THE COMMUNICATION MAP OF A TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

The components of a transformer communicate by writing to and reading from a shared residual stream, and the mechanistic interpretability literature has mapped these connections by hand, one circuit at a time. We present the *communication map*, which charts every potential communication channel from the geometry of the model's weights alone, generalizing the composition score of Elhage et al. (2021) into a single *coupling coefficient* covering all 18 connection classes, from head-to-head to neuron-to-neuron and everything in between. We provide an account of the properties of the coupling coefficient, including its geometric interpretation and its exact chance level. The census finds that 70–89% of head pairs are oriented far from chance, some coupled strongly and others actively avoiding each other. We demonstrate the communication map in two novel applications. In Application 1, we recover the known induction circuits blind from the strongest head-to-head couplings and group the heads into communities, and ablating one such community destroys the model's in-context copying. In Application 2, we pool the coupling coefficients of every head to identify a distinct two-dimensional residual stream subspace, whose deletion abolishes the induction capability in six models up to Pythia-6.9B. We show that this subspace is different from those identified by either activation PCA or outlier dimensions. We release the map, the statistical machinery, and the intervention suite.[1]

1 INTRODUCTION

A transformer (Vaswani et al., 2017) connects all of its components through a single shared medium of communication, the *residual stream* (Elhage et al., 2021). Every attention head and every neuron reads from the residual stream and writes its result back into it for downstream consumption. Superposition allows this d-dimensional stream to hold exponentially many nearly orthogonal directions, far more than d, so the residual stream acts as a public information "highway" carrying many millions of "channels" of communication between components (Elhage et al., 2022).

However, little is known about how the hundreds of attention heads, the orders of magnitude more neurons, and the other components of a transformer communicate over this information highway. *Who is talking to whom? Do heads in one layer mainly talk to the heads in the next layer up or do they communicate long-range across multiple layers? Do they form subnetworks? Do they share common channels in a "broadcast" style, or do they prefer one-to-one communication over more "private" channels?* These questions go to the heart of understanding the inner workings of transformers but, to our knowledge, largely remain unanswered in the mechanistic interpretability literature (Ferrando et al., 2024; Rai et al., 2024). This paper answers these questions by building the first communication map of an entire transformer from the geometry of the model's weight matrices alone.

Mechanistic interpretability has examined individual circuits and connections of a transformer. Elhage et al. (2021) defined *composition scores* measuring whether one attention head reads what another wrote, and used them to explain *induction heads* (Olsson et al., 2022). Wang et al. (2023) hand-traced a circuit of some two dozen heads and identified indirect-object-identification (IOI) circuits. Merullo et al. (2024) showed that specific pairs of distant heads communicate over low-rank channels. In every case, the analysis starts from a known behavior and works backward to a handful of components. To our knowledge, no full transformer-wide communication maps across all

[1] Code and results are included in the supplementary material.

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

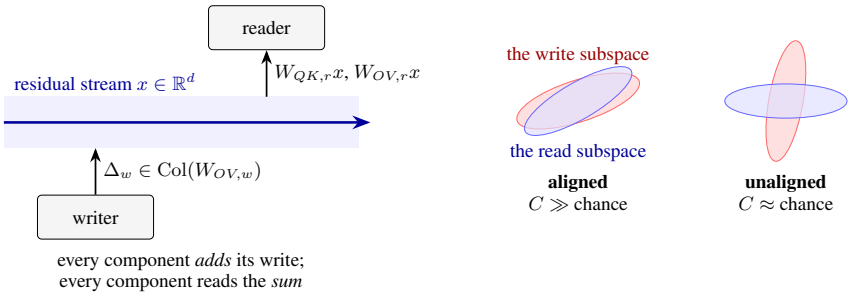

Figure 1: **The communication map's unit of analysis.** *Left:* components interact only by adding writes to, and linearly reading from, a shared residual stream (Eq. 1). *Right:* an edge is a writer–reader pair whose write and read subspaces align far beyond chance.

components have ever been built. The closest attempt is by Yamagiwa et al. (2026), which computes a subspace-affinity metric between attention heads, but it reports only the top twenty pairs and excludes MLPs entirely.

The residual stream is akin to a wide radio band that can accommodate millions of nearly interference-free channels of communication simultaneously (Figure 1). Many signals ride on the same d dimensions, yet each component transmits and receives in its own particular subspaces, like their own "frequencies." For instance, when the directions that a head in Layer 1 writes to align closely with the directions that a head in Layer 5 reads from, the two heads can pass information directly. The weights of a trained model's matrices therefore contain, implicitly, a complete directory of who can talk to whom, over which subspaces, and how strongly. *Using that directory to reconstruct the communication map of an entire transformer and to demonstrate the map's usefulness in two new applications is the central goal of this paper.*

To demonstrate this mapping technique, we build full-transformer communication maps for GPT-2 small, GPT-2 medium, and GPT-2 large (Radford et al., 2019), and for the Pythia family (160M, 2.8B, 6.9B) (Biderman et al., 2023). In the case of GPT-2 small, we map every connection among its 144 heads, its 36,864 MLP neurons, and the embedding, positional, and unembedding matrices. Every directed writer–reader pair is scored from weights with the *coupling coefficient* $C = \|RW\|_F / (\|R\|_F \|W\|_F)$, generalized from Elhage et al. (2021) for reader–writer pairs of any shape, where W is the writer's write matrix and R the reader's reading matrix. The result is a graph of *potential connectivity* for the whole model. Charting the communication across an entire transformer is challenging not only because of the billions of candidate writer–reader pairs, but also because of the need for statistical significance testing. Since any two random directions in $\mathbb{R}^{768}$ overlap somewhat, a statistical procedure is needed to separate genuinely aligned pairs from chance overlap. After charting the map, we conduct two ablation experiments to demonstrate the applications of the map, yielding novel findings.

We contribute to the literature in the following areas:

Full map. First, we construct the first full communication map of a transformer (Section 3), including all 18 connection classes and billions of candidate connections or "edges" within minutes, and provide the statistical machinery for separating real connections from chance alignment. Among the results, we discover that attention heads communicate with each other over long ranges spanning many layers and heads form communities (Application 1). One community, in particular, holds all five induction heads and most of the IOI circuit, and ablating it destroys 93.8% of the model's in-context copying.

Theory. Second, we provide an account of the properties of the coupling coefficient, C, which generalizes the composition score of Elhage et al. (2021) to all 18 classes of read-write connections. We discuss its geometric interpretation, its random chance level, and rotational invariance.

Induction-critical subspace. Third, in our Application 2 (Section 5), we use the pooled coupling coefficient C of all heads of a transformer to identify a 2-dimensional stream subspace critical to induction of the whole model. Deleting this subspace destroys 82–96% of the induction capability on six models from 124M to 6.9B parameters, well above those achieved by the prior literature's methods (Mu & Viswanath, 2018; Timkey & van Schijndel, 2021).

Release. Fourth, we release the map, the statistical machinery, and the intervention suite, with fixed seeds.

2 BACKGROUND AND RELATED WORK

The residual stream is the communication backbone of a transformer. Attention heads and MLP neurons in different layers address one another by writing to and reading from the stream, each through fixed linear maps into subspaces of its own, and all writes to the residual stream combine additively (Elhage et al., 2021). At every token position i, layer ℓ receives a stream vector $x_i \in \mathbb{R}^d$ (d=768 for GPT-2 small) that is exactly

$$x_i \;=\; \mathrm{emb}(t_i) \;+\; \mathrm{pos}(i) \;+\; \sum_{c \,\in\, \text{components before } \ell} \Delta_c(i), \tag{1}$$

where $\Delta_c(i)$ is the write of component c. Every reader's input therefore decomposes exactly into contributions from identifiable writers, all of which share the same d dimensions (Elhage et al., 2022). As Figure 1 shows, each head or MLP neuron is tuned to its own subspaces of the stream, and one upstream writer component can influence a downstream reader component precisely when the writer's subspaces align with those of the reader. Furthermore, the stream's *interface matrices*, where the model meets its initial inputs and final outputs, are the token embedding W_E and the positional embedding W_{pos}, which write into the stream before layer 0, and the unembedding W_U, which reads the stream after the last layer and converts the result to next-token logits.

Elhage et al. (2021) establishes that an upstream writer head can influence a downstream reader head through three channels, known as the *K-composition*, *Q-composition*, and *V-composition*. The *K-composition* is $W_{QK,r}W_{OV,w}$ (Olsson et al., 2022) (with $W_{QK} = W_Q W_K^\top$ and $W_{OV} = W_O W_V^\top$, the head's attention-pattern form and write operator), the *Q-composition* is $W_{QK,r}^\top W_{OV,w}$, and the *V-composition* is $W_{OV,r}W_{OV,w}$, where the subscript w marks the writer head and r the reader.[2] The definitions of these matrices, along with the rest of this paper's conventions, shapes, and the attention block derivation, are all in Appendices A.1–A.2.

These three compositions form the conceptual foundation of single-circuit analyses (Olah et al., 2020), whose findings span induction circuits (Olsson et al., 2022), indirect-object identification (IOI) via manual tracing (Wang et al., 2023), and individual low-rank head-pair channels (Merullo et al., 2024). Each MLP neuron is also an individual rank-one reader and writer, reading in one direction through $a_m = r_m^\top x$ and writing in another direction via $\Delta_m = g(a_m)\, w_m$ (Geva et al., 2021). Neuron-level structure has its own literature, which probes what individual neurons encode and finds families of functionally universal neurons recurring across models (Gurnee et al., 2023; 2024). That literature asks what neurons represent, whereas the map asks whom they talk to.

A writer–reader pair whose subspaces align far beyond chance is a connection or "edge" in the communication graph, and we quantify connection strength with the coupling coefficient C in Section 3.

A second paradigm of research focuses on the activations a model produces rather than the geometry of its weight matrices. These works include automated circuit discovery and attribution (Conmy et al., 2023; Syed et al., 2024; Ameisen et al., 2025; Marks et al., 2025), causal interventions (Vig et al., 2020; Meng et al., 2022; Chan et al., 2022; Zhang & Nanda, 2024), head ablations (Michel et al., 2019; Voita et al., 2019), and scaling studies (Lieberum et al., 2023). They characterize traffic on particular inputs, one behavior at a time, increasingly over learned feature dictionaries rather than neurons (Huben et al., 2024; Dunefsky et al., 2024).

More recently, Tang et al. (2026) patch label-aligned subspaces of the residual stream, showing that a roughly 70-dimensional subspace carries most of the causal effect of in-context task information in Llama-3-8B. Xu (2026) clusters attention heads into communities by co-activation and validates them by ablation against matched random controls. In comparison, our work on head communities in Section 4 follows the same validation logic but is built entirely from the geometry of model weights instead of activations, and maps neurons and the interface matrices in addition to attention heads.

[2] A fourth term, in which the writer enters the keys and the queries at once, is bounded by the two one-sided channels and thus adds no independent channel of its own (see Appendix A.2 for proof).

The closest weight-geometry work to ours is by Yamagiwa et al. (2026), who rank head-pair subspace affinities in GPT-2 small and visualize the top twenty, but they do not create the full communication map of a transformer, as they exclude MLPs, discard the operators' singular values (i.e., the channel "gains"), and compare pooled score distributions to a random-subspace reference rather than testing individual pairs.

Assembling these primitives into a whole-model map is also a statistics problem. Screening all of a model's candidate channels, for example 6.3×10^8 in GPT-2 small and 1.3×10^{11} in Pythia-6.9B, is large-scale simultaneous inference, and statisticians developed the required toolbox for exactly this setting in genomic screening. We apply empirical null distributions fitted to the data's own geometry (Efron, 2004) and the false-discovery-rate (FDR) control procedure of Benjamini & Hochberg (1995).

3 THE COMMUNICATION MAP: THE COUPLING COEFFICIENT AND THE CENSUS

To create the full-transformer communication map, we compute a *coupling coefficient* for every ordered, causally eligible writer–reader pair within a model. The census of these coefficients forms the basis of the map. Derivations are in Appendix A. Exact computational methods are detailed in Appendix B.

The coupling coefficient (C). Generalizing the composition score of Elhage et al. (2021) from head pairs to arbitrary writer–reader matrices in a transformer, we define the *coupling coefficient* of a writer with write matrix W and a reader with read matrix R as

$$C \;=\; \frac{\|R\,W\|_F}{\|R\|_F\,\|W\|_F} \qquad \in [0,1]. \tag{2}$$

The reader and writer can be any component that writes to or reads from the residual stream. For example, W equals $W_{OV,w}$ for a head, and w_m for a neuron; R equals $W_{QK,r}$ for K-composition, $W_{QK,r}^\top$ for Q-composition, $W_{OV,r}$ for V-composition, and $r_m^\top$ for a neuron. The token embedding matrix W_E and, when it exists, the positional embedding matrix W_{pos}, are also writers, whereas the unembedding matrix W_U is the final reader.[3] For head–head pairs, C is exactly the composition score of Elhage et al. (2021). The generalization lets one formula cover every connection class of the map, from the $d \times d$ head circuits W_{QK} and W_{OV} down to rank-one neurons, and every feasible type of connection in between.

Geometric interpretation. Let p_k be the fraction of the reader's total squared weight $\|R\|_F^2$ carried by its k-th principal input direction (its k-th squared singular value over their sum), q_ℓ the same for the writer's output directions, and $\theta_{k\ell}$ the angle between the two directions. Then Proposition 1 (Appendix A.3) says:

$$C^2 \;=\; \sum_{k,\ell} p_k\, q_\ell\, \cos^2\theta_{k\ell}. \tag{3}$$

To use a radio analogy, $\cos^2\theta_{k\ell}$ asks whether the two stations are tuned to the same frequency, q_ℓ measures how strongly the sender transmits on that frequency, and p_k measures how much the receiver amplifies the signal. Since the direction vectors on each side are orthonormal (Appendix A), the double sum in Eq. 3 counts every transmit–receive band pair exactly once, and since p and q each sum to one, C^2 is a weighted average of $\cos^2\theta_{k\ell}$. The coupling is therefore high only when the sender's transmission strength and the receiver's amplification concentrate on the same band pairs and those band pairs are tuned alike.

Properties of the coupling coefficient. Three properties of C support the census (proofs in Appendix A.3).

1. *Rotation invariance.* C depends on the pair only through the Grams $G = R^\top R$ and $H = WW^\top$ (Proposition 2, Corollary 3). Thus, a rotation of the d_{head}-dimensional head space,

[3] From here on, we call the three matrices (W_E, W_U, and W_{pos} if it exists) *interface matrices*, as they handle the model's token inputs and final token outputs.

Table 1: **Candidate-edge summary of 7 mapped models**. Counts are causally eligible ordered pairs, and *wires* are defined as neuron→neuron pairs with $|\cos| \geq 0.5$. The rotary Pythia models lack the 4 classes where the positional embedding matrix is the writer. (ifc = interface matrices).

Model	head–head	ifc↔head	head↔neuron	ifc↔neuron	neuron–neuron	Total	Wires
GPT-2	28,512	1,008	1.02×10^7	110,592	6.23×10^8	6.33×10^8	2,066
GPT-2-medium	211,968	2,688	7.39×10^7	294,912	4.63×10^9	4.70×10^9	3,124
GPT-2-large	756,000	5,040	2.62×10^8	552,960	1.65×10^{10}	1.68×10^{10}	6,001
GPT-Neo-125M	28,512	1,008	1.02×10^7	110,592	6.23×10^8	6.33×10^8	376
Pythia-160m	28,512	576	1.02×10^7	73,728	6.23×10^8	6.33×10^8	725
Pythia-2.8B	1,523,712	4,096	6.61×10^8	655,360	5.20×10^{10}	5.27×10^{10}	2,389
Pythia-6.9B	1,523,712	4,096	1.06×10^9	1,048,576	1.33×10^{11}	1.34×10^{11}	1,837

which changes a head's internal bookkeeping but not its effect on the residual stream, leaves both Gram matrices and hence C unchanged.

2. *Scale.* $C \in [0, 1]$, and for rank-one pairs it collapses to $|\cos|$, so neuron wires and head pairs are measured on a common scale (Corollary 4).
3. *Chance level.* Under a Haar rotation of either side, $\mathbb{E}[C^2] = 1/d$ exactly, for any singular values, giving one baseline for all classes of connections at once (Proposition 5). For rank-one writer–reader pairs (i.e., neuron–neuron connections), C^2 follows a $\mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$ distribution (Lemma 6).

Computing the map. We compute the map across seven LLMs of varying scales in three model families (see summary in Table 1): GPT-2, GPT-Neo, and Pythia. Each pair's C^2 is standardized against its rotation null distribution, whose mean and variance are closed-form (Appendix A.4). We use the cyclic property of the trace to compute C efficiently, collapsing every class to the much smaller d_{head}-wide products instead of naively multiplying two $d \times d$ matrices for each of the billions of candidate pairs (see Appendix B for details). This makes mapping whole transformers feasible. The full map costs only 15 seconds for GPT-2 small and 11 minutes for Pythia-6.9B on one consumer GPU.

Table 2: **Head-to-head pairs against the theoretical null distribution**, for every mapped model. Share of head pairs significantly above ($z \geq 2$) and below ($z \leq -2$) chance orientation, per composition channel, with the median z of each group in parentheses. Here $z = \left(C^2 - 1/d\right)/\sqrt{\mathrm{Var}[C^2]}$, and both moments are closed-form (Appendix A.4).

Model	K-composition above / below	Q-composition above / below	V-composition above / below
GPT-2	55% (+16) / 34% (−8)	61% (+18) / 27% (−7)	25% (+15) / 65% (−10)
GPT-2-medium	54% (+12) / 28% (−7)	62% (+15) / 20% (−6)	25% (+13) / 58% (−9)
GPT-2-large	63% (+13) / 23% (−6)	70% (+14) / 15% (−5)	32% (+12) / 51% (−7)
GPT-Neo-125M	40% (+8) / 39% (−6)	48% (+9) / 33% (−7)	24% (+8) / 57% (−8)
Pythia-160m	50% (+7) / 27% (−7)	65% (+9) / 16% (−5)	21% (+9) / 63% (−10)
Pythia-2.8B	55% (+9) / 14% (−4)	52% (+8) / 12% (−3)	27% (+9) / 50% (−7)
Pythia-6.9B	84% (+14) / 4% (−4)	50% (+7) / 10% (−3)	31% (+11) / 50% (−7)

Table 2 reveals several patterns. Most K- and Q-compositions have C^2 values that are either significantly above ($z \geq 2$) or below ($z \leq -2$) the chance level, suggesting these writer–reader head pairs are either *super-coupled* or actively *avoidant* of each other by communicating in orthogonal subspaces. For instance, 89% of GPT-2 small's head pairs in the K-composition deviate beyond two standard deviations from the chance level, with 55% *super-coupled* and 34% *avoidant*. Q-composition follows a similar divide. The same pattern replicates across all seven mapped models.

The V-composition, in contrast, is more avoidant (65% in GPT-2 small) than super-coupled (25%). Why V-composition holds more avoidant pairs than super-coupled ones is an open question for future research. Elhage et al. (2021) observed that V-composition is rare in their toy-scale attention-only models. Our evidence is consistent with that of Elhage et al. (2021) as we find active suppression in the V-composition channel at every scale we map, from GPT-2 small to Pythia-6.9B.

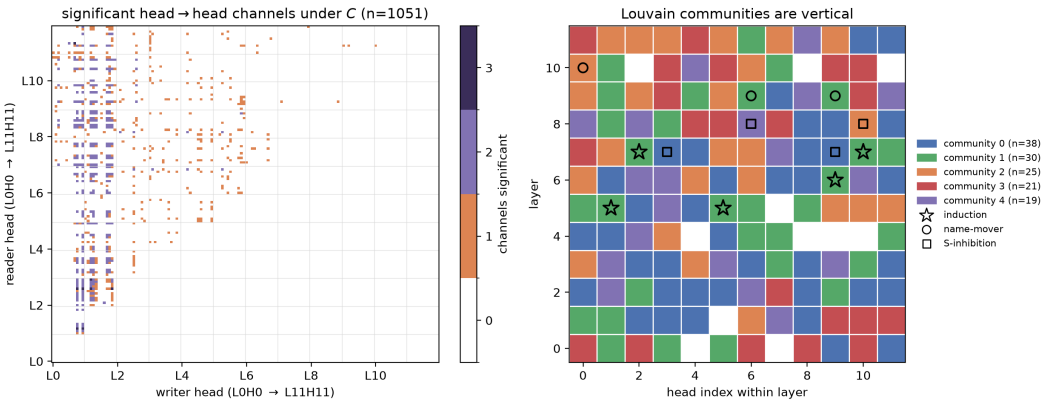

Figure 2: **The head-to-head subgraph of the communication map (GPT-2).** *Left:* adjacency of the 1,051 selected head→head channel edges under C, one cell per ordered pair, darker where more of the three channels are selected, layer 0 at the lower left. *Right:* the seed-0 Louvain partition of the same graph (assignment archived), one cell per head, with the five induction heads (⋆), the IOI name-movers (•), and the S-inhibition heads (■) marked. White cells are heads outside the giant component.

Neuron wires. The mapping census extends to the 6.2×10^8 neuron→neuron pairs of GPT-2 and the 1.3×10^{11} of Pythia-6.9B. The result indicates that neuron-to-neuron pairings have couplings far exceeding what random directional alignment would produce. Under the null distribution of $C = |\cos|$, the expected number of pairs with $|\cos| \geq 0.2$ is 14, and the expected number of pairs with $|\cos| \geq 0.23$ is essentially zero (Corollary 7). In reality, GPT-2 small holds 140,927 pairs whose $|\cos| \geq 0.2$, and 78,861 whose $|\cos| \geq 0.23$. So neurons are directly wired to each other. Similar patterns persist at every model scale tested (Table 1; Figure 3, Appendix A.4).

4 Application 1: Head-to-Head Communities

In the first application of the communication map, we investigate the graph formed by the strongest head-to-head couplings of an entire transformer, and check whether the subnetworks of this graph correspond to known circuits.

Edge selection. We select, within each stratum of head-to-head pairs (Appendix C.1), the pairs coupled far beyond the typical degree of coupling for that stratum, and use those pairs as the edges of the head graph. The theoretical null cannot serve as the selection standard, since 47% of GPT-2's head pairs exceed it at $z \geq 2$ (Table 2). Instead, we ask which pairs are exceptionally strong against each stratum's *empirical null distribution* (Efron, 2004), using the robust z-score:

$$z_e = \frac{C_e - \mathrm{med}_s}{1.4826 \cdot \mathrm{MAD}_s}, \qquad \mathrm{MAD}_s = \mathrm{med}_{e' \in s} \left| C_{e'} - \mathrm{med}_s \right|, \tag{4}$$

where in each stratum s, the numerator of edge e is C_e minus the stratum's median med_s, and the robust standard deviation is 1.4826 times the median absolute deviation (MAD) (Hampel, 1974; Rousseeuw & Croux, 1993). Within each connection class, Benjamini–Hochberg control at $q = 0.05$ (Benjamini & Hochberg, 1995) then sets the class's selection threshold adaptively, so that the edges the map draws carry a false discovery rate (FDR) of at most 5% per class. On GPT-2 the resulting head-class thresholds fall between $z = 2.8$ and $z = 3.2$, the 99.75th to 99.94th percentiles of the fitted empirical null distributions.

Head graph. Figure 2 (Left) shows the resulting head-to-head communication graph for GPT-2's 144 attention heads. Each ordered head pair carries three separately tested channels (K-, Q-, and V-composition), and 1,051 of the 28,512 candidate edges are selected at FDR $q = 0.05$ under C. Layers 0–1 mostly hold *broadcasting* heads whose writes are read by many later heads, while middle and later layers mostly hold *receiver* heads that read from the earlier layers with little broadcasting of their own.

Head communities. We then compute a Louvain partition of the head-to-head subgraph (Blondel et al., 2008), which yields the five communities shown in Figure 2 (Right). The partition is stable across ten Louvain seeds. Three patterns emerge: *First*, the communities are vertical. They span the depth of the model rather than a single layer, and no community spans fewer than eight of GPT-2's twelve layers. *Second*, communication across heads appears long-range. The median writer-to-reader separation is five layers against four for the eligible pairs, and the rate of wired pairs roughly doubles with distance (Table 11), consistent with the predictions of Elhage et al. (2021). *Third*, known circuits concentrate in certain communities. Figure 2 (Right) shows that the green community with 30 heads holds all five induction heads and two of the three IOI name-mover heads in ten of ten seeds. The induction and IOI heads are identified behaviorally following the protocols of prior work (Olsson et al., 2022; Wang et al., 2023). The previous-token head L4H11 is the top-ranked K-composition writer into every induction head.

The community ablation experiment. We conduct an experiment to investigate the effect of ablating the induction-concentrated community (e.g., the green community in Figure 2) on a model's in-context copying capability. We measure induction capability by the *induction gain*, which is the drop in next-token loss between a block of random tokens and an immediately following copy of it. Full results are presented in Table 9 and Table 10 in Appendix C.

On GPT-2 small, freezing the full thirty-head community's outputs at corpus means destroys 93.8% of the model's induction gain, against a median of 2.5% for random control sets matched in size and per-layer head counts, suggesting that the ablated community of heads is indeed responsible for virtually all induction in the model. Furthermore, the five behaviorally identified induction heads collectively account for only 60.8% of the induction gain, so the community's remaining members, individually near-inert, carry redundant machinery that single-head identification misses. The ablation results replicate across six models as shown in Table 9, reaching 97.1% against a 1.0% control median at Pythia-2.8B.

5 Application 2: induction-critical subspace

In this second application, we apply the communication mapping method to the problem of identifying induction-critical dimensions in the residual stream of a transformer. Instead of focusing on individual heads or neurons, here we demonstrate how the same mapping method can be used to identify communication channels affecting the entire transformer at once. This application also connects the map to the existing literature on important directions in the residual stream, activation PCA, and outlier dimensions (see Baselines below).

Reader–writer PCA. We develop a method, called *reader–writer PCA* (RW-PCA), to identify an induction-critical subspace in the residual stream of a transformer from the communication map directly. We exclude the MLP neurons as we focus primarily on attention heads in this analysis. RW-PCA differs from activation PCA (Mu & Viswanath, 2018; Ethayarajh, 2019) in that the eigendecomposition is applied to a transformer's reader and writer matrices rather than sampled activations the model generates in forward passes. In other words, RW-PCA is based on the inherent *geometry of a transformer's matrices*.

The RW-PCA method. An attention head touches the stream through its four factor matrices $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d\times d_{\text{head}}}$, each reading or writing a d_{head}-dimensional slice of it (Appendix A.1), and the interface matrices touch it directly. For any one of these matrices X, scored against a unit stream direction v playing the writer, Eq. 2 squares to $C^2(X, v) = \|X^\top v\|^2/\|X\|_F^2 = v^\top G_X\, v/\operatorname{tr}(G_X)$ with $G_X := XX^\top$, a quadratic form in the unit-trace Gram. The stream directions that matter globally are those maximizing this coupling summed over every attention-head factor and interface matrix, $\sum_X C^2(X, v) = v^\top S v$, which defines the *pooled coupling matrix*

$$S = \sum_X \frac{G_X}{\operatorname{tr}(G_X)}. \tag{5}$$

The trace normalization gives each factor one vote (Appendix D.1). Maximizing $v^\top S v$ over orthonormal directions is the Rayleigh quotient problem, so the maximizers are the eigenvectors of S.

Table 3: **The deletion protocol and subspace comparison at six scales.** Condition cells are percentages of induction gain destroyed / Pile ΔNLL in nats, for RW-PCA (ours), activation PCA (Mu & Viswanath, 2018; Ethayarajh, 2019), and outlier dimensions (Timkey & van Schijndel, 2021; Kovaleva et al., 2021; Dettmers et al., 2022). The last two columns are the two principal cosines between the RW-PCA subspace and each baseline's subspace (1 = identical, 0 = orthogonal). The control deletes a random 2-dimensional subspace (median of five draws). GPT-Neo is excluded because random deletion already removes 60–98.1% of its induction gain.

					Cosines: ours vs	
Model	RW-PCA (ours)	Activation PCA	Outlier dims	Random	Act. PCA	Outliers
GPT-2	92.3% / +1.6	87.4% / +7.0	57.5% / +4.2	0.2% / +0.1	.29, .07	.08, .00
GPT-2-medium	82.5% / +3.4	88.5% / +7.7	71.2% / +5.1	1.3% / +0.1	.25, .01	.16, .00
GPT-2-large	81.8% / +1.3	86.6% / +3.4	−19.0% / +1.8	0.1% / +0.0	.50, .05	.05, .01
Pythia-160m (rotary)	95.8% / +8.1	42.0% / +3.8	35.4% / +2.4	3.0% / +0.2	.66, .12	.59, .00
Pythia-2.8B (rotary)	85.9% / +9.7	78.0% / +7.2	83.5% / +6.9	0.2% / +0.0	.26, .02	.21, .01
Pythia-6.9B (rotary)	89.5% / +8.6	84.5% / +8.2	36.5% / +8.4	0.2% / +0.0	.83, .26	.84, .34

These eigenvectors can be regarded as the global communication *bands* that most heads share, and the ten leading bands are the candidate pool for later analysis.

Selecting the most induction-critical bands. Induction is *positional*, as induction circuits generally match the current token to its previous occurrence and copy what followed (Elhage et al., 2021; Olsson et al., 2022), so whatever performs induction must read and write positional signals. Therefore, from the ten global bands identified above, we select as the induction-critical pair the two most position-specific bands, those with the highest ratio of positional coupling to token coupling, defined as

$$\mathrm{PosRatio}(v_i) = \frac{C^2_{pos}(v_i)}{C^2_{tok}(v_i)}. \tag{6}$$

The numerator and denominator are band v_i's squared *positional coupling* with the positional matrix Π and its squared *token coupling* with the token embedding matrix W_E,

$$C^2_{\mathrm{pos}}(v_i) \;=\; \frac{\|\Pi^\top v_i\|^2}{\|\Pi\|_F^2}, \qquad C^2_{\mathrm{tok}}(v_i) \;=\; \frac{\|W_E^\top v_i\|^2}{\|W_E\|_F^2}. \tag{7}$$

We divide C^2_{pos} by C^2_{tok} in Eq. 6 because the deletion is meant to be selective. Deleting content-carrying directions degrades prediction indiscriminately (Kovaleva et al., 2021; Dettmers et al., 2022), so the ratio reflects the stated goal, selecting the most positional and least content-carrying directions. For GPT-2, the *positional matrix* Π is just the positional embedding matrix W_{pos}. But for models using RoPE, which have no positional embedding matrix at all, we estimate Π by sampling the residual stream with 32 sequences of 128 random tokens each. The full implementation details are in Appendix D.2.

Induction-critical subspace deletion. Once the two-dimensional induction-critical subspace is identified, we delete it via a projection in the following three steps: (1) Stack the two chosen directions as the columns of an orthonormal matrix $V \in \mathbb{R}^{d\times 2}$. (2) Remove this subspace V from the residual stream at the input of every layer and before the final unembedding, by replacing the residual stream's activation x by $x - VV^\top x$. (3) Measure the induction gain of the transformer without this two-dimensional subspace, as well as the rise in natural-text loss on the Pile.

Baselines. As baselines, we report two other key methods from the prior literature for identifying important directions in the residual stream. *Activation PCA* takes the top two principal components of the mid-layer residual stream on natural text, and removing such components is an established postprocessing operation (Mu & Viswanath, 2018; Ethayarajh, 2019). *Outlier dimensions* are the two stream coordinates with the largest activation RMS, the rogue dimensions of the outlier literature, documented for GPT-2 as dimensions 447 and 138 (Timkey & van Schijndel, 2021; Kovaleva et al., 2021; Dettmers et al., 2022). Subspace deletions are run on these two baselines, and the results are reported in Table 3.

Distinct directions. Table 3 (last two columns) reports the two principal cosines (Björck & Golub, 1973) between the RW-PCA plane and each of these planes on all six models. The two directions identified by our method are correlated with both baselines, to varying degrees, and also distinct from both. While the two baselines nearly coincide with each other, with first cosines of 0.93–0.998 on five of six models, RW-PCA's second principal cosine against either baseline never exceeds 0.35 and is usually below 0.1. On every model, our method identifies at least one principal direction not captured by either baseline, confirming that the proposed mapping method can identify induction-critical dimensions distinct from those of the prior literature.

Deletion outcomes. As Table 3 illustrates, deleting our two induction-critical directions destroys 82–96% of the induction gain on every model. In comparison, the baseline methods are either inconsistent or less effective at destroying induction. In the Pythia-6.9B case, projecting out only 2 of its $4{,}096$ dimensions, merely 0.05% of the residual stream's width, abolishes 89.5% of its induction capability. Activation PCA achieves 84.5% and the outlier-dimension method only 36.5%. The contrast is starker in Pythia-160m, where RW-PCA destroys 95.8% of induction, but activation PCA destroys only 42% and the outlier-dimension method destroys only 35.4%. Random subspace deletions, as a control, show zero or near-zero effect across all six models. Furthermore, a deletion-dose curve in the released results confirms that setting the subspace dimension to two is a minimal and effective choice at the scales we tested.

In summary, the map identifies a two-dimensional subspace, distinct from those of the prior literature, that carries a transformer's in-context copying capability at every scale tested. Unlike activation PCA, whose components depend on the text corpus and the layer sampled, RW-PCA derives its bands from the model's weights alone, and it is the only one of the three methods whose deletion effect holds at every scale.

6 Conclusion

In this paper, we build a communication map of an entire transformer, covering every head, neuron, and eligible channel, from the geometry of the transformer's weight matrices. We then demonstrate two novel applications of the map. In the first application, we construct a head-to-head communication graph and find that heads group into vertical communities with induction and IOI circuits concentrated in one community. In the second application, we identify a few global communication bands in the residual stream that most of the transformer's heads share, and find that two bands in particular form an induction-critical subspace. Our findings replicate across six models, ranging from GPT-2 small to Pythia-6.9B.

This research has several limitations. First, the map charts potential connectivity from model weights alone, showing which pairs are geometrically wired to communicate. It does not show, however, how much they actually communicate on any given input. While our ablation results show that deleting map-drawn structures destroys induction capability in a model, measuring which channels carry traffic on a given input, and how much, requires activation-level methods and is left to future work. Second, our applications validate the map against a single capability, in-context copying, and focus on head-to-head structure. The neuron wires and the remaining communities the map exposes await investigation of their own. Third, our mapping and intervention experiments are limited to the sub-10B-parameter model scale. Fourth, the map's nodes are heads and neurons, not learned features. A feature-level map, using sparse-autoencoder or transcoder units (Bricken et al., 2023; Huben et al., 2024; Dunefsky et al., 2024), is a natural extension for future research.

Reproducibility statement

All models are public checkpoints accessed via TransformerLens (Nanda & Bloom, 2022). Natural-text corpora are fixed Pile samples (Gao et al., 2020). The full pipeline is included in the supplementary material, with fixed seeds, and every number regenerates from the released code and map tables. Appendix B lists decisions, corpora, and compute. All experiments run on a single NVIDIA RTX 5090 (32 GB). The full GPT-2 map builds in 15 s, the Pythia-6.9B map in 11 minutes, and the ablation experiments run in minutes per model. The 6.9B rows need the 32 GB for fp32 forwards. Every other number reproduces on smaller cards.

AI USE STATEMENT

In this work, we used generative AI tools for assistance in the writing of proofs, providing critical ingredients for proving mathematical claims, providing feedback on research methodology, interpreting results, and efficient implementation of methods (algorithms). We have not used generative AI tools for formulating mathematical claims, designing research methodology or experiments, or assisting with translation. Furthermore, generating synthetic data sets, developing theoretical models or conceptual frameworks, proposing or refining hypotheses, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for coding assistance, creating figures in the paper, identifying and summarizing the relevant literature, copyediting and formatting. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. https://doi.org/10.5281/zenodo.5297715, 2021.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, et al. Causal scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, et al. All circuits lead to rome: Rethinking functional anisotropy in circuit and sheaf discovery for LLMs. *arXiv preprint arXiv:2605.12671*, 2026.

Benoît Collins and Piotr Śniady. Integration with respect to the haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006.

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*, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, 2024.

Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

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*, 2021. `https://transformer-circuits.pub/2021/framework/index.html`.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, 2019.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, 2021.

Zhiren Gong, Zihao Zeng, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. *arXiv preprint arXiv:2607.01940*, 2026.

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.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*, 2024.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025. arXiv:2403.19647.

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

Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*. Cambridge University Press, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.09519.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, 2019.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

Neel Nanda and Joseph Bloom. TransformerLens. `https://github.com/TransformerLensOrg/TransformerLens`, 2022.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

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.

Wei Tang, Xinyan Jiang, Fakhri Karray, and Lijie Hu. In-context learning operates as concept subspace learning. *arXiv preprint arXiv:2605.18830*, 2026.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of EMNLP*, 2021.

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

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL*, 2019.

Kevin Ro 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.

Yongzhong Xu. Closure-validated circuit discovery in attention heads: Co-activation proposes, ablation disposes. *arXiv preprint arXiv:2606.09607*, 2026.

Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. Measuring affinity between attention-head weight subspaces via the projection kernel. *arXiv preprint arXiv:2601.10266*, 2026.

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

A MATHEMATICAL APPENDIX

A.1 CONVENTIONS

Stream vectors $x \in \mathbb{R}^d$ are columns, and all maps act by left multiplication. All four per-head weight matrices are stored uniformly as factors $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d \times d_{\text{head}}}$. Reading projects into head space through a transpose ($W_X^\top x \in \mathbb{R}^{d_{\text{head}}}$ for $X \in \{Q, K, V\}$), and writing returns through left multiplication ($W_O z \in \mathbb{R}^d$). Consequently $W_{QK} = W_Q W_K^\top$ and $W_{OV} = W_O W_V^\top$ are $d \times d$ with rank at most d_{head}. The letters are mnemonic and mark a type distinction. W_{QK} is a bilinear *form* (two reading bases, analyzed by SVD, never applied as an operator), whereas W_{OV} is a stream-to-stream *map* (analyzed by eigendecomposition, which is what makes copying scores possible). Common implementations, including TransformerLens (Nanda & Bloom, 2022), store activations as row vectors with weights multiplying on the right, where W_Q, W_K, W_V already have the factor shape and W_O must be transposed. Elhage et al. (2021) write the same operator as $W_{OV} = W_O W_V$ with map-shaped factors, the identical object. Our loading code performs the conversion once and asserts the resulting W_{QK} and W_{OV} against TransformerLens's own factored QK and OV circuits, so a silently transposed map is caught at load time. Table 4 collects the shapes, ranks, and roles of every per-component matrix in one place.

Table 4: **Shapes, ranks, and roles of the per-component matrices** (column convention). Ranks are generic, meaning that with trained weights all d_{head} singular values are numerically nonzero, though *effective* ranks are often far smaller.

Matrix	Shape (general)	Shape (GPT-2)	Rank	Role
W_Q	$d \times d_{\text{head}}$	768×64	$\le d_{\text{head}}$	query factor: $q = W_Q^\top x$
W_K	$d \times d_{\text{head}}$	768×64	$\le d_{\text{head}}$	key factor: $k = W_K^\top x$
W_V	$d \times d_{\text{head}}$	768×64	$\le d_{\text{head}}$	value factor: $v = W_V^\top x$
W_O	$d \times d_{\text{head}}$	768×64	$\le d_{\text{head}}$	output factor: head space $\to$ stream, $\Delta = W_O z$
$W_{QK} = W_Q W_K^\top$	$d \times d$	768×768	$\le d_{\text{head}}$	QK circuit: bilinear form on position pairs
$W_{OV} = W_O W_V^\top$	$d \times d$	768×768	$\le d_{\text{head}}$	OV circuit: stream $\to$ stream operator
r_m	$d \times 1$	768×1	1	neuron read direction: $a_m = r_m^\top x$
w_m	$d \times 1$	768×1	1	neuron write direction: $\Delta_m = g(a_m)\, w_m$

A.2 THE ATTENTION BLOCK, DERIVED TO THE FACTORED FORM

Stacking positions as columns of $X \in \mathbb{R}^{d \times n}$, the head is $\Delta X = W_{OV,h}\, X A^\top$, where the attention matrix A acts only on the position axis and is recomputed on every forward pass, while $W_{QK,h}$ (through the logits) and $W_{OV,h}$ act only on the feature axis and are frozen weights. The map charts the feature axis (Section 2).

The four-term logit split behind K- and Q-composition. The reader's attention logit $\ell_{ij} = x_i^\top W_{QK,r}\, x_j / \sqrt{d_{\text{head}}}$ is bilinear in the two stream vectors x_i and x_j, and an upstream writer contributes to both, since it writes at every position. Let $W_{OV,w} u_i$ denote the writer's output at position i, with u_i its attention-averaged input there, and split $x_i = \bar{x}_i + W_{OV,w} u_i$ into the writer's write and the rest of the stream (same for x_j). Bilinearity gives, exactly,

$$\begin{aligned}\sqrt{d_{\text{head}}}\, \ell_{ij} \;=\; & \bar{x}_i^\top W_{QK,r}\, \bar{x}_j \;+\; \bar{x}_i^\top W_{QK,r}\, (W_{OV,w} u_j) \\ & + (W_{OV,w} u_i)^\top W_{QK,r}\, \bar{x}_j \;+\; (W_{OV,w} u_i)^\top W_{QK,r}\, (W_{OV,w} u_j).\end{aligned} \tag{8}$$

The first term does not involve the writer. The second is the writer's entry through the key side, governed by the operator $W_{QK,r} W_{OV,w}$ (K-composition). The third is the writer's entry through the query side, governed by $W_{QK,r}^\top W_{OV,w}$ (Q-composition). The two cross terms are not exclusive cases but coexisting channels of one exact expansion. The fourth term routes the writer into both sides at once, through the operator $W_{OV,w}^\top W_{QK,r} W_{OV,w}$.

The both-sides term adds no channel of its own. Its operator contains the writer's matrix twice, so its normalized score carries $\|W_{OV,w}\|_F^2$ in the denominator. The Frobenius norm is submultiplicative, $\|XY\|_F \le \|X\|_F \|Y\|_F$ (Horn & Johnson, 1991), and applying it to the grouping

$W_{OV,w}^\top(W_{QK,r}W_{OV,w})$ gives

$$\frac{\|W_{OV,w}^\top W_{QK,r}\, W_{OV,w}\|_F}{\|W_{QK,r}\|_F\,\|W_{OV,w}\|_F^2} \le \frac{\|W_{OV,w}\|_F\,\|W_{QK,r}W_{OV,w}\|_F}{\|W_{QK,r}\|_F\,\|W_{OV,w}\|_F^2} = \frac{\|W_{QK,r}W_{OV,w}\|_F}{\|W_{QK,r}\|_F\,\|W_{OV,w}\|_F} = C_K, \tag{9}$$

after one factor of $\|W_{OV,w}\|_F$ cancels, so the both-sides score is bounded by exactly the K-composition coupling. The other grouping, $(W_{OV,w}^\top W_{QK,r})\,W_{OV,w}$, gives the mirror bound by the Q-composition coupling C_Q. A vanishing one-sided channel therefore forces the both-sides term to zero, so the map scores two QK edge classes per head pair rather than three. Numerically, the bound holds with a wide margin. Across all $9{,}504$ head pairs of GPT-2 the normalized both-sides score never exceeds its bound, and its ratio to the smaller of the two one-sided couplings, $\min(C_K, C_Q)$, is at most 0.35 (script released).

A.3 The coupling coefficient

Proposition 1 (the weighted-cosine identity)**.** *Let $R = U\Sigma V^\top$ and $W = ASB^\top$ be the two SVDs, with singular values σ_k and s_ℓ, and let $\cos\theta_{k\ell} := v_k^\top a_\ell$ pair the reader's input directions with the writer's output directions. With the gain shares $p_k := \sigma_k^2/\sum_{k'}\sigma_{k'}^2$ and $q_\ell := s_\ell^2/\sum_{\ell'} s_{\ell'}^2$,*

$$C^2 \;=\; \sum_{k,\ell} p_k\, q_\ell\, \cos^2\theta_{k\ell}.$$

Proof. The composed operator factorizes as $RW = U\,\Sigma\,(V^\top A)\,S\,B^\top = U\,(\Sigma\Theta S)\,B^\top$ with the cosine table $\Theta := V^\top A$, $\Theta_{k\ell} = \cos\theta_{k\ell}$. The outer factors have orthonormal columns and preserve the Frobenius norm, since $\mathrm{tr}(BX^\top U^\top UXB^\top) = \mathrm{tr}(X^\top X)$ by cyclicity, so $\|RW\|_F^2 = \|\Sigma\Theta S\|_F^2$. The left diagonal factor Σ scales the rows of Θ, and the right diagonal factor S scales the columns of Θ. Thus $(\Sigma\Theta S)_{k\ell} = \sigma_k\cos\theta_{k\ell}\,s_\ell$, and summing squared entries gives $\|RW\|_F^2 = \sum_{k,\ell}\sigma_k^2 s_\ell^2\cos^2\theta_{k\ell}$. The same norm preservation applied to each SVD alone gives the denominators, $\|R\|_F^2 = \|\Sigma\|_F^2 = \sum_k\sigma_k^2$ and $\|W\|_F^2 = \|S\|_F^2 = \sum_\ell s_\ell^2$, and dividing the squared numerator by both gives Eq. 3. □

Proposition 2 (Gram form)**.**

$$C^2 \;=\; \frac{\|RW\|_F^2}{\|R\|_F^2\,\|W\|_F^2} \;=\; \frac{\mathrm{tr}(GH)}{\mathrm{tr}(G)\;\mathrm{tr}(H)}, \qquad G := R^\top R, \quad H := WW^\top.$$

Proof. By cyclicity of the trace,

$$\|RW\|_F^2 = \mathrm{tr}(W^\top R^\top RW) = \mathrm{tr}\,\big(\underbrace{R^\top R}_{G}\,\underbrace{WW^\top}_{H}\big),$$

and the normalizers are $\mathrm{tr}\,G = \mathrm{tr}(R^\top R) = \|R\|_F^2$ and likewise $\mathrm{tr}\,H = \|W\|_F^2$. □

The numerator is in fact the Frobenius inner product of the two Grams, since $\mathrm{tr}(GH) = \mathrm{tr}(G^\top H) = \langle G, H\rangle_F$ by the symmetry of G. The denominator is a choice. Normalizing by the traces, rather than by Cauchy–Schwarz with $\|G\|_F\,\|H\|_F$, is what yields the weighted-cosine reading of Eq. 3 and the universal chance level of Appendix A.4. Under the trace normalization a perfectly matched isotropic pair, $R = W = I_d$, scores $C^2 = 1/d$, exactly the chance level of Proposition 5, while the Cauchy–Schwarz normalization would award the same pair its maximal score. The trace verdict is the intended one, because a pair coupled equally over every direction of the stream exhibits none of the selective alignment that defines an edge (Section 2).

Corollary 3 (invariance)**.** *C depends on the pair only through (G, H). It satisfies $C(\alpha R, \beta W) = C(R, W)$ for any nonzero scalars and $C(UR,\, WV) = C(R, W)$ for any orthogonal U, V.*

Proof. $(UR)^\top(UR) = G$ and $(WV)(WV)^\top = H$, and scalars cancel in the ratio. □

These invariances are the appropriate blindness for a communication statistic. U rotates the reader's internal encoding after reading, and V rotates the writer's internal channels. Neither is visible on the residual stream, so neither affects an edge score. Only the listened-to directions (G) and the written-onto directions (H) survive.

Corollary 4 (range and extremes)**.** $C \in [0, 1]$. *Moreover: (i)* $C = 0$ *if and only if* $RW = 0$*, equivalently* $\mathrm{Row}(R) \perp \mathrm{Col}(W)$*; (ii)* $C = 1$ *if and only if* R *and* W *both have rank one and the reader's input direction coincides with the writer's output direction up to sign; (iii) for rank-one pairs,* $C = |\cos \angle(r_m, w_{m'})|$.

Proof. By Proposition 1, $C^2 = \sum_{k,\ell} p_k\, q_\ell \cos^2 \theta_{k\ell}$ is a convex combination of the values $\cos^2 \theta_{k\ell}$, since the weights $p_k q_\ell$ are nonnegative and sum to one, so $C \in [0, 1]$. Part (i) is immediate from the definition, because $\|RW\|_F = 0$ if and only if $RW = 0$. For (ii), the convex combination equals 1 if and only if $\cos^2 \theta_{k\ell} = 1$ for every pair with $p_k q_\ell > 0$. If p had two nonzero entries $k \neq k'$, the directions v_k and $v_{k'}$ would both coincide with the same unit vector a_ℓ up to sign, contradicting their orthogonality. Hence p and q each have a single nonzero entry, both matrices have rank one, and the two surviving directions satisfy $\cos^2 \theta = 1$, meaning they coincide up to sign. For (iii), the SVD of a vector has a single singular direction, so the sum has one term with $p_1 = q_1 = 1$. □

At $C = 0$ the reader receives nothing the writer transmits, and the rank-one case is what puts neuron wires and head circuits on one common scale.

Why the K-composition reader is $W_{QK,r}$, not W_K. Eq. 3 also justifies the choice of reading matrix. The key-side reading directions are the right singular vectors v_k of $W_{QK,r}$, and each is weighted by σ_k, which measures how strongly the paired query direction u_k listens. A key channel the query side never matches has $\sigma_k \approx 0$ and contributes nothing to logits, and Eq. 2 counts it as nothing. A reader built from W_K alone would count it fully.

A.4 Nulls

The rotation null distribution of the main text is, in the terminology of Efron (2004), the *theoretical null distribution*. Application 1's selection standard, the *empirical null distribution*, is documented with its tooling in Appendix C.1.

Proposition 5 (universal chance level)**.** *Under a Haar rotation of either side,* $\mathbb{E}[C^2] = 1/d$ *exactly, for any reader and writer singular values.*

Proof. Writing $\hat{H} = H/\operatorname{tr}(H)$, invariance of the Haar measure forces $\mathbb{E}_Q[Q\hat{H}Q^\top] = I/d$, a classical averaging identity (Collins & Śniady, 2006, Cor. 3.4; Meckes, 2019, §2.1), so, by linearity of the trace and of expectation, $\mathbb{E}[C^2] = \operatorname{tr}(\hat{G}\,\mathbb{E}_Q[Q\hat{H}Q^\top]) = 1/d$. □

Proposition 5 pins down the mean of the null distribution and nothing more. How C^2 scatters around $1/d$ from one rotation to the next still depends on the singular values. A writer whose singular values are all equal has $\hat{H} = I/d$, a matrix that every rotation leaves unchanged, so C^2 equals $1/d$ with no fluctuation at all. A rank-one pair is the opposite extreme. All of the gain lies in one band on each side, the statistic is the squared cosine between a fixed direction and a uniformly random one, and in this case the whole null distribution, not only its mean, has a classical closed form.

Lemma 6 (rank-one null law; classical, see Meckes, 2019, §2.1, Prop. 2.5)**.** *For a fixed unit vector* u *and* v *uniform on the unit sphere in* $\mathbb{R}^d$*, the squared cosine* $(u^\top v)^2$ *follows* $\mathrm{Beta}\big(\frac{1}{2}, \frac{d-1}{2}\big)$*, and under this distribution the standard deviation of the cosine is* $1/\sqrt{d} = 0.036$ *at* $d = 768$.

Corollary 7 (chance-law exceedance counts)**.** *By Lemma 6 the two-sided tail is* $P(|\cos| > t) = I_{1-t^2}\big(\frac{d-1}{2}, \frac{1}{2}\big)$*, the regularized incomplete beta function. Among the* $N = 622{,}854{,}144$ *neuron-pair draws from the null distribution, the expected count* $N \cdot P(|\cos| > t)$ *of pairs beyond* $t = 0.2$ *is therefore* 13.8*, beyond* 0.23 *it is* 0.07*, and beyond the wire threshold* 0.5 *it is* 4×10^{-41}*. The maximum of the* N *draws consequently has median* 0.217 *and 95th percentile* 0.231*, solving* $N \cdot P(|\cos| > t) = \ln 2$ *and* $-\ln 0.95$ *respectively.*

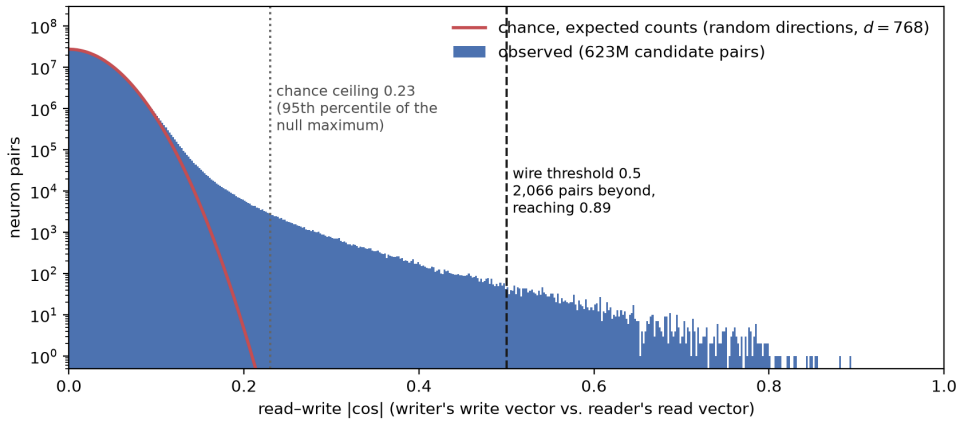

Figure 3: **The neuron→neuron coupling distribution.** Read–write $|\cos|$ over all 622,854,144 candidate neuron pairs (blue, log count scale) against the exact chance density for random directions in $d = 768$ (red), scaled to the same number of pairs. Chance and observation agree through the bulk. The dotted line marks 0.23, beyond which the chance law expects 0.07 pairs. The observed distribution continues to 0.89. The 2,066 pairs beyond 0.5 are the *wires*.

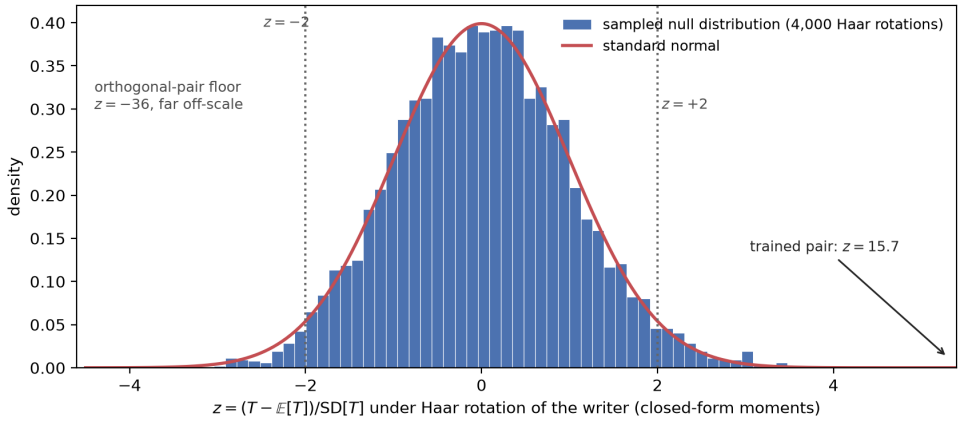

Figure 4: The sampled rotation null distribution of one head pair (writer L4H7 → reader L8H10 of GPT-2, 4,000 Haar rotations), standardized by the closed-form moments of Equation 11. The rank-64 null is a near-symmetric bump with mass in both tails (sampled mean 0.02, SD 1.00, skew +0.20), so the below-chance regime of Table 2 is reachable. The perfectly orthogonal pair sits at $z = -36$, far off-scale, in contrast with the rank-one law of Figure 3, whose floor of -0.7 makes $z \le -2$ unreachable for neuron pairs. The trained pair scores $z = 15.7$, far outside its own null distribution.

The wire tail persists at every scale. At 6.9B the largest observed $|\cos|$ is 0.78 against a 0.016 chance scale. On natural text the wired pairs co-activate at median $|\text{corr}| = 0.31$ against 0.03 for random pairs, so they are not dormant. Table 5 sets the observed exceedance counts against the chance law at each threshold.

Closed-form moments and the census. Lemma 6 covers the rank-one classes. A subspace pair of rank above one has no closed-form null law, because the rotation-null distribution of C depends on the singular values of both sides. Its first two moments are nevertheless closed-form, and the z-scores of the census need nothing more. Fix one head pair as trained, with reader Gram G and writer Gram H. Under the rotation null the squared statistic is $C^2 = T/(\operatorname{tr} G \operatorname{tr} H)$ with $T = \operatorname{tr}(GQHQ^\top)$ and Q Haar-random. The mean $\mathbb{E}[C^2] = 1/d$ is Proposition 5. The variance of T is the degree-two case of the same Weingarten calculus (Collins & Śniady, 2006, Cor. 3.4):

$$\operatorname{Var}[T] \;=\; \frac{2}{(d-1)(d+2)}\Big(\operatorname{tr} G^2 - \tfrac{(\operatorname{tr} G)^2}{d}\Big)\Big(\operatorname{tr} H^2 - \tfrac{(\operatorname{tr} H)^2}{d}\Big), \tag{10}$$

Table 5: **Exceedance counts over the neuron→neuron chance law (GPT-2).** For each threshold t: the single-pair tail $P(|\cos| > t)$ under the exact chance law (Lemma 6), the resulting expected number of pairs beyond t among all $N = 622{,}854{,}144$ candidate pairs, and the observed count in the trained weights.

t	$P(\lvert\cos\rvert > t)$	Expected $N \cdot P$	Observed
0.15	3.0×10^{-5}	1.8×10^{4}	511,605
0.20	2.2×10^{-8}	13.8	140,927
0.23	1.1×10^{-10}	0.07	78,861
0.25	2.0×10^{-12}	1.3×10^{-3}	55,991
0.30	1.9×10^{-17}	1.2×10^{-8}	26,100
0.50	7.0×10^{-50}	4.4×10^{-41}	2,066

and dividing by the constant $(\operatorname{tr} G \ \operatorname{tr} H)^2$ turns it into $\operatorname{Var}[C^2]$. Thus, we compute the standardized score (z-score) for C^2 of each writer–reader pair as

$$z \;=\; \frac{C^2 - 1/d}{\sqrt{\operatorname{Var}[T]} \,/\, (\operatorname{tr} G \ \operatorname{tr} H)} \;=\; \frac{T - \operatorname{tr} G \ \operatorname{tr} H/d}{\sqrt{\operatorname{Var}[T]}}, \tag{11}$$

with both moments exact, so no approximation enters. The second form is the one the released code computes. All four invariants collapse to inner-Gram products, $\operatorname{tr} G = \operatorname{tr}(g_Q g_K)$ and $\operatorname{tr} G^2 = \operatorname{tr}((g_Q g_K)^2)$ for a K-composition reader and likewise with $g_O g_V$ for V-composition and for writers, so every pair's z-score costs a few d_{head}-wide products at any model size. On GPT-2 the closed-form census reproduces a 500-rotation Monte Carlo census to a tenth of a percentage point on every fraction (verification below). Table 2 (Section 3) is the resulting census. Pairs suppressed below even the empirical selection threshold in magnitude concentrate almost exclusively in the V channel (28 on GPT-2, 5,229 at 2.8B, 4,179 at 6.9B).

The shares in Table 2 use only these exact moments and assume nothing about the shape of the null distribution. For the head classes the sampled null is in practice close to Gaussian, as the many independent subspace overlaps aggregate, while the rank-one classes follow the Beta law of Lemma 6, which is why the neuron census reports exceedance counts rather than z-tails.

Shared-rotation null distributions for the interface-matrix classes. Interface classes violate the sparsity assumption behind the empirical null distribution (Appendix C.1), because nearly every head genuinely reads the embedding, so their signal is dense rather than sparse. They are scored against shared-rotation null distributions. The interface matrix's Gram H is conjugated by Haar-random rotations, $H \to QHQ^\top$ with $Q \in O(d)$, the statistic is recomputed for each rotation, and each edge is scored as a z-score against the mean and standard deviation of that ensemble. The rotations are shared across all readers of an interface matrix, so the cost is a few hundred $d \times d$ conjugations rather than per-edge Monte Carlo. The ensemble is the isotropic reference, which is anti-conservative under anisotropy, so these classes are reported as effect-size rankings rather than sparse discoveries.

Monte Carlo confirmation. Both closed forms are verified by sampling the null distribution directly (scripts released). For a pair with Grams G and H, Haar-random rotations $Q_1, \ldots, Q_n \in O(d)$ give the ensemble

$$C_i \;=\; \sqrt{\frac{\operatorname{tr}\big(G\, Q_i H Q_i^\top\big)}{\operatorname{tr}(G)\ \operatorname{tr}(H)}}, \qquad i = 1, \ldots, n, \tag{12}$$

which keeps both sides' singular values and randomizes only the relative orientation of the read and write subspaces, so $\{C_i\}$ samples the pair's theoretical null distribution exactly. The ensemble's standard deviation, written s_{theory}, is the spread that chance alignment would produce for the pair. Across 20 combinations of reader and writer singular-value profiles at $d = 768$, sampled Haar rotations give $\mathbb{E}[C^2] = 1/d$ within sampling error ($|z| \le 1.6$), and profiles with equal singular values attain $1/d$ exactly with zero variance, confirming that the singular values control only the fluctuations. The rank-one cosine law matches $\operatorname{Beta}(\frac{1}{2}, \frac{d-1}{2})$ by Kolmogorov–Smirnov at $d = 3, 8, 768$ ($p = 0.55, 0.71, 0.71$), with the first four moments correct to four decimals.

B COMPUTING THE MAP: ALGORITHMS AND REPRODUCIBILITY

B.1 WEIGHT PREPROCESSING: LAYERNORM FOLDING

With layer normalization $\mathrm{LN}(\cdot)$, a transformer's sublayers read the transformed activation $\mathrm{LN}(x)$ from the residual stream, not the raw activation x. LayerNorm decomposes into a fixed linear map followed by an input-dependent scalar and a constant shift,

$$\mathrm{LN}(x) \;=\; \frac{1}{\sigma(x)}\,\mathrm{diag}(\gamma)\Big(I - \tfrac{1}{d}\mathbf{1}\mathbf{1}^\top\Big)\,x \;+\; \beta, \tag{13}$$

where $\sigma(x)$ is the standard deviation of the coordinates of x at that position and $\mathbf{1}$ is the all-ones vector. Both pieces of the linear factor are input-independent, so they fold into the weights. To account for LayerNorm, every matrix W that reads the stream through this LayerNorm is replaced by

$$\widetilde{W} \;=\; \Big(I - \tfrac{1}{d}\mathbf{1}\mathbf{1}^\top\Big)\,\mathrm{diag}(\gamma)\,W, \qquad \text{so that} \qquad W^\top \mathrm{LN}(x) \;=\; \frac{1}{\sigma(x)}\,\widetilde{W}^\top x \;+\; W^\top\beta. \tag{14}$$

The rescaling $\mathrm{diag}(\gamma)$ and the centering projector are thus absorbed into every reading matrix, and the constant $W^\top\beta$ folds into the sublayer's bias, which the map's statistics never touch. After folding, the only input dependence separating a read from the raw stream is the scalar gain $1/\sigma(x)$.

Writing matrices are treated with the complementary operation. Every write into the stream is centered along the stream dimension, removing its component in the $\mathbf{1}$ direction. This changes nothing about any coupling, because the projector now inside every reader annihilates that component anyway. Its purpose is to make the stored weights honest. A writer's matrix then contains only what some reader could ever receive, so the norms and couplings the map computes from writers measure transmissible signal and no dead mass. All couplings in the map are computed exactly on these folded and centered weights.

What folding cannot remove is the gain $1/\sigma(x)$, one scalar per position multiplying every coordinate of a read. Two consequences follow. A common scalar rescales all of a reader's incoming couplings equally, so the ranking of edges into a fixed reader is unaffected by it. What does vary with input is the overall strength of each write–read hop, so path strengths chained across layers carry a per-hop multiplicative uncertainty set by the spread of $\sigma(x)$.

B.2 FROM C^2 TO ALGORITHMS

This subsection and the three that follow show how the statistic of Eq. 2 becomes the released algorithms. Throughout, a *connection class* means one type of reader paired with one type of writer, for example head $\to$ neuron, every eligible pair of that type scored and calibrated together (Table 8 lists all 18). This subsection tabulates every reader and writer with its Gram and its normalizer. The next three each cover one shape of pair, both sides full matrices (Appendix B.3), one side a single vector (Appendix B.4), and both sides single vectors (Appendix B.5). These subsections introduce no new mathematics, only the definition of C unwound by trace cyclicity and applied at scale.

Ingredients. Every score in the map is assembled from a small set of precomputed ingredients. Each head contributes four factors $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d\times d_{\mathrm{head}}}$ (768×64 on GPT-2 small). From these we first precompute and cache the four $d_{\mathrm{head}} \times d_{\mathrm{head}}$ *inner Grams* per head, $g_Q := W_Q^\top W_Q$, $g_K := W_K^\top W_K$, $g_V := W_V^\top W_V$, $g_O := W_O^\top W_O$. The inner Grams are the reusable blocks from which the full Grams, the normalizers, and the mixed-class quadratic forms are all assembled. The factors also compose into the two per-head circuits, the QK circuit $W_{QK} := W_Q W_K^\top$, which scores query-key pairs, and the OV circuit $W_{OV} := W_O W_V^\top$, which is the head's write operator. The interface matrices contribute the embedding and positional writers W_E and W_{pos} and the unembedding reader W_U. Each neuron m contributes its read and write vectors $r_m, w_m \in \mathbb{R}^d$, used unit-normalized as $\hat{r}_m, \hat{w}_m$ throughout.

Assembling an edge. Every edge in the map is scored the same way. Pick one reader row of Table 6 and one writer row, check that the pair is causally eligible (the masks are listed in Table 8),

Table 6: **Every reader and writer of the map**, as instantiations of $G = R^\top R$ and $H = WW^\top$. Only the three interface-matrix Grams are built as arrays (the rotation-null sampling conjugates them). A dash marks a Gram never formed, collapsed through the inner-Gram identities of Appendix B.3; neuron normalizers are folded into the unit vectors up front.

Component	Matrix	Gram	Normalizer	Rank
Head, reader side (one row per channel)				
K-composition	$R = W_{QK,r}$	—	$\|W_{QK}\|_F^2 = \mathrm{tr}(g_Q g_K)$	$\le d_{\mathrm{head}}$
Q-composition	$R = W_{QK,r}^\top$	—	same as K-composition	$\le d_{\mathrm{head}}$
V-composition	$R = W_{OV,r}$	—	$\|W_{OV}\|_F^2 = \mathrm{tr}(g_O g_V)$	$\le d_{\mathrm{head}}$
Head, writer side (one row serves all three channels)				
head write	$W = W_{OV,w}$	—	$\|W_{OV}\|_F^2 = \mathrm{tr}(g_O g_V)$	$\le d_{\mathrm{head}}$
Neurons				
neuron read	$R = r_m^\top$	—	$\|r_m\|_2^2 = r_m^\top r_m$, folded into $\hat{r}_m$	1
neuron write	$W = w_m$	—	$\|w_m\|_2^2 = w_m^\top w_m$, folded into $\hat{w}_m$	1
Interfaces				
embedding write	$W = W_E$	$H = W_E W_E^\top$	$\|W_E\|_F^2 = \mathrm{tr}\, H$	$\le d$
positional write	$W = W_{\mathrm{pos}}$	$H = W_{\mathrm{pos}} W_{\mathrm{pos}}^\top$	$\|W_{\mathrm{pos}}\|_F^2 = \mathrm{tr}\, H$	$\le d$
unembedding read	$R = W_U^\top$	$G = W_U W_U^\top$	$\|W_U\|_F^2 = \mathrm{tr}\, G$	$\le d$

and evaluate $C^2 = \mathrm{tr}(GH)/(\mathrm{tr}\, G \cdot \mathrm{tr}\, H)$ with that reader's G and that writer's H, through the collapsed forms of Appendix B.3. The numerators of Table 7 are the gain-weighting of Proposition 1 in computational form. The K-composition numerator weights the cross factor by the *query* inner Gram g_Q, even though K-composition reads through the key factor, a deliberate reversal discussed in Appendix A.3. Two normalizers serve all four head rows, $\mathrm{tr}(g_Q g_K) = \|W_{QK}\|_F^2$ for the K- and Q-composition readers and $\mathrm{tr}(g_O g_V) = \|W_{OV}\|_F^2$ for both the V-composition reader and the writer.

B.3 Full-rank pairs, factored traces through the inner Grams

This subsection covers the computation of C when both the reader and the writer are full matrices, head $\to$ head in its three channels and the interface-matrix $\leftrightarrow$ head pairs. In the case of K-composition, the numerator $\|RW\|_F^2 = \|W_{QK,r} W_{OV,w}\|_F^2$ is needed for every causally eligible pair, 9,504 per channel on GPT-2. Computing it naively would require multiplying two $d \times d$ matrices per pair, $d^3 \approx 4.5 \times 10^8$ multiply-adds each at GPT-2's width and growing cubically with the stream, about 10^{13} operations for the head-to-head classes alone. The goal of this subsection is to evaluate the same number using only d_{head}-wide objects, where the largest per-pair cost is the $d\, d_{\mathrm{head}}^2$ multiplication that builds the cross factor, smaller by a factor of $(d/d_{\mathrm{head}})^2$, which is 144 on GPT-2 and 1,024 on Pythia-6.9B.

Substitute the factored forms $W_{QK,r} = W_{Q,r} W_{K,r}^\top$ and $W_{OV,w} = W_{O,w} W_{V,w}^\top$ into $\|M\|_F^2 = \mathrm{tr}(M^\top M)$:

$$\begin{aligned} \|W_{QK,r} W_{OV,w}\|_F^2 &= \mathrm{tr}\left(W_{OV,w}^\top W_{QK,r}^\top W_{QK,r} W_{OV,w}\right) \\ &= \mathrm{tr}\left(W_{V,w} W_{O,w}^\top W_{K,r} W_{Q,r}^\top W_{Q,r} W_{K,r}^\top W_{O,w} W_{V,w}^\top\right) \\ &= \mathrm{tr}\left(X^\top g_Q X\, g_V\right), \qquad X := W_{K,r}^\top W_{O,w}. \end{aligned} \tag{15}$$

The last step is one cyclic move and a regrouping. Cycling the leading $W_{V,w}$ to the end joins it with its transpose as the inner Gram $g_V := W_{V,w}^\top W_{V,w}$; the central pair $W_{Q,r}^\top W_{Q,r}$ is g_Q; and the two mixed products flanking g_Q are $X^\top = W_{O,w}^\top W_{K,r}$ and X. Every matrix in the final expression is $d_{\mathrm{head}} \times d_{\mathrm{head}}$, and the stream width d survives only inside the single product that builds X. The denominators collapse by the same move, $\|W_{QK,r}\|_F^2 = \mathrm{tr}(W_{K,r} W_{Q,r}^\top W_{Q,r} W_{K,r}^\top) = \mathrm{tr}(g_Q\, g_K)$ and $\|W_{OV,w}\|_F^2 = \mathrm{tr}(g_O\, g_V)$, so one elementwise division and a square root finish C for every pair. Table 7 instantiates the numerator for all three channels.

Table 7: **The factored numerator for each head-to-head channel.** The writer always contributes $W_{O,w}$ and $g_{V,w}$; the channels differ in which reader factor forms the cross factor X and which reader inner Gram weights it.

Channel	Reader R	Cross factor X	$\|R\,W_{OV,w}\|_F^2$
K-composition	$W_{QK,r}$	$W_{K,r}^\top W_{O,w}$	$\mathrm{tr}(X^\top g_{Q,r}\,X\,g_{V,w})$
Q-composition	$W_{QK,r}^\top$	$W_{Q,r}^\top W_{O,w}$	$\mathrm{tr}(X^\top g_{K,r}\,X\,g_{V,w})$
V-composition	$W_{OV,r}$	$W_{V,r}^\top W_{O,w}$	$\mathrm{tr}(X^\top g_{O,r}\,X\,g_{V,w})$

Every ingredient of the score is now a cached d_{head}-wide object, and the coefficient assembles as, for a K-composition edge,

$$C(r,w) \;=\; \sqrt{\frac{\mathrm{tr}\big(X^\top g_{Q,r}\,X\,g_{V,w}\big)}{\mathrm{tr}(g_{Q,r}\,g_{K,r})\;\mathrm{tr}(g_{O,w}\,g_{V,w})}}, \qquad X = W_{K,r}^\top W_{O,w}, \tag{16}$$

with the other channels substituting their row of Table 7. The four inner-Gram stacks are computed once per model and reused by every pair, and batched tensor operations evaluate the formula for all ordered pairs of the model's N_h heads at once. The released code is the reference for the exact batching.

Against the naive route's $\sim 10^{13}$ operations, evaluating all three GPT-2 channels this way costs about 2×10^{11} multiply-adds, and the gap widens as $(d/d_{\mathrm{head}})^2$ with width, the same comparison at Pythia-6.9B being $\sim 2 \times 10^{14}$ against $\sim 10^{17}$. The memory gap is as decisive as the arithmetic one, because the naive route's composed operators are d^2 floats each, 137 GB in total at Pythia-6.9B against the card's 32 GB, while the factored route's resident tensors, the factor and Gram stacks and one cross-factor block, total a few gigabytes at the same scale. Measured, the C tables of all three channels fill in 0.1 seconds for GPT-2 and 14 seconds for Pythia-6.9B on one GPU.

The interface-matrix $\leftrightarrow$ head classes evaluate the same master formula with the interface matrix's materialized Gram (Table 6) on one side. The head side's Gram is the sandwich $G_r = W_{K,r}\,g_{Q,r}\,W_{K,r}^\top$ (K-composition; the other channels analogously), so $\mathrm{tr}(G_r H) = \mathrm{tr}\left(g_{Q,r}\,(W_{K,r}^\top H\,W_{K,r})\right)$ and the head Gram is again never formed. These classes are always causally legal because the embedding and positional writers act before the first layer and the unembedding reader acts after the last.

B.4 Mixed pairs, quadratic forms through the skinny factors

This subsection covers the algorithm for computing C when exactly one side of the pair is a neuron, the head $\leftrightarrow$ neuron channels and the interface-matrix $\leftrightarrow$ neuron pairs. A neuron reads and writes through single vectors, so one side of every such pair is a single row or column whose Gram has rank one, and the score collapses from the dense machinery of the previous subsection to a small quadratic form. We derive head $\rightarrow$ neuron in full, because the remaining mixed classes follow the same pattern.

The edge asks how strongly the write operator of head w overlaps the input direction of neuron m, so the reader is the single row $R = \hat{r}_m^\top$ and the writer is $W = W_{OV,w}$. Their product $RW = \hat{r}_m^\top W_{OV,w}$ is a $[1 \times d][d \times d]$ multiplication, hence a single row of length d, and the Frobenius norm of a single row is that vector's Euclidean length, so $\|RW\|_F = \|W_{OV}^\top \hat{r}_m\|_2$. Evaluating this norm never requires the $d \times d$ operator itself. Substituting $W_{OV}^\top = W_V W_O^\top$ and multiplying from the right,

$$W_{OV}^\top \hat{r}_m = W_V\big(W_O^\top \hat{r}_m\big) = W_V\,x_m, \qquad x_m := W_O^\top \hat{r}_m \in \mathbb{R}^{d_{\mathrm{head}}}, \tag{17}$$

the computation passes through the d_{head}-vector x_m, which is the neuron's read direction expressed in the head's output channels. The squared numerator then collapses onto the cached inner Gram, $\|W_V x_m\|_2^2 = x_m^\top g_V\,x_m$. Both denominators are already precomputed as well, since $\|\hat{r}_m\|_2 = 1$ by construction and $\|W_{OV}\|_F^2 = \mathrm{tr}(g_O g_V)$, so the edge's coefficient assembles entirely from cached d_{head}-wide objects,

$$C(w,m) \;=\; \sqrt{\frac{x_m^\top g_{V,w}\,x_m}{\mathrm{tr}(g_{O,w}\,g_{V,w})}}, \qquad x_m = W_{O,w}^\top\,\hat{r}_m, \tag{18}$$

the mixed-pair analog of Eq. 16.

Batching. To efficiently compute the squared numerator $\|W_V x_m\|_2^2 = x_m^\top g_V\, x_m$ between one head and all $N_t = L\, d_{\mathrm{mlp}}$ neurons in a single pass (36,864 on GPT-2 small), stack every neuron's unit read vector $\hat{r}_m^\top = r_m^\top / \|r_m\|$ as the rows of $\hat{R} \in \mathbb{R}^{N_t \times d}$. One matrix multiplication $X = \hat{R} W_O$ then delivers all the projections at once, row m being $x_m^\top$, and batched tensor operations evaluate the N_t quadratic forms $x_m^\top g_V\, x_m$ from it directly, the released code being the reference. The denominators need no batching at all. The neuron side of every score is $\|\hat{R}\|_F = 1$ by construction, since the rows of $\hat{R}$ are unit vectors, and the head side is the single constant $\|W\|_F = \|W_{OV}\|_F = \sqrt{\mathrm{tr}(g_O g_V)}$ shared by all N_t scores of the row, so finishing C is one division of the row by one number. Per head this is about 2×10^9 multiply-adds, and nothing larger than $N_t \times d_{\mathrm{head}}$ is ever held.

The head $\leftrightarrow$ neuron classes are computed one head per batch, each pass scoring that head against all N_t neurons as one row of an $[N_h \times N_t]$ table, iterating through the heads until the table is full. The causal mask, applied to the finished table, allows equality here, $l_w \le l_r$, because within a block attention writes to the residual stream before the MLP reads it, so a head can feed the neurons of its own layer.

Neuron $\rightarrow$ head. These channels ask the reverse question, whether a neuron's write vector lands where a later head reads, and the derivation is the mirror image of head $\rightarrow$ neuron with the writer and reader roles swapped, so we state the result directly. For K-composition,

$$C(m, r) \;=\; \sqrt{\frac{y^\top g_{Q,r}\, y}{\mathrm{tr}(g_{Q,r}\, g_{K,r})}}, \qquad y = W_{K,r}^\top\, \hat{w}_m, \tag{19}$$

and the Q- and V-composition channels substitute $y = W_{Q,r}^\top \hat{w}_m$ against $g_{K,r}$ and $y = W_{V,r}^\top \hat{w}_m$ against $g_{O,r}$, with the normalizers of Table 6. Batching is identical to head $\rightarrow$ neuron with the roles reversed. The mask here is strict, $l_w < l_r$, because an MLP writes after everything else in its layer.

Interface matrices $\leftrightarrow$ neurons. These classes pair a neuron vector with an interface matrix's Gram, and no factoring is needed at all, because these are the three Grams the implementation genuinely materializes (Table 6), $H_{\mathrm{emb}} = W_E W_E^\top$, $H_{\mathrm{pos}} = W_{\mathrm{pos}} W_{\mathrm{pos}}^\top$, and $G_{\mathrm{unemb}} = W_U W_U^\top$, each computed once as a $d \times d$ array when the weights are loaded. The neuron side is a unit vector, so the coefficient is a normalized quadratic form in the materialized Gram,

$$C(H, m) \;=\; \sqrt{\frac{\hat{r}_m^\top H\, \hat{r}_m}{\mathrm{tr}\, H}}, \qquad C(m, G_U) \;=\; \sqrt{\frac{\hat{w}_m^\top G_U\, \hat{w}_m}{\mathrm{tr}\, G_U}}, \tag{20}$$

the left form for the embedding and positional writers $H \in \{H_{\mathrm{emb}}, H_{\mathrm{pos}}\}$ against the neuron's read vector, the right form for the neuron's write vector against the unembedding reader $G_U := G_{\mathrm{unemb}}$, evaluated for all N_t neurons at once by batched tensor operations. The same materialized arrays also supply their eigenvalues to the exact null distribution below.

Exact rotation null distributions. A rank-one reader also makes the rotation null distribution exact and inexpensive to sample. Under the null hypothesis the read direction is a uniformly random unit vector, and in the eigenbasis of H the statistic $v^\top Q H Q^\top v = \sum_i \lambda_i u_i^2$ is a weighted sum of the squared coordinates of a uniform unit vector u, with λ_i the eigenvalues of H, so draws from the null distribution sample u directly and never form a $d \times d$ rotation matrix.

B.5 Rank-one pairs, neuron $\rightarrow$ neuron streamed

This subsection covers the computation of C for the last and largest class, neuron $\rightarrow$ neuron, where both the reader and the writer are single vectors and no general machinery is needed at all. The product $RW = r^\top w$ is a 1×1 matrix, the Frobenius norm of a single row or column is the vector's length, and so the score is a plain cosine directly from the definition, $C = |r^\top w| / (\|r\|\, \|w\|) = |\cos\theta|$, with θ the angle between r and w. The master formula agrees, since the middle scalar factors out of the trace, $\mathrm{tr}(r\, r^\top w\, w^\top) = (r^\top w)^2$. What makes this class hard is not the formula but the count, 6.2×10^8 candidate pairs on GPT-2 small and 1.3×10^{11} at Pythia-6.9B, far too many to hold in memory at once. The denominators are folded in up front by unit-normalizing each vector

Table 8: **The 18 connection classes.** Reader–writer pairing, causal mask, and candidate-count formula, instantiated for GPT-2 small. The head$_{\{K,Q,V\}}$ rows aggregate three separately scored channel classes over the same pairs.

Connection class	Mask	Count formula
head $\to$ head$_{\{K,Q,V\}}$	$l_w < l_r$	$3 \times 12^2 \times 66$
embedding $\to$ head$_{\{K,Q,V\}}$	always	3×144
positional $\to$ head$_{\{K,Q,V\}}$	always	3×144
head $\to$ unembedding	always	144
head $\to$ neuron	$l_w \leq l_r$	$(12 \times 3072) \times 78$
neuron $\to$ head$_{\{K,Q,V\}}$	$l_w < l_r$	$3 \times (3072 \times 12) \times 66$
embedding $\to$ neuron	always	36,864
positional $\to$ neuron	always	36,864
neuron $\to$ unembedding	always	36,864
neuron $\to$ neuron	$l_w < l_r$	$3072^2 \times 66$

once, after which one matrix product of a layer's d_{mlp} unit write vectors against a later layer's d_{mlp} unit read vectors emits the full $d_{\mathrm{mlp}} \times d_{\mathrm{mlp}}$ tile of cosines for that layer pair. The computation runs one tile at a time through the $L(L-1)/2$ ordered layer pairs (66 tiles of 38 MB and 9.4×10^6 cosines each on GPT-2 small). A single pass suffices, because the class is censused rather than selected (Section 3). Each tile contributes its per-span histogram for Figure 3, its pairs beyond the wire threshold $|\cos| \geq 0.5$, and its maximum. Peak memory is one 38 MB tile, and the 6.2×10^8 scores never exist at once.

Summary. Our efficient implementation of the coupling coefficient C varies by the type of connection, as follows:

1. When both sides are full matrices, factored traces through the inner Grams score all pairs of the class at once (Appendix B.3).
2. When one side is a neuron, quadratic forms score one component against every neuron at once (Appendix B.4).
3. When both sides are neurons, the cosines stream through one tile at a time.

B.6 Census and run configuration

Table 8 defines the 18 connection classes with their causal masks and candidate-count formulas. Table 1 gives the candidate census for every mapped model. Selection applies only to the head-to-head classes (Application 1, Appendix C.1). Interface classes are read as effect-size rankings against their rotation null distributions, and the neuron classes are censused against the exact chance law (Section 3).

Weights are loaded with LayerNorm folded and writer matrices centered (Appendix B.1). All decompositions are float64. Natural text is a fixed public sample of pile-10k, 32 sequences of 256 tokens for the ablation experiments and of 128 tokens for the subspace deletions. Repeated-block sequences use 128-token uniform-random blocks. All seeds are fixed. Every experiment in this paper runs on one consumer GPU. The full GPT-2 map builds in 15 s, the cross-model ablation experiments run in minutes per model, and the full maps of Pythia-2.8B and 6.9B build in 4 and 11 minutes.

C Application 1: selection and the community ablation

C.1 The selection standard

The stratum-fitted null distribution of Eq. 4 is, in the terminology of Efron (2004), the *empirical null distribution*, Application 1's selection standard.

Strata. A *stratum* is the set of candidate edges that are calibrated together, and every candidate edge belongs to exactly one. Two coordinates define it. The first is the connection class, because different kinds of pairs have different chance behavior. A cosine between two rank-one

neuron vectors, a head circuit read against a neuron vector, and two head circuits are geometrically different statistics. The second is the layer separation between writer and reader, because the stream's geometry drifts with depth, so nearby components share more ambient structure than distant ones. "K-composition edges whose writer sits three layers below the reader" is one stratum, and "neuron→neuron pairs seven layers apart" is another. On GPT-2 each head–head channel contributes eleven strata, one per separation, and each stratum receives its own empirical null distribution, its own median and MADN, so an edge is always judged against chance for its own kind of pair at its own distance. Pooling the separations would instead judge every pair against a baseline dominated by the most numerous separations.

The empirical null distribution. The stratum's empirical spread is the robust estimate of Eq. 4, $s_{\text{emp}} = 1.4826 \cdot \text{MAD}_s$. For the neuron–neuron class, where Lemma 6 supplies the theoretical spread $s_{\text{theory}} = 1/\sqrt{d}$ in closed form, the empirical spread agrees within 3% (script released). For rank-one pairs the empirical and theoretical null distributions therefore coincide, and for subspace pairs they diverge, so selection standardizes against the empirical null distribution, fitted per stratum (edge class $\times$ layer span) by central matching (Efron, 2004), in which the central quantiles of the stratum's score distribution estimate the center and the standard deviation of the null distribution on the assumption that true channels are sparse in the bulk.

Why robust estimators. The robust estimators of Eq. 4 are a measured necessity rather than a convention. Refitting the same strata with the stratum mean and standard deviation lets the signal tail inflate the estimated standard deviation of the null distribution by up to $5\times$ and removes 61% of the head–head map's edges, the signal corrupting its own null distribution (script released). At scale, this selection standard keeps $115{,}164$ of the $1{,}523{,}712$ head–head candidates at 2.8B and $115{,}793$ at 6.9B.

C.2 The community ablation

Protocol. The readout is the *induction gain*, the drop in next-token loss between a block of random tokens and an immediately following copy of it, 12.40 nats for clean GPT-2.[4] Head sets are ablated by freezing their outputs at corpus means (mechanics below), each set against five random control sets matched in size and per-layer head counts, with Pile loss tracked as a check on nonspecific damage.

Non-additivity. The whole community destroys far more than its parts sum to (Table 10), and such non-additivity is the signature of redundancy, documented as backup heads, self-repair, and non-unique task circuits (Wang et al., 2023; McGrath et al., 2023; Gong et al., 2026; Chen et al., 2026). The decomposition places that redundant machinery, invisible to single-head identification, inside a boundary the map drew from weights alone.

Table 9: **The community ablation replicates across models.** Percent of the clean induction gain destroyed by mean-ablating each head set. "Holds" is how many of the model's top-five behavioral induction heads fall in the community. Ctrl median is over the whole-community row's five matched random sets. Communities come from each model's own selected head graph (protocol below). Pythia-6.9B is absent due to our hardware's RAM constraint.

Model	Community (holds)	Five heads	Comm. minus five	Whole comm.	Ctrl median
GPT-2	n=30 (5/5)	60.8%	72.9%	94.7%	2.9%
GPT-2-medium	n=47 (5/5)	4.7%	75.4%	93.4%	2.0%
GPT-2-large	n=184 (4/5)	3.6%	102.0%	108.5%	16.2%
GPT-Neo-125M	n=32 (5/5)	63.7%	100.5%	101.3%	86.8%
Pythia-160m	n=25 (3/5)	60.0%	92.1%	86.2%	27.4%
Pythia-2.8B	n=443 (4/5)	41.3%	96.9%	97.1%	1.0%

[4]Each prompt is a 128-token block of uniform-random tokens followed by a copy of itself. The gain is the mean loss on the first block minus the mean loss on the second, averaged over 32 prompts with independently drawn blocks. Random tokens cannot be predicted from the weights, but every token of the second block appeared 128 positions earlier in the same prompt, so the gain isolates in-context copying.

The nested decomposition. The GPT-2 decomposition behind the community ablation (Section 4) ablates five nested head sets, from the five induction heads alone, through the previous-token head L4H11 and the 24 remaining community heads, up to the full community (Table 10). If induction were carried solely by the five induction heads, removing them would abolish the capability and removing the rest of the community would have little effect. Table 10 rejects both halves of that null hypothesis.

Table 10: **Decomposing the induction-community knockout in GPT-2.** Fraction of the clean induction gain (12.40 nats) destroyed by mean-ablating each head set. Each ablated set is compared with its own five random control sets, matched to it in size and per-layer head counts. Pile ΔNLL is the natural-text loss increase in nats. All numbers are stable across corpus seeds (largest change 1.1 points).

Ablated set	n	Gain destroyed	Ctrl median	Ctrl max	Pile ΔNLL
five induction heads	5	60.8%	0.5%	1.4%	+0.09
previous-token head L4H11	1	1.9%	0.1%	0.6%	+0.03
remaining community heads	24	16.1%	3.2%	21.7%	+0.22
community minus induction heads	25	77.3%	3.7%	12.1%	+0.35
whole community	30	93.8%	2.5%	5.1%	+0.53

Ablation mechanics. Heads are removed by *mean-ablation* of the value summary, not by zeroing. For head h, let $z_h(t) \in \mathbb{R}^{d_{\text{head}}}$ denote its attention-weighted value vector at position t, the quantity the output factor W_O turns into the head's write (Appendix A.1). A clean forward pass over a reference corpus of N prompts of T positions defines the frozen summary $\bar{z}_h = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} z_h^{(n)}(t)$, and the ablation replaces $z_h(t) \to \bar{z}_h$ at every position, so the head still writes the constant $W_O\bar{z}_h$ into the stream but carries no input-dependent signal. The reference corpus is matched to the readout, the repeated-block corpus for the induction gain and the Pile sample for the natural-text loss, so the preserved average is the right one for the distribution being measured. Mean-ablation is the conservative choice, because a head's average output acts as a standing bias that downstream LayerNorms and readers are calibrated to, and zeroing removes that bias along with the signal, producing off-distribution damage that inflates apparent importance (Wang et al., 2023; Zhang & Nanda, 2024). The mean is taken over positions as well as prompts, so any positional structure in the head's output is removed rather than preserved, the stricter choice for a position-critical capability. Control sets are ablated identically.

Cross-model protocol. The decomposition replicates across six models spanning three families and both positional schemes (Radford et al., 2019; Biderman et al., 2023; Black et al., 2021; Table 9). Every row uses the unified pipeline. Communities are the Louvain partition (seed 0) of the model's own selected head graph, the top five induction heads are identified behaviorally by induction-offset attention on repeated random blocks, and three head sets are mean-ablated per model, the five behavioral heads, the community minus those heads, and the whole community, each against matched random controls as above. The GPT-2 row repeats the protocol of Table 10 and agrees with it within the documented corpus-seed variation. Pythia-6.9B is absent for a hardware reason. Assembling the runnable fp32 model requires the HuggingFace copy and the TransformerLens processing copy simultaneously, about 55 GB against our machine's 45 GB of memory. Map building avoids this by streaming weights, but ablation requires the assembled model, so the 6.9B map's ablation probe is Application 2's deletion (Section 5).

Cross-model observations. Two regularities replicate in every model. Ablating the five behaviorally identified induction heads never abolishes the capability, destroying between 3.6% and 63.7% of the gain, with the larger models the extreme cases of compensation (4.7% at GPT-2-medium, 3.6% at GPT-2-large). And the parts overlap heavily, since the two component sets sum to well over the whole community's effect in every model. Three model-specific observations accompany the table. At Pythia-2.8B the specificity margin is the widest, 97.1% against a 1.0% control median. At GPT-2-large the seed-0 community is a quarter of the model's heads ($n = 184$ of 720), so its controls are themselves elevated and its ablation overshoots (108.5%, at +4.8 nats of natural-text cost). In GPT-Neo the matched random controls reach 86.8%, consistent with Neo's fragile stream geometry, in which 43% of residual variance lies in a single dimension (cf. Timkey & van Schijndel, 2021), so Neo carries the caveat rather than the evidence.

C.3 ADDITIONAL RESULTS

Layer-separation profile. Table 11 gives the wired-pair counts by writer-to-reader layer separation behind the long-range finding (Section 4).

Table 11: Wired head pairs by writer-to-reader layer separation. Eligible pairs are all ordered pairs with the writer in a strictly earlier layer. A pair is wired if at least one of its three channels is selected under C (FDR $q = 0.05$).

layer separation	1	2	3	4	5	6	7	8	9	10	11
eligible pairs	1584	1440	1296	1152	1008	864	720	576	432	288	144
wired pairs	102	88	77	65	59	69	83	68	48	40	18
wired (%)	6.4	6.1	5.9	5.6	5.9	8.0	11.5	11.8	11.1	13.9	12.5

The recovered induction edges. The per-edge statistics behind the induction-wiring recovery, as (C, z, q) per induction head: L5H1 (0.102, 6.3, $9{\times}10^{-9}$), L5H5 (0.097, 5.8, $2{\times}10^{-7}$), L6H9 (0.103, 6.2, $2{\times}10^{-8}$), L7H2 (0.087, 5.6, $6{\times}10^{-7}$), L7H10 (0.090, 5.9, $1{\times}10^{-7}$).

Copying scores. The receivers are copiers. By the copying score (Elhage et al., 2021), $\mathrm{copy}(h) = \sum_k \mathrm{Re}(\lambda_k)/\sum_k|\lambda_k| \in [-1, 1]$ over the eigenvalues of $W_{OV,h}$, the top receivers score at least 0.994. The S-inhibition head L8H10 is among the strongest anti-copiers (-0.982).

D APPLICATION 2: THE SUBSPACE DELETION ACROSS MODELS AND SCALES

This appendix provides the detailed procedures for selecting the top global bands in the residual stream that are the most relevant to a model's induction capability. Most of the algorithm is identical for every model, with a minor difference depending on whether the model has learned positional embeddings or rotary embeddings (discussed in Appendix D.2).

D.1 SELECTING TOP GLOBAL BANDS

The construction uses the attention-head factors and the interface matrices, each head's $W_{Q,h}, W_{K,h}, W_{V,h}, W_{O,h} \in \mathbb{R}^{d\times d_{\mathrm{head}}}$ together with W_E, W_{pos} when present, and $W_U^\top$ as standalone factors, and no MLP neuron weight enters the estimate. Each factor reads the stream through $X^\top$ or writes it through X, and the construction asks which stream directions this population of readers and writers uses. Taking the factor as the reader, $R = X^\top$, and a unit direction v as the writer, Eq. 2 squares to the Gram fraction of Section 5, $C^2(X, v) = \|X^\top v\|^2/\|X\|_F^2 = v^\top G_X\, v/\,\mathrm{tr}(G_X)$, the fraction of X's squared weight lying along v. The roles also swap freely, since setting $R = v^\top$ and $W = X$ produces the transpose of the same vector, so either orientation gives the same C^2.

The bands are the orthonormal directions maximizing this coupling summed over every factor, $\sum_X C^2(X, v) = v^\top S v$ with the pooled coupling matrix of Eq. 5. Each unit-trace Gram has eigenvalues summing to 1, one vote per factor, so no factor dominates by size. The bands are the successive maximizers of $v^\top S v$ over unit directions orthogonal to their predecessors, the Rayleigh quotient problem, so they are the leading eigenvectors of S, ordered by eigenvalue, and $\lambda_i = \sum_X C^2(X, v_i)$ is the total amount of coupling band v_i captures across the entire transformer. Since each factor contributes exactly trace 1 to S, $\mathrm{tr}(S)$ equals the number of factors, so the eigenvalues λ_i partition the model's total coupling budget among the bands, and $\lambda_i/\,\mathrm{tr}(S)$ is the fraction of the total load carried by band v_i. The eigenvalues decay smoothly with no privileged cutoff (Figure 5), and we screen the ten leading bands $v_1, \ldots, v_{10}$ as candidates for the selection rule below.

The two most position-specific bands, ranked by the ratio of Eq. 6, are the deleted pair, and the couplings entering the ratio are those of Eq. 7, with the positional matrix Π defined for each positional style in Appendix D.2. These are the writer roles W_E and W_{pos} already play in the map's interface-matrix classes.

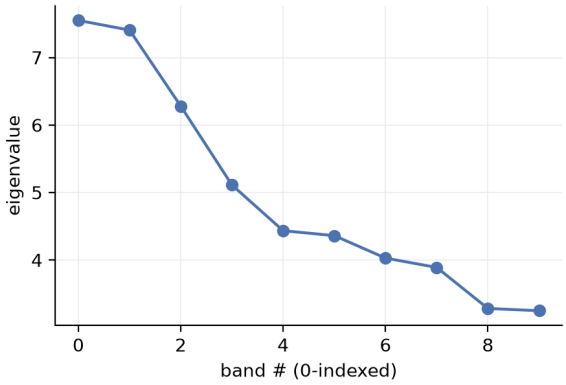

Figure 5: **Eigenvalues of the pooled coupling matrix** (GPT-2 small). Smooth decay with no sharp cutoff. The ten leading bands are the candidate pool for the selection rule of Section 5.

D.2 Positional matrix for learned positional embeddings (GPT-2) and rotary embeddings (Pythia)

GPT-2 has a learned positional embedding matrix W_{pos}, which directly writes position into the stream at layer 0 as a vector. In this case, the positional matrix is just the positional embedding matrix itself, $\Pi = W_{\text{pos}}$. Therefore,

$$C^2_{\text{pos}}(v_i) \;=\; \frac{\|W_{\text{pos}}^\top v_i\|^2}{\|W_{\text{pos}}\|_F^2}. \tag{21}$$

But in modern transformers that rely on rotary embeddings, as in the Pythia family, positions enter the stream only by rotating queries and keys inside attention, so there is no W_{pos} matrix to read. However, the stream still carries positional information. So we measure, or "reconstruct," the positional matrix Π empirically from the residual stream's activations instead, and call the estimated positional matrix $\hat{\Pi}$. Specifically, we feed the model 32 sequences of 128 tokens (a BOS token followed by 127 uniform-random tokens), so token content is pure noise and position is the only systematic variable, and cache the residual-stream state $h_t^{(s)} \in \mathbb{R}^d$ at the middle layer's input for every sequence s and position t. The resulting estimated positional matrix $\hat{\Pi} \in \mathbb{R}^{d\times T}$ has columns

$$\hat{\Pi}_t \;=\; \frac{1}{32}\sum_{s=1}^{32} h_t^{(s)} \;-\; \bar{h}, \qquad \bar{h} \;:=\; \frac{1}{32\,T}\sum_{s=1}^{32}\sum_{t=1}^{T} h_t^{(s)}, \tag{22}$$

with $T = 128$. Averaging over sequences cancels token content and keeps exactly what depends on position, and subtracting the mean $\bar{h}$ removes the position-independent component every position shares. With our estimated positional matrix $\hat{\Pi}$,

$$C^2_{\text{pos}}(v_i) \;=\; \frac{\|\hat{\Pi}^\top v_i\|^2}{\|\hat{\Pi}\|_F^2}. \tag{23}$$

The same construction runs unchanged at every rotary scale, from Pythia-160m through 2.8B to 6.9B, with only the stream width d and the middle-layer index adapting to the exact architecture.

D.3 Robustness

First, we test selecting the top two directions by the highest eigenvalues alone, without regard to whether the direction carries more positional or token information, on Pythia-160m. We find that deleting the two leading eigendirections destroys only 46% of the induction gain, against the 95.8% that the $\mathrm{PosRatio}$ selection achieves on the same model. This is consistent with the fact that induction is largely a positional phenomenon and confirms the importance of $\mathrm{PosRatio}$ as a selection criterion.

Second, we test the sensitivity of the rotary estimated positional matrix $\hat{\Pi}$ to its sample size, building it from 4 to 128 random-token sequences at three seeds per size on Pythia-160m. Every run selects

the identical pair of top induction-critical directions, with the selected bands' positional couplings never below $125\times$ chance and no unselected band above $28\times$. This shows that our estimation protocol for $\hat{\Pi}$ is reliable and not sensitive to sample size.

Third, GPT-Neo-125M is absent from Table 3 because random 2-dimensional deletions already destroy 60–98% of its induction gain. So no 2-dimensional deletion's effect is measurable, and the subspace deletion experiment is not meaningful for this specific model.

E GLOSSARY

Term	Definition
residual stream	the d-dimensional per-position vector every component reads from and adds to (Eq. 1)
QK / OV circuit	a head's pair-scoring form $W_{QK,h}$ / content-transforming operator $W_{OV,h}$
K/Q/V-composition	the three channels by which one head's writes enter another's keys / queries / values
potential connectivity	a wired channel: write/read subspaces aligned in the weights, regardless of use
edge; wire	a head pair selected by Application 1 (FDR $q = 0.05$); a neuron→neuron pair with $\lvert\cos\rvert > 0.5$
bands	leading eigendirections of the pooled coupling matrix, the directions most components read and write
copying score	normalized signed eigenvalue mass of $W_{OV,h}$ ($+1$ copy, -1 suppress)
theoretical null distribution	a pair's coupling distribution under Haar rotation of one side, singular values fixed (Prop. 5, Eq. 12)
empirical null distribution	the stratum's null distribution, fitted robustly from the observed scores (Eq. 4); Application 1's selection standard
previous-token score	a head's mean attention to the immediately preceding position
induction gain	first-copy minus second-copy loss on repeated random blocks (nats)
RW-PCA	reader–writer PCA, the eigendecomposition of the pooled coupling matrix S (Eq. 5), PCA applied to the model's read/write directions rather than to activations
positional coupling	$C^2_{\mathrm{pos}}(v) = \lVert\Pi^\top v\rVert^2/\lVert\Pi\rVert_F^2$, a direction's squared coupling with the positional matrix Π (Eq. 7), reported in multiples of the $1/d$ chance level; the token coupling C^2_{tok} is the analog with W_E
$\mathrm{PosRatio}$	a band's squared positional coupling divided by its squared token coupling (Eq. 6); Application 2 deletes the top two bands by this ratio