation for a fixed query and key position is independent of content, so

$$s = \frac{(R_q W_Q x_q) \cdot (R_k W_K x_k)}{\sqrt{d_{\text{head}}}} = x_q^\top M x_k, \tag{3}$$

and splitting both sides into named sources gives $s = \sum_{ij} s_{ij}$ with no path integral. Exactness here is structural. Manually recomputed pre-softmax scores reproduce the model's attention probabilities to 0.000×10^{0}, and source closure holds to 1.3×10^{-5}.

A second-order readout. Translating a write into a behavior change uses a local linear readout, and its accuracy is lowest for the largest pairs (calibration slope 0.628 for cycles and 0.475 for months). Adding the curvature term $\frac{\alpha^2}{2} \Delta F^\top H \, \Delta F$ costs one Hessian-vector product per pair, the same correction concurrent work arrives at for single-patch estimates

(Zhang & Wang, 2026), and improves top-five rank fidelity in all five behaviors, with months rising from 0.548 to 0.888 and cycles from 0.675 to 0.924. Magnitude calibration moves in both directions across behaviors, so we claim rank fidelity here rather than calibrated magnitudes.

5 What a first-order method misses

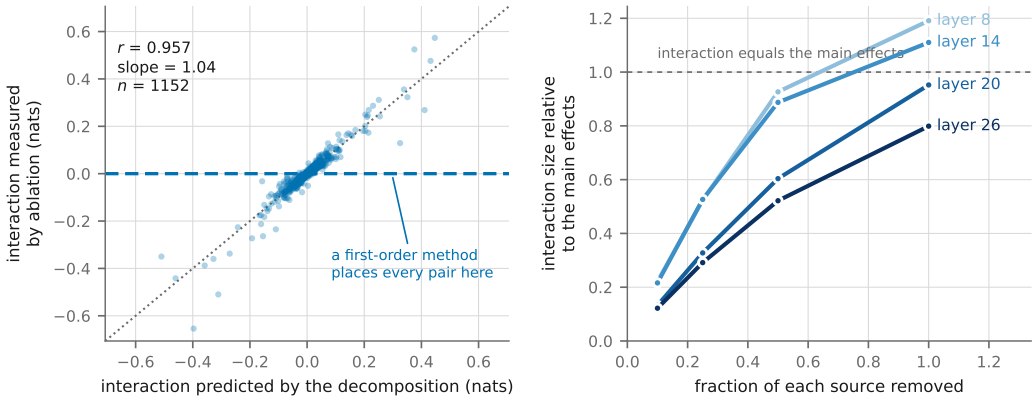

Figure 3: A first-order method assigns zero to every pairwise interaction, and the interactions are as large as the main effects it recovers accurately. Left: the interaction the decomposition predicts against the interaction a 2×2 ablation measures, pooled over four layers, with a tenth of each source removed. The dotted line is equality and the dashed line is where an estimate linear in the patch places every pair. Right: the size of the measured interaction relative to the mean of the two main effects, as the ablation grows from a tenth of each source to all of it.

Define the interaction of two sources as the second difference

$$I_{ij}(\alpha) = F(x) - F(x - \alpha r_i) - F(x - \alpha r_j) + F(x - \alpha r_i - \alpha r_j). \tag{4}$$

An estimate that is linear in the patch assigns this quantity the value zero for every pair and every α, since the two single-source terms already account for the joint one. Measuring I_{ij} says how much such an estimate leaves out, and comparing it against $\Delta F(r_i, r_j)$ says whether the decomposition recovers it. Equation 4 is built from ablations alone and contains no path, so the comparison also tests the pair term against a quantity defined outside the construction.

We measure both at four layers of SmolLM2-135M on 24 WikiText blocks, taking at each position the six pairs with the largest aligned attribution together with six pairs drawn at random from the rest, for 288 pairs per layer. With a tenth of each source removed the decomposition predicts the measured interaction at Pearson correlation between 0.950 and 0.974 and slope between 0.96 and 1.13 across the four layers (Figure 3, left). Agreement survives removing the sources in full, at Pearson 0.785 to 0.919. Rank correlation over the pooled sample is 0.963 against a Pearson correlation of 0.957, so the agreement rests on the whole sample rather than on a few large pairs. Taking only the pairs drawn at random, the correlation is 0.933 pooled against 0.960 for the pairs the method ranked highest, so the agreement holds on pairs the method had no hand in choosing.

The interaction is large enough to matter. Against the mean of the two main effects it runs from 0.12 to 0.22 at a tenth of each source and from 0.80 to 1.19 at full removal (Figure 3, right). At the scale of a complete ablation the part a first-order method assigns to the pair is the same size as the part it assigns to the two sources separately.

First-order estimates stay accurate for the quantity they target. Attribution patching at the same sites recovers the measured main effects at Pearson 0.796 to 0.883 and slope 0.739 to 0.967 (Syed et al., 2024), so the comparison turns on which quantity is estimated rather than on the quality of the estimator.

Carrying the prediction downstream. Read as a change in next-token log-probability, the decomposition predicts the measured interaction at Pearson 0.833, 0.842, and 0.879 at layers 14, 20, and 26, and at 0.387 at layer 8. Substituting the exactly computed interaction for the predicted one moves these to 0.853, 0.885, 0.965, and 0.438, which places the loss in the linearization of the layers between the site and the readout rather than in the decomposition. Independent work reaches the same conclusion about where first-order patching estimates lose fidelity (Zhang & Wang, 2026).

6 Patching a write, and what it establishes

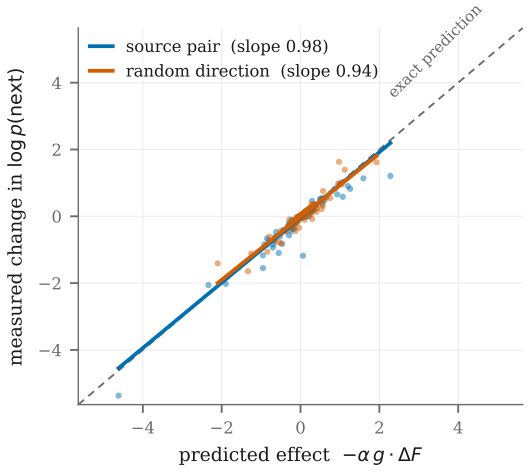

Figure 4: Patching a pair's write changes the next-token log-probability by the amount the decomposition predicts. Each point is one patch at one wikitext position, layer 20, with the readout taken from the token the corpus actually continues with. Source pairs follow the identity line at slope 0.98 ($R^2 = 0.91$). Random directions of matched norm follow it at slope 0.94, which is the behavior a correct first-order attribution produces.

Exactness in composition leaves open whether the decomposition guides intervention at the output. For each pair we patch the MLP output by $-\alpha\,\Delta F(r_i, r_j)$ with $\alpha = 0.2$ and record the change in $\log p$(next token), using the token the corpus continues with so that no target is chosen by hand. Source pairs lie on the identity line at slope 0.977 with 95% interval $[0.945, 1.010]$ and $R^2 = 0.912$ over 336 patches (Figure 4). The quantity the decomposition assigns to a pair is the quantity an intervention at the output recovers.

The same measurement fixes what an intervention of this kind shows. To first order the model's response to any perturbation δ at the MLP output is $g \cdot \delta$, which follows from the readout and holds for every δ. The attribution assigned to a pair is $g \cdot \Delta F(r_i, r_j)$, so a random write with the same projection produces the same response whenever the attribution is right. That is what we observe: random directions of matched norm calibrate at slope 0.944 with interval $[0.912, 0.975]$, overlapping the real pairs within sampling error ($z = 1.43$, $p = 0.15$).

Two things follow. A matched-write control that reproduces a pair's effect is evidence that the attribution is correct. And a write patched at the output carries the information $g \cdot \delta$ about whatever produced it, so establishing that two particular sources were multiplied calls for an intervention on the sources themselves. Section 5 supplies one at corpus scale and Section 7 supplies one conditioned on behavior. The distinction was raised in review of an earlier version of this work, and it bears on how activation patching is read more generally (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024).

A control to avoid. Matching the first-order effect by rescaling a random direction calls for a factor of 10 to 24 in 576 dimensions, which makes the resulting write a far larger

perturbation and overshoots the real effect by roughly nine times. Norm-matched and doubly matched controls behave as intended.

7 Factorial interventions on the sources

To test whether the internal multiplication occurred we remove the sources themselves from the input rather than the write from the output, and read the four-cell interaction

$$I = m(\cdot) - m(-q_i) - m(-k_j) + m(-q_i, -k_j). \tag{5}$$

This quantity is zero when the two sources act additively on the behavior. The argument of Section 6 does not apply to it, because the intervention acts on the input rather than on the output.

At the attention locus the interval excludes zero in four of five behaviors. At the MLP locus the result is clear for capital facts, where the selected pair gives $I_s = 1.599$ against 0.160 for norm-matched real pairs, and the remaining behaviors support a block-level reading. We report this asymmetry directly: the locus where output-write patching is least informative is the locus where the internal test carries the most weight.

8 Ranking the terms of the decomposition

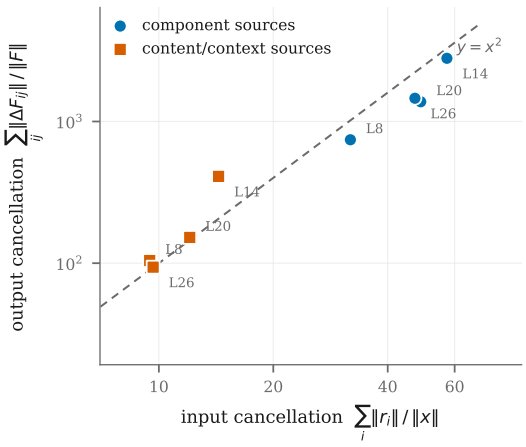

Figure 5: Output cancellation grows as the square of input cancellation. Each point is one layer under one source decomposition. The relationship follows from the form being bilinear and holds across two decades and two different splits of the same input.

The first thing to do with an exact decomposition is to read off its largest terms, and how much that tells you depends on how the size of a term relates to the output. Let input cancellation be $\sum_i \|r_i\|/\|x\|$ and output cancellation be $\sum_{ij} \|\Delta F_{ij}\|/\|F\|$. The form is bilinear, so the second grows as the square of the first, which Figure 5 confirms across two decades and two different splits of the same input.

For the component decomposition at four layers the named sources sum to between 31.6 and 57.4 times the norm of the input they compose, and their pair writes sum to between 724 and 2868 times the norm of the output they compose. Under the split into a position-local part and an attention-routed part, the two parts are close to antiparallel, with cosine between -0.976 and -0.991 and each about five times the size of their sum. That split connects to the result that the effect of context on a transformer block maps to a rank-one patch of its MLP weights (Goldwaser et al., 2025).

Ranking by size reflects this scale structure. The energy matrix $\|\Delta F_{ij}\|$ agrees with its own best rank-one approximation at Spearman 0.991, so ordering pairs by the magnitude of their write recovers which sources are large.

An attribution that accounts for cancellation. Projecting each pair write onto the layer output direction $\hat{F} = F(x)/\|F(x)\|$ gives

$$a_{ij} = \Delta F(r_i, r_j) \cdot \hat{F}, \qquad \sum_{i \le j} a_{ij} = \|F(x)\|, \tag{6}$$

an exact signed attribution of the output with cancellation accounted for by construction. The readout is the layer's own output, so no target is chosen, and the signed terms recover the output norm exactly, which is what makes this an attribution of the output rather than of an arbitrary scalar. Over 512 corpus positions at layer 20 the summed term mass falls from 1402 times the output norm, with 95% interval $[1365, 1442]$, to a median of 80. Its ranking is statistically independent of the magnitude ranking (Spearman -0.000). Pairs that carry a large share of the output appear far down the magnitude ranking: at layer 20, L0_mlp × L0_mlp carries 47.3% of the output at magnitude rank 50, and L15_attn × L20_attn carries 18.3% at rank 234.

9 Behaviors, natural text, and scale

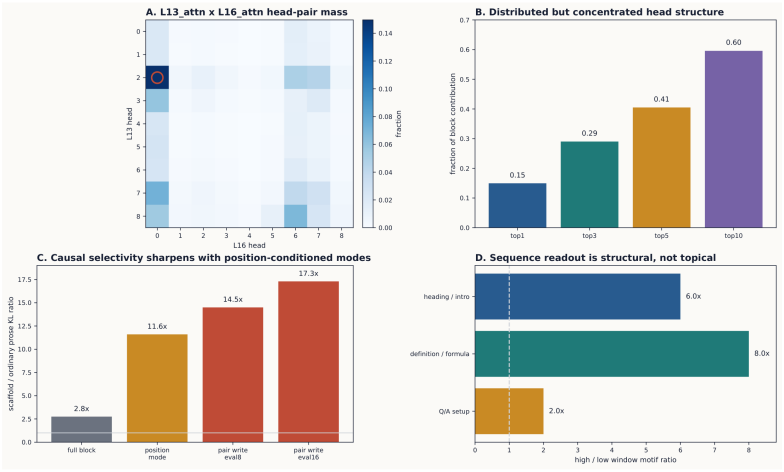

Figure 6: A natural-text source-pair interaction at layer 20, stable across splits and decomposable into native head-pair terms.

Applied to SmolLM2-135M (Ben Allal et al., 2025), the method identifies source-pair circuits for capital facts, cycle continuation, month continuation, opposites, and category facts. Predicted and measured removal effects agree across 105 pairs per behavior (Pearson 0.776 to 0.983, sign accuracy 0.930 to 0.978). Discovery is reproducible: bootstrap selection frequency is 1.00 for capitals and 0.965 to 0.975 for months and opposites, with top-five inclusion 0.97 to 1.00. Dedicated contextual units in middle layers have been reported by sparse probing (Gurnee et al., 2023). On natural text the method finds a stable layer-20 interaction between layer-13 and layer-16 attention writes whose effects concentrate on structured continuation positions.

Scaling to Qwen2.5-0.5B, capital facts are supported at layer 8 by an interaction between the layer-5 and layer-6 MLP writes, chosen over attention-block candidates. That pair is selected in 100% of 200 bootstrap resamples and carries significant patch and internal factorial effects across four prompt templates. The layer-selection rule transfers poorly in both models, and we report that alongside the positive result.

10 Scope and limitations

The decomposition is second order. It reads the interaction between a pair of sources, and a computation that requires three or more sources jointly appears distributed over the pairs

Behavior	Circuit	Actual effect	Control	Readout
Capital facts	$L12H0 \times L13H6$	margin -0.188, CE $+0.122$	margin $+0.004$, CE -0.0003	schema binding
Cycle continuation	$L16H0/H2/H1 \times L18H6$	margin about -0.085 per top pair	near zero	phase and retrieval
Month continuation	$L18H2 \times L20H5$	margin -1.556, CE $+0.401$	margin -0.046, CE $+0.054$	order and suffix
Opposites	$L16H0 \times L20H0$	margin -0.257, CE $+0.034$	margin -0.065, CE $+0.008$	query concept cue
Category facts	$L18_{\mathrm{MLP}} \times L19_{\mathrm{attn}}$	margin -1.393, CE $+0.532$	margin -0.055, CE $+0.035$	block-level only

Figure 7: Behavior-conditioned circuits in SmolLM2-135M layer 20, with margin and cross-entropy effects on held-out examples.

they form. The construction assumes a bias-free gated MLP, which covers the Llama-style feed-forward block used by both models studied here, and leaves out blocks carrying biases and mixture-of-experts routing.

Behavior-conditioned modes are selected by their effect on a chosen target, so what such a mode encodes is a question the selection alone cannot settle. The corpus-scale measurements in Sections 5, 6, and 8 read out the token the corpus continues with and carry no selection of that kind.

Claims about category and type facts are block-level. The layer-18 MLP and layer-19 attention interaction we identify moves the target margin, and the controls we ran place the effect in those blocks rather than in a category-specific mechanism inside either one.

The appendix records the RMSNorm path convention and its sensitivity analysis, the quadrature sweep, the static bilinear replacement that positions the method as its constant-Hessian limit, and the two top-ten statistics whose denominators differ.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big. data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Advances in Neural Information Processing Systems, volume 36, 2023. URL https://arxiv.org/abs/2304.14997.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits, 2024. URL https://arxiv.org/abs/2406.11944.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, and Christopher Olah. A mathematical framework for transformer circuits, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5484–5495, 2021.

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

Adrian Goldwaser, Michael Munn, Javier Gonzalvo, and Benoit Dherin. Equivalence of context and parameter updates in modern transformer blocks, 2025. URL https://arxiv.org/abs/2511.17864.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. Transactions on Machine Learning Research, 2023. URL https://arxiv.org/abs/2305.01610.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. Journal of Machine Learning Research, 22(104): 1–54, 2021.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Tom Henighan, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Michael T. Pearce, Thomas Dooms, and Alice Rigg. Weight-based decomposition: A case for bilinear MLPs, 2024. URL https://arxiv.org/abs/2406.03947.

Michael T. Pearce, Thomas Dooms, Alice Rigg, Jose M. Oramas, and Lee Sharkey. Bilinear MLPs enable weight-based mechanistic interpretability, 2025. URL https://arxiv.org/abs/2410.08417.

Lee Sharkey. A technical note on bilinear layers for interpretability, 2023. URL https://arxiv.org/abs/2305.03452.

Noam Shazeer. GLU variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, pp. 3319–3328, 2017. URL https://arxiv.org/abs/1703.01365.

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, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In Advances in Neural Information Processing Systems, volume 33, 2020. URL https://arxiv.org/abs/2004.12265.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024. URL https://arxiv.org/abs/2309.16042.

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

A Derivation of the source-pair decomposition

Equation 1 in the main text is stated for a vector-valued output and for an arbitrary additive split of the input. This appendix derives it, fixes the symmetry convention that produces the factor of two on off-diagonal terms, and matches each factor to the released implementation.

A.1 Setting

Let the target sublayer be a bias-free gated MLP

$$F(x) = W_{\text{down}}\big(\varphi(W_{\text{gate}}x) \odot W_{\text{up}}x\big), \qquad W_{\text{gate}}, W_{\text{up}} \in \mathbb{R}^{d_{\text{ff}} \times d_m}, \quad W_{\text{down}} \in \mathbb{R}^{d_m \times d_{\text{ff}}}, \tag{7}$$

with $\varphi = \text{SiLU}$. Write g_h and u_h for the h-th rows of W_{gate} and W_{up}, and v_h for the h-th column of W_{down}, so that

$$F(x) = \sum_{h=1}^{d_{\text{ff}}} v_h \, \varphi(g_h^\top x)\, (u_h^\top x). \tag{8}$$

Two properties of this form supply the baseline that the integrated-Hessian identity asks for. Since $\varphi(0) = 0$ and the block carries no biases, $F(0) = 0$. Differentiating one channel gives

$$\nabla_x \big[\varphi(g_h^\top x)(u_h^\top x)\big] = \varphi'(g_h^\top x)(u_h^\top x)\, g_h + \varphi(g_h^\top x)\, u_h, \tag{9}$$

and both terms vanish at $x = 0$, so $J_F(0) = 0$. The two properties rest on the absence of biases and on $\varphi(0) = 0$, so they hold for SiLU, for GELU, and for the bilinear (identity) gate alike.

A.2 The integrated Hessian

Let $\psi(t) = F(tx)$ on $[0, 1]$, so that $\psi(0) = 0$ and $\psi(1) = F(x)$. Taylor's theorem with integral remainder gives $\psi(1) = \psi(0) + \psi'(0) + \int_0^1 (1 - t)\psi''(t)\, dt$, where $\psi'(t) = J_F(tx)\, x$ and $\psi''(t) = H_F(tx)[x, x]$. Here $H_F(y)[a, b] \in \mathbb{R}^{d_m}$ denotes the vector whose c-th entry is $a^\top \nabla^2 F_c(y)\, b$, so the identity holds componentwise across all d_m outputs. The two boundary properties remove the first two terms:

$$F(x) = \int_0^1 (1 - t)\, H_F(tx)[x, x]\, dt. \tag{10}$$

This is the vector-valued form of the integrated-Hessian identity (Janizek et al., 2021; Sundararajan et al., 2017) with the origin as baseline. The origin is where F and its Jacobian both vanish, so it is the baseline the identity requires rather than a reference input chosen by hand.

Substituting (8) into the Hessian gives the explicit form

$$H_F(y)[a, b] = \sum_h v_h \Big[\varphi''(g_h^\top y)\, (u_h^\top y)\, (g_h^\top a)(g_h^\top b) + \varphi'(g_h^\top y)\big((g_h^\top a)(u_h^\top b) + (g_h^\top b)(u_h^\top a)\big)\Big]. \tag{11}$$

Each summand is symmetric in (a, b) and has rank at most two as a quadratic form, with range spanned by g_h and u_h.

A.3 An arbitrary additive split, and the factor of two

Let $x = \sum_{i=1}^n r_i$ be any decomposition of the MLP input. Define the symmetric bilinear map

$$T(a, b) = \int_0^1 (1 - t)\, H_F(tx)[a, b]\, dt, \tag{12}$$

which depends on the input through the integration path and on (a, b) through the two arguments of the Hessian. Substituting the split into (10) and using bilinearity in each

argument separately,

$$F(x) = T\Big(\sum_i r_i,\ \sum_j r_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} T(r_i, r_j), \tag{13}$$

where the double sum runs over ordered index pairs. Because T is symmetric, the ordered pairs (i, j) and (j, i) with $i \neq j$ contribute the same vector. Collecting them,

$$F(x) = \sum_i T(r_i, r_i) + 2\sum_{i<j} T(r_i, r_j). \tag{14}$$

Storing one entry per unordered pair therefore calls for the convention

$$\Delta F(r_i, r_j) = (2 - \delta_{ij})\, T(r_i, r_j), \tag{15}$$

under which $F(x) = \sum_{i \leq j} \Delta F(r_i, r_j)$, which is Equation 1. The factor of two counts the ordered pairs collapsed into one stored entry, and it applies to every off-diagonal term.

In the implementation, integrated_hessian_write and batch_integrated_hessian_writes evaluate T and multiply by two when the off_diagonal flag is set. source_pair_writes sets that flag exactly for $i \neq j$, and pair_score applies the same factor to the scalar operator. The stored writes sum back to $F(x)$ to a mean relative error of 1.4×10^{-7} (Table 2).

A.4 Behavior conditioning, and what is aggregated

For a scalar s read downstream with gradient $g = \partial s / \partial F$ at the target position, the contribution of a pair is $g \cdot \Delta F(r_i, r_j)$. Moving g inside (12) gives the symmetric matrix

$$Q = \int_0^1 (1 - t)\, \nabla^2\big(g^\top F\big)\,(tx)\, dt \in \mathbb{R}^{d_m \times d_m}, \qquad a^\top Q\, b = g \cdot T(a, b), \tag{16}$$

and pair scores are read from Q under the same convention (15). integrated_hessian_matrix builds Q, and integrated_hessian_block projects it into a pair of source bases without forming it.

The behavior scalar used throughout is the next-token margin between the target token and the mean logit over a fixed distractor list, evaluated at the final prompt position, and g is its gradient with respect to the target MLP output at that position, obtained in one backward pass (behavior_margin_gradient). The operator Q is defined per example, and every intervention uses the per-example operator. Where one operator is reported for a behavior, it is the mean of the per-example projected blocks over the discovery examples (mean_integrated_hessian_block), so the aggregation is over examples alone.

B Independence of the path parametrization

Equation (10) integrates along the straight segment from the origin to x, and it is fair to ask how much of the resulting attribution comes from that convention. Within its own class the convention is forced: every monotone reparametrization of the segment gives the same numbers.

Lemma. Let $h : [0, 1] \to [0, 1]$ be continuously differentiable and nondecreasing with $h(0) = 0$ and $h(1) = 1$, and let $\gamma(t) = h(t)x$. For F with $F(0) = 0$ and $J_F(0) = 0$, and for all $a, b \in \mathbb{R}^{d_m}$,

$$\int_0^1 h'(t) \left[\int_0^t h'(s)\, H_F\big(h(s)x\big)[a, b]\, ds\right] dt = \int_0^1 (1 - u)\, H_F(ux)[a, b]\, du. \tag{17}$$

In particular the pair writes $\Delta F(r_i, r_j)$ and the operator Q take the same value for every such h.

Proof. Fix a, b and write $G(s) = H_F(h(s)x)[a, b]$. The integrand on the left is continuous on the triangle $\{0 \le s \le t \le 1\}$, so the order of integration may be exchanged:

$$\int_0^1 h'(t) \int_0^t h'(s) G(s)\, ds\, dt = \int_0^1 h'(s)\, G(s) \left[\int_s^1 h'(t)\, dt\right] ds = \int_0^1 h'(s)\, G(s) \big(1 - h(s)\big)\, ds, \tag{18}$$

using $h(1) = 1$. Substituting $u = h(s)$ with $du = h'(s)\, ds$, and using $h(0) = 0$ and $h(1) = 1$ for the endpoints, gives $\int_0^1 (1 - u) H_F(ux)[a, b]\, du$, in which h has disappeared. □

Where the left side comes from. Along γ the fundamental theorem gives $F(x) = \int_0^1 h'(t) J_F(h(t)x)\, x\, dt$, and $J_F(0) = 0$ gives $J_F(h(t)x)\, x = \int_0^t h'(s) H_F(h(s)x)[x, x]\, ds$. Substituting $x = \sum_i r_i$ and expanding by bilinearity produces the left side of (17) with $(a, b) = (r_i, r_j)$, summed over ordered pairs. The lemma therefore says that the decomposition and each of its terms are unchanged by the reparametrization.

Two choices do fix the numbers: the baseline, which the boundary conditions of Appendix A supply, and the RMSNorm convention of Appendix D. The speed at which the segment is traversed leaves them alone, so the straight-line convention is the canonical representative of its class.

Numerical confirmation. We reimplemented the quadrature under $\gamma(t) = h(t)x$ with $h(t) = t^2$, which moves the evaluation point from t to t^2 and the integration weight from $(1 - t)$ to $2t(1 - t^2)$, and compared the resulting pair writes against the straight-line implementation. At SmolLM2-135M layer 20 on four WikiText positions under the full 42-source component split, covering all 903 unordered pairs at each position, the largest relative disagreement is 3.6×10^{-15} at matched $K = 32$ and 1.8×10^{-15} with the alternative parametrization at $K = 64$. Taking $h(t) = t^3$ at $K = 128$ gives 1.2×10^{-14}. Both conventions reproduce the layer's true output to the same relative error at every position, between 1.2×10^{-6} and 3.7×10^{-6}, which is the single-precision floor of the forward pass they are compared against. The disagreement between conventions therefore sits at the level of double-precision accumulation.

C The channel form

Equation (11) is a sum over d_{ff} hidden channels, and each summand touches the input through two scalars. Making that explicit reduces the cost of the whole operator at one token to a single pass over the hidden layer, which is what makes the corpus-scale measurements of Section 8 affordable, and it isolates the quantities through which context reaches the operator.

Forward form. Since $\text{SiLU}(z) = z\, \sigma(z)$ with σ the logistic function, $\varphi(\gamma) \odot \upsilon = \gamma \odot \upsilon \odot \sigma(\gamma)$ for the two pre-activation vectors $\gamma = W_{\text{gate}} x$ and $\upsilon = W_{\text{up}} x$, written g_x and u_x in Equation 2.

Operator form. Setting $y = tx$ in (11) gives $g_h^\top y = t\gamma_h$ and $u_h^\top y = t\upsilon_h$, so the whole t dependence of channel h runs through those two scalars. Carrying out the integral in (12) channel by channel,

$$T(a, b) = \sum_h \upsilon_h\, a^\top M_h\, b, \qquad M_h = A_h\, g_h g_h^\top + B_h \big(g_h u_h^\top + u_h g_h^\top\big), \tag{19}$$

$$A_h = \upsilon_h \int_0^1 t(1 - t)\, \varphi''(t\gamma_h)\, dt, \qquad B_h = \int_0^1 (1 - t)\, \varphi'(t\gamma_h)\, dt, \tag{20}$$

with $\varphi'(z) = \sigma(z) + z\sigma(z)(1 - \sigma(z))$ and $\varphi''(z) = \sigma(z)(1 - \sigma(z))\big(2 + z(1 - 2\sigma(z))\big)$. The behavior-conditioned operator of (16) is $Q = \sum_h (W_{\text{down}}^\top g)_h\, M_h$.

Reading. Everything in (19) other than A_h and B_h is a weight. B_h depends on the context only through the gate pre-activation γ_h, and A_h is the same kind of quantity scaled by the up pre-activation v_h, so two scalars per channel carry all of it. One pass over the hidden layer produces (A, B), after which every pair at that token costs one matrix product, and (A, B) is a behavior-free summary of the multiplicative structure at that token. Q is therefore a weighted sum of d_{ff} forms of rank at most two, one per hidden channel. integrated_hessian_matrix_from_channels accepts an arbitrary per-channel weighting in place of $W_{\text{down}}^{\top} g$, which supplies the matched null that any claim about the spectrum of Q needs.

Agreement with the implementation. channel_coefficients returns (A, B). At SmolLM2-135M layer 20, where $d_{\text{ff}} = 1536$, on two WikiText positions, rebuilding pair writes from A and B alone and comparing against integrated_hessian_write gives a worst relative difference of 3.9×10^{-16}, and the forward channel form matches the gated forward pass to 2.4×10^{-16}. At these tokens A ranges over $[-0.24, 0.23]$ and B over $[0.03, 0.45]$.

D RMSNorm source attribution

The decomposition needs an additive split of the MLP input, and the MLP input is the RMSNorm output rather than the residual stream itself. The residual stream is additive by construction and the normalizer rescales it jointly, so the split across the normalizer is a convention that has to be stated. The paper uses two, and both close exactly.

Let $r = \sum_i r_i$ be the residual stream at the target position, w the RMSNorm gain, ϵ the stabilizer, $M = \text{mean}(r^2)$, and $N(r) = w \odot r\,(M + \epsilon)^{-1/2}$.

Frozen denominator (endpoint_scale). Evaluate $\rho = (M + \epsilon)^{-1/2}$ once at the full residual and set $n_i = \rho\,(w \odot r_i)$. Then $\sum_i n_i = N(r)$ exactly, each source keeps its direction, and the attribution is linear in the source under a single shared scale. The influence each source has on the denominator stays inside that shared factor. We report this convention as the sensitivity analysis.

Path-integrated (joint_zero). Set $n_i = \int_0^1 J_N(tr)\, r_i\, dt$. Since $N(0) = 0$, summing over i gives $\int_0^1 J_N(tr)\, r\, dt = N(r)$, so this convention also closes exactly. Every measurement reported in the paper uses this convention, and the agreement table below gives the frozen-denominator values alongside. With

$$J_N(y)\,\nu = w \odot \Big[\nu \left(\text{mean}(y^2) + \epsilon\right)^{-1/2} - y\ \text{mean}(y \odot \nu)\left(\text{mean}(y^2) + \epsilon\right)^{-3/2}\Big], \tag{21}$$

both integrals along $y = tr$ are elementary:

$$n_i = w \odot \left[c_1\, r_i - c_2\, \text{mean}(r \odot r_i)\, r\right], \tag{22}$$

$$c_1 = \frac{\operatorname{asinh}\sqrt{M/\epsilon}}{\sqrt{M}}, \qquad c_2 = \frac{\operatorname{asinh}\sqrt{M/\epsilon} - \sqrt{M}\big/\sqrt{M+\epsilon}}{M^{3/2}}, \tag{23}$$

with limits $c_1 \to \epsilon^{-1/2}$ and $c_2 \to \frac{1}{3}\epsilon^{-3/2}$ as $M \to 0$, which the implementation uses near zero. The lemma of Appendix B applies here in its first-order form, so this convention is likewise insensitive to reparametrization of the segment.

A third convention (sequential) adds sources in a fixed order and assigns each the telescoping difference $N(\sum_{k \le i} r_k) - N(\sum_{k<i} r_k)$. It closes by construction and depends on the order, and it is available as an order-sensitivity control.

Sensitivity. Both conventions were evaluated from the same forward pass for all five behaviors at SmolLM2-135M layer 20, with $K = 32$ and a source-PCA basis capped at rank 64 retaining 0.999 of the energy, over the full set of 903 source pairs.

The two conventions agree on the broad ordering, at Spearman 0.844 to 0.929, and on the leading mode of the operator, at singular-vector cosines 0.884 to 0.981. They select the same

Table 1: Agreement between the two RMSNorm conventions at layer 20. Spearman is taken over all 903 pair scores. The last column reports the absolute cosine between the left and between the right leading singular vectors of the two operators.

behavior	pair-ranking Spearman	top-five overlap	same top pair	leading-mode cosine
capitals	0.844	3/5	no	0.967 / 0.921
cycles	0.877	4/5	no	0.962 / 0.884
months	0.929	2/5	yes	0.979 / 0.915
opposites	0.899	4/5	no	0.930 / 0.981
category	0.907	4/5	no	0.976 / 0.975

top-ranked pair in one of five behaviors. The convention is therefore part of the specification whenever a single pair is named, and the main text states it wherever that happens.

E Attention scores need no path integral

The same source-pair reading applies to attention logits, where it holds exactly by structure. The reason marks the boundary of the path-integral machinery.

For a query position q and a key position k in one head, the pre-softmax logit is

$$s = \lambda\,(R_q W_Q x_q)^\top (R_k W_K x_k) = x_q^\top M_{qk}\, x_k, \qquad M_{qk} = \lambda\, W_Q^\top R_q^\top R_k W_K, \tag{24}$$

with λ the attention scale and R_p the RoPE rotation at position p. Each R_p is a fixed block-diagonal rotation determined by the position alone, so M_{qk} carries no dependence on content, and $R_q^\top R_k = R_{k-q}$ makes it a function of the offset. The map $(x_q, x_k) \mapsto s$ is therefore bilinear globally, and substituting the two source splits $x_q = \sum_i q_i$ and $x_k = \sum_j k_j$ gives

$$s = \sum_i \sum_j q_i^\top M_{qk}\, k_j. \tag{25}$$

Every term is available in closed form, with floating point as the only source of error.

Two differences from the MLP case follow from the form. First, M_{qk} is asymmetric and its two arguments sit at different positions, so the pairs are ordered: a split into n_q query sources and n_k key sources gives $n_q n_k$ terms, and the factor of two in (15) has no analogue. Second, exactness here is structural. Measured on the five behaviors, source closure holds to 1.3×10^{-5}, and pre-softmax scores recomputed from the decomposition reproduce the model's attention probabilities to 0.

The bilinear object is the logit. Softmax and the value path sit outside it, so a statement read from this decomposition reaches as far as the logit and stops there. The interventions respect the same boundary: qk_source_removal hooks the query and key projections and leaves the value path intact, so the four-cell factorial of Section 7 acts on the bilinear form alone. Grouped-query attention changes bookkeeping only, with the key head index obtained from the query head index by integer division by the group size. The derivation assumes bias-free query and key projections, which holds for SmolLM2-135M, where the attention results are measured.

F Quadrature convergence

The resolution K is the one numerical parameter in the MLP decomposition. The integrand of (12) is $(1-t)H_F(tx)[a,b]$, smooth in t on $[0,1]$, and it is evaluated with a K-node Gauss-Legendre rule mapped from $[-1,1]$ to $[0,1]$. All reported results use $K = 32$.

Closure error settles by $K = 16$ and is unchanged from $K = 32$ to $K = 64$, which places the residual at the single-precision floor of the forward pass the writes are compared against. The $K = 8$ mean is carried by one behavior, month continuation, whose mean closure error there is 1.3×10^{-5}; the other four stay within a factor of two of their $K = 32$ values at

Table 2: Quadrature sweep at SmolLM2-135M layer 20 over five behaviors with six held-out examples each (30 examples), at $\alpha = 0.2$ and with $K = 32$ as the reference. Closure error is $\| \sum_{i \le j} \Delta F_{ij} - F(x) \| / \| F(x) \|$. The last column is the largest change, across behaviors, in the mean held-out patch effect of the configured pair. Spearman entries are rounded to three decimals; the smallest value anywhere in the sweep is 0.99999998.

K	closure error		ranking	top-five	change in
	mean	worst	Spearman	overlap	patch effect
8	2.7×10^{-6}	2.3×10^{-5}	1.000	5/5	7.0×10^{-6}
16	1.48×10^{-7}	1.71×10^{-7}	1.000	5/5	3.5×10^{-6}
32	1.36×10^{-7}	1.51×10^{-7}	1.000	5/5	0
64	1.36×10^{-7}	1.60×10^{-7}	1.000	5/5	2.5×10^{-6}

$K = 8$ as well. Everything downstream is stable: the pair ranking, the top-five set, and the held-out patch effect agree across all four resolutions to six significant figures. We keep $K = 32$ as the default so that the released code matches every recorded artifact.

G The static bilinear replacement as the constant-Hessian limit

Replacing the gate with the identity turns (7) into a bilinear sublayer, and the decomposition of Appendix A collapses onto it. With $\varphi = \mathrm{id}$ the second-derivative term in (11) drops and $\varphi' \equiv 1$, so

$$H_F(y)[a, b] = \sum_h v_h \big((g_h^\top a)(u_h^\top b) + (g_h^\top b)(u_h^\top a) \big) \tag{26}$$

holds at every y. The Hessian is constant, the integral in (12) reduces to the factor $\int_0^1 (1 - t)\, dt = \frac{1}{2}$, and $T(a, b) = \frac{1}{2} H_F[a, b]$ with no quadrature and no dependence on the input. This is what the identity activation option evaluates, and it recovers the static quadratic form that bilinear MLPs expose (Sharkey, 2023; Pearce et al., 2024; 2025). The behavior-conditioned operator of pretrained gated MLPs is the same object evaluated locally.

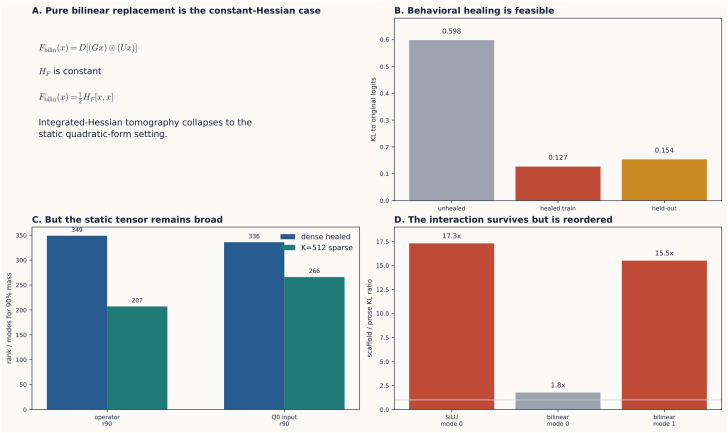

Figure 8: Pure bilinear replacement gives the constant-Hessian limit of tomography. Healed static bilinear spectra stay broad, which favors local behavior-conditioned operators for pretrained gated MLPs.

To see how close that limit sits to the trained block, we replace the layer-20 MLP of SmolLM2-135M with a bilinear block, heal it on WikiText, and measure the healed model against the original on held-out validation text (64 sequences of 128 tokens, 8,128 scored positions). Perplexity moves from 34.89 to 40.64, a ratio of 1.16; next-token top-token agreement is 0.804; mean token-level KL from the original is 0.156; and the replaced layer's

output has mean cosine 0.765 with the original layer's output. On the five behavior families (60 prompts) candidate accuracy moves from 0.767 to 0.717 and the mean margin from 4.95 to 4.50, a retention of 0.910.

The spectral result is the point of the comparison: the healed static spectra stay broad (Figure 8), which is what favors a local behavior-conditioned operator over a single global quadratic form for a pretrained gated MLP. We use the replacement to position the method at its constant-Hessian limit, and the numbers above set how far the limit sits from the trained block.

H Two top-ten head-pair statistics

Figure 6 resolves the layer-13 and layer-16 attention sources into their nine heads each, giving an 81-entry head-pair matrix. Two different fractions of that matrix are quoted in the results, and their denominators differ, so we define both.

The direct head-pair score D_{ab} is the behavior-conditioned pair score between head a of the left source and head b of the right source, evaluated from the integrated-Hessian block and averaged over samples, with the off-diagonal factor of (15) applied because the two source layers differ. Its top-ten fraction is $\sum_{\text{top }10} |D_{ab}| / \sum_{ab} |D_{ab}|$, so the denominator is the mass of the whole matrix. It is 0.417 on discovery and 0.383 held out.

The leading-mode mass is $M_{ab} = \sigma\, \ell_a \rho_b$, built from the leading singular triple (σ, ℓ, ρ) of the source block with the head writes projected onto ℓ and ρ. Its top-ten fraction has the same form with M in place of D, so the denominator is the mass of that one rank-one mode. It is 0.564 on discovery and 0.572 held out. A rank-one matrix concentrates its mass by construction, which is why the second number is the larger one. We name which of the two is meant wherever either appears.

I Behavior atlas and controls

Each of the five prompt families induces its own source-pair operator at the same layer, and Figure 9 shows all five side by side. Figure 10 shows the two conditions that bound what those operators support: an operator built with no behavior target, and the topic readout used to test how far a semantic reading carries.

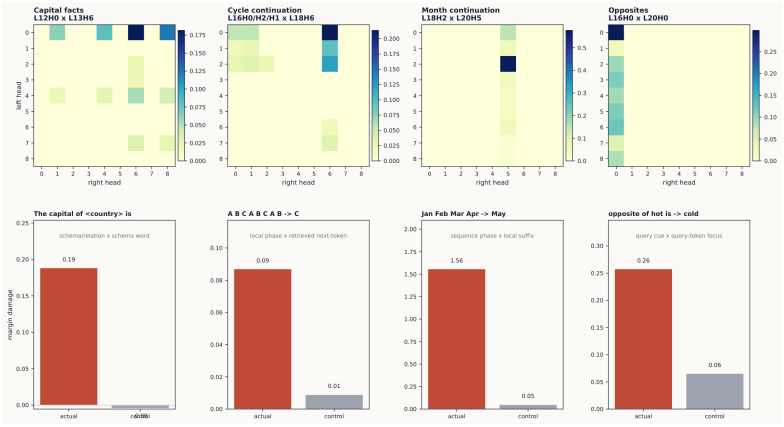

Figure 9: Behavior-conditioned circuit atlas. Each prompt family induces a different source-pair operator at the same layer.

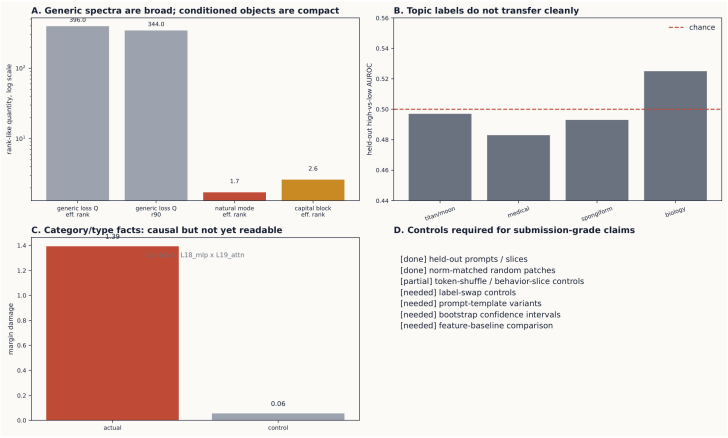

Figure 10: Controls and boundaries. Generic operators are broad and topic labels constrain the readable claim.

Table 3: Models, pinned by revision. Both have bias-free gate, up, and down projections, which the derivation of Appendix A requires; the weight-extraction helper raises if a bias is present.

	SmolLM2-135M	Qwen2.5-0.5B
layers	30	24
d_m / d_{ff}	576 / 1536	896 / 4864
attention heads	9 over 3 key/value	14 over 2 key/value
gate	SiLU	SiLU
RMSNorm ϵ	10^{-5}	10^{-6}

identifier	revision
HuggingFaceTB/SmolLM2-135M	93efa2f097d58c2a74874c7e644dbc9b0cee75a2
Qwen/Qwen2.5-0.5B	060db6499f32faf8b98477b0a26969ef7d8b9987

J Experimental settings

J.1 Models and runtime

All experiments run in float32 on CPU. Models are loaded with attn_implementation set to eager. The second-order readout reported in the main text requires it, because the fused CPU attention kernel has no double backward. Elsewhere the setting is numerically inert: first-order predicted-versus-actual Pearson under the eager and the fused kernels differ by less than 10^{-6} on all five behaviors, with a largest difference of 8.8×10^{-7}.

Software: Python 3.12.13, torch 2.13.0, transformers 5.14.1, datasets 5.0.1, safetensors 0.8.0, macOS on Apple silicon with 12 torch threads. Every result artifact records this block alongside its numbers.

J.2 Data

Behavior probes are five prompt families (capital facts, cycle continuation, month continuation, opposites, category facts) of twelve prompts each, split six for discovery and six held out. The split is written into the behavior configuration file and frozen before any intervention runs. Each family carries a fixed distractor list, and the behavior scalar is the target logit minus the mean distractor logit at the final prompt position.

Natural text is Salesforce/wikitext with configuration wikitext-103-raw-v1, at revision b08601e04326c79dfdd32d625aee71d232d685c3. It is streamed and packed into non-overlapping fixed-length blocks, with an end-of-text token between documents.

Table 4: Corpus settings per experiment. Blocks are packed deterministically from the stream; a seed appears where the stream is shuffled or a random control is drawn.

experiment	split	block length	blocks	seed
calibration (Figure 4)	train	40	24	0
cancellation (Figure 5)	train	48	10	none
component cancellation ratios	train	48	12	none
natural-text tomography (Figure 6)	train / validation	128	256 each	0
bilinear healing diagnostics	validation	128	64	31415

Natural-text tomography reads positions from index 8 with stride 8, uses an uncentered rank-2 source PCA basis, and streams with a shuffle buffer of 10,000. The cancellation measurements cover layers 8, 14, 20 and 26 under both the component split and the position-local versus attention-routed split.

J.3 Tomography and interventions

Quadrature resolution is $K = 32$ throughout. The intervention scale is $\alpha = 0.2$, applied as a patch of $-\alpha\,\Delta F$ at the MLP output. Source-PCA bases cap at rank 64 and retain 0.999 of the energy. RMSNorm closure is checked at every trace, with tolerance 3×10^{-5} for component sources, 5×10^{-5} when attention heads are resolved, and 10^{-4} for Qwen; a trace that fails the check raises rather than being recorded.

Predicted-versus-actual evaluation uses the five top-ranked pairs plus one hundred further pairs per behavior, 105 in total, patched in batches of twenty-five. Interval estimates are 2,000-resample percentile bootstraps at 95%, seeded per behavior from 1729. Discovery stability uses 200 resamples of the discovery examples, re-running selection inside each resample.

The calibration of Section 6 patches at layer 20 on the WikiText blocks of Table 4, reading at the second-to-last position the token the corpus continues with, and takes fourteen pairs per position spread evenly across the attribution ranking, for 336 patches per write type. The random-direction and shuffled-source arms draw from a generator seeded at 0.

The layer sweep covers all thirty SmolLM2 layers under the same behaviors and settings. The Qwen2.5-0.5B replication targets layer 8, selected by a separate discovery-only sweep over all layers, and evaluates four prompt templates (canonical, possessive, capital city, in country) with six held-out examples each, 200 discovery bootstraps, and thirty-two gradient-matched controls per example.

J.4 Compute

Every experiment reported in the main text runs on CPU. The full set of 110 result artifacts totals 2,566 seconds of wall time, and the most expensive single run, the thirty-layer sweep across five behaviors, takes 134.5 seconds. A GPU is used only to train the static bilinear replacement and its healing checkpoints, which position the method rather than producing any tomography result.

K Reproducibility statement

All models and datasets are pinned by revision: SmolLM2-135M and Qwen2.5-0.5B at the commits in Table 3, and WikiText-103 (raw) at b08601e04326c79dfdd32d625aee71d232d685c3. Every result artifact records the software versions and the accelerator that produced it, together with the quadrature resolution, the intervention scale, the RMSNorm convention, and the seeds. Discovery and validation

splits are disjoint and frozen before any intervention is run: source pairs and target layers are selected on discovery data alone, and every reported causal effect is measured on held-out examples. Interval estimates are seeded example-cluster bootstraps. The release includes the tomography implementation, the intervention and control harnesses, the behavior specifications, and a configuration file for every experiment reported here, together with the machine-readable artifact that each figure and table is rendered from, and an audit script that regenerates the comparison between the reported numbers and those artifacts.