WHEN IS AN ATTENTION HEAD A COMPUTATIONAL UNIT? SHARED CIRCUITS DESPITE ORTHOGONAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic analyses of transformers often proceed head-by-head, looking for attention maps that correspond to computational roles. But when should an attention head be interpreted as a computational unit? In principle, a single computational function may be distributed across heads, and a single head may participate in several computations, with invariants appearing only after OV routing and residual-stream summation. In language models, where the underlying features and computations are unknown, an apparently uninterpretable head may be polysemantic, part of a distributed circuit, or a sign that the analyst chose the wrong decomposition. We therefore study transformers trained on factored sequence processes, a controlled setting where the target computation is analytically specified: multiple latent factors, represented in orthogonal residual-stream subspaces, receiving independent predictive updates. This lets us ask whether the attention circuits implementing these updates are themselves factorized. We find that they need not be. Per-factor updates constrain only the aggregate contribution across heads, leaving substantial freedom in how computation is distributed. Heads may specialize to individual factors, compose with other heads to implement one factor, or contribute polysemantically across factor boundaries. The regime that emerges depends on the generator's spectrum, the number of available heads, and training dynamics. We introduce *effective subspace attention*, a scalar quantity that combines attention patterns with OV circuits to measure the amount that heads collectively route from a source position into a given factor subspace. In our controlled setting, this aggregate view recovers the theoretically predicted computation even when per-head maps are illegible. Because only the aggregate is constrained, computation may be best understood at the level of task-relevant subspaces rather than individual heads. We then ask whether effective attention can support circuit discovery in real models and find that its natural application is closely related to existing circuit analysis techniques. When applied to GPT-2, it recovers known and novel circuits on two contrastive next-token tasks.

1 INTRODUCTION

When should an attention head be interpreted as a computational unit? Mechanistic analyses often proceed head-by-head, looking for attention maps that correspond to distinct roles. But multi-head attention need not organize computation this cleanly: a single computation may be distributed across heads (Jermyn et al., 2023), and a single head may contribute to multiple computations (Janiak et al., 2023). In such cases, the invariant object is not an individual attention map, but the aggregate residual-stream update produced after OV routing and summation.

In natural-language models we cannot easily tell these cases apart, because we lack ground truth for what the underlying features are and what computations the model is meant to implement. To study the question precisely, we need a setting where the functional units of the computation are known in advance and are controllable. We can then ask if and how those known functions distribute across heads, and why.

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.

We use transformers trained on *factored sequence processes*: sequences generated by several independently evolving computational units, called factors. Prior work shows that transformers trained on such processes represent each factor in its own orthogonal subspace of the residual stream (Shai et al., 2026). In this setting, we know the computational units (the factors), the subspaces they occupy (orthogonal subspaces of the residual stream), and the per-factor update that attention must implement at each layer, which we derive analytically from the data-generating process. We then ask: when the computation and corresponding representations of a network decompose into separable computational units, do the attention circuits naturally factor in the same way?

We find that they do not. Even when the factors are independent and assigned to orthogonal subspaces of the residual stream, individual heads can be illegible while the collective routed contribution matches theory exactly. A computation downstream from the attention block simply receives the full residual stream, and whether the information within was written by one or multiple heads is irrelevant.

Our formal results apply to the setting of single-layer transformers trained on factored processes, where ground truth is available; we then test whether these results remain useful in the more natural setting of a pre-trained multilayer language model. Our main contributions are:

1. We derive a **predictive theory for the collective update an attention layer must write** into each factor-specific subspace of the residual stream, along with the minimal QK/OV division of labor it requires. From this we predict the minimum number of heads needed to minimize loss on tested factored sequence processes.
2. We show that this subspace update admits **many head-level decompositions**. Heads can be specialized, collaborative, and polysemantic, all while achieving the same loss.
3. We introduce **effective subspace attention**, a quantity which decomposes the collective contribution of multiple attention heads into each subspace when individual attention maps are illegible.

2 Related Work

Interpretations of attention depend on what attention is meant to explain. Early work often treated attention weights as token-level explanations (Yang et al., 2016), but follow-up work showed that these weights often do not agree with other measures of feature importance which may admit functionally similar alternatives (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Mechanistic interpretability instead studies attention heads as components of larger circuits, since a head's effect depends on its attention pattern together with its value vectors, output projection, and downstream residual-stream computation (Elhage et al., 2021; Kobayashi et al., 2020). This perspective has identified "canonical" heads with relatively clean roles in induction, indirect-object identification, greater-than comparison, modular addition, copy suppression, and function vectors (Olsson et al., 2022; Wang et al., 2023; Hanna et al., 2023; Nanda et al., 2023; McDougall et al., 2024; Todd et al., 2024). At the same time, heads can be polysemantic, redundant, compensatory, or represented in superposition, with no clear roles (Janiak et al., 2023; Jermyn et al., 2023; Michel et al., 2019; Voita et al., 2019; Rushing & Nanda, 2024; McGrath et al., 2023; Wynroe & Sharkey, 2024; Kamath et al., 2025). Thus, heads can be useful units of analysis without being the fundamental units of computation.

Our work is closest to approaches that explicitly recover computation spanning multiple heads. Parameter decompositions can identify functionally meaningful subcomponents of attention (Christensen & Riggs, 2025), while Bushnaq et al. (2026) find weight-space components in language models that span heads and separate QK from OV contributions. Their results provide a natural-language analogue of the mechanisms studied here. Our setting lets us solve for the target computation analytically and ask when, and why, it aligns with individual heads rather than only with their routed aggregate.

3 Background

3.1 Data-generating processes and belief state geometry

We consider training data generated by edge-emitting hidden Markov models (HMMs). An HMM is defined by a token alphabet $\mathcal{X}$, a set of hidden states $\mathcal{S}$, an initial distribution $\boldsymbol{\eta}^{(\varnothing)}$ over hidden states,

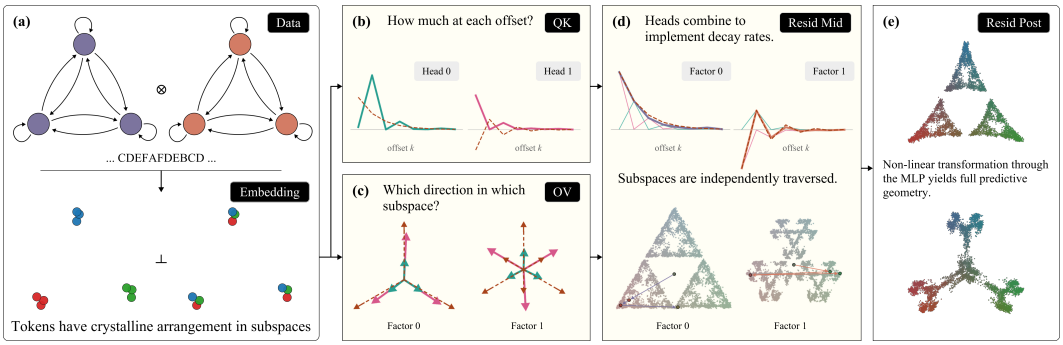

Figure 1: **Mechanistic overview of a transformer trained on a factored process.** **(a)** Tokens form factor-specific clusters in orthogonal residual-stream subspaces. **(b)** QK sets source weights by temporal lag, while **(c)** OV routes token-dependent updates into factor subspaces. **(d)** Heads combine to implement the factor-specific spectral decays predicted by theory. **(e)** The MLP maps the constrained predictive geometry to the full predictive geometry. We focus on the attention circuit (**b**–**d**).

and a collection of token-labeled transition matrices $(T^{(x)})_{x\in\mathcal{X}}$, where $T^{(x)}_{i,j} = \Pr(x, s_j \mid s_i)$ gives the joint probability of emitting token x and transitioning to state s_j given current state s_i.

An optimal Bayesian observer who knows the HMM structure but not the current hidden state maintains a *belief state* $\boldsymbol{\eta}$ over hidden states. Updates upon observing a token x and the belief state after observing history $x_{1:\ell}$ are:

$$\boldsymbol{\eta}' = \frac{\boldsymbol{\eta}T^{(x)}}{\boldsymbol{\eta}T^{(x)}\mathbf{1}} \qquad (1) \qquad\qquad \boldsymbol{\eta}^{(x_{1:\ell})} = \frac{\boldsymbol{\eta}^{(\emptyset)}T^{(x_1)}\cdots T^{(x_\ell)}}{\boldsymbol{\eta}^{(\emptyset)}T^{(x_1)}\cdots T^{(x_\ell)}\mathbf{1}} \qquad (2)$$

For stationary processes, the optimal initial belief state is the stationary distribution $\boldsymbol{\eta}^{(\varnothing)} = \boldsymbol{\pi}$, the left eigenvector of $T = \sum_x T^{(x)}$ associated with eigenvalue 1.

Prior work shows that transformers trained on next-token prediction linearly represent the belief states induced by the data-generator in the residual stream (Shai et al., 2024; Riechers et al., 2025).

Factored representations The Factored World Hypothesis (Shai et al., 2026) formalizes generators with independently evolving latent factors, and shows that transformers represent the per-factor beliefs in approximately orthogonal residual-stream subspaces. We call a process *independent* if each token-labeled transition operator factorizes as $T^{(x)} = \bigotimes_{n=1}^{N} T_n^{(x)}$, where factor n has hidden state space $\mathcal{S}_n$ and transition dynamics $T_n^{(x)}$. The observed token x is determined by the joint output of latent subtokens $z^{(n)}$, though this decomposition is not exposed in the training data. Under independence, product beliefs remain product beliefs under Bayesian updating: if $\boldsymbol{\eta}^{(x_{1:\ell})} = \bigotimes_n \boldsymbol{\eta}_n^{(x_{1:\ell})}$, then $\boldsymbol{\eta}^{(x_{1:\ell+1})} = \bigotimes_n \boldsymbol{\eta}_n^{(x_{1:\ell+1})}$. Thus, the joint belief can be represented losslessly as per-factor beliefs with factor n occupying its own $(d_n - 1)$-dimensional subspace, where $d_n = |\mathcal{S}_n|$.

3.2 An Architectural Division of Labor

We train pre-norm, decoder-only Transformer models on tokens generated by HMMs as described in Section 3.1, where each factor is represented by a three-state HMM, 'Mess3', (Marzen & Crutchfield, 2017) with a single nontrivial decaying eigenvalue.

We are primarily concerned with *attention*. Let x_s denote the normalized residual stream and σ the softmax function. For head h, the attention scores and residual stream writes are given by:

$$A^{(h)}_{d,s} = \sigma\left(\frac{(W_Q^{(h)}x_d)^\top (W_K^{(h)}x_s)}{\sqrt{d_{\text{head}}}}\right), \qquad \text{head}_h(d) = \sum_{s\le d} A^{(h)}_{d,s} W_O^{(h)} W_V^{(h)} x_s.$$

We refer to the former as the *QK circuit*, which determines where the head attends, and to $W_O^{(h)} W_V^{(h)}$ as the *OV circuit*, which determines what is written. The attention-layer output is the sum of these contributions across heads. We define $k = d - s$ as the source-destination lag and say attention is lag-dependent when the scores depend primarily on k, regardless of token identity, and content-dependent when they depend primarily on the source token's representation.

The controlled setting we work in has a structurally manifest decomposition in the data generator itself, and that decomposition maps cleanly onto distinct components of the attention architecture. This lets us state up front what the architecture *must* implement, and which parts of the architecture are capable of implementing it. We describe the intuition here.

For a factored generator, each factor transition matrix $T_n^{(x)}$ has its own independent state space, so optimal beliefs evolve independently by factor. Each update requires three pieces of information: (i) *which factor* to update, set by the generator's tensor-product structure; (ii) *how strongly* to weight each temporal offset, set by the factor spectrum; and (iii) *what token-dependent vector* to write into each factor's subspace. This maps directly onto attention as a minimal division of labor: OV chooses what to write into each factor subspace, and QK then need only choose the lag-dependent weights.

OV carries factor identity. (Figure 1c) For any attention pattern, head h writes only in the column space of $W_O^{(h)} W_V^{(h)}$: attention reweights value vectors but cannot change their span. Thus, if the residual stream decomposes into orthogonal factor subspaces $\bigoplus_n \mathbf{V}_{F_n}$ (Shai et al., 2026), the OV maps must route contributions into these subspaces. Factor identity is therefore enforced by OV, not by the attention weights.

QK need only be lag-dependent, not content-dependent. (Figure 1b) In our toy model, each source contribution decomposes into a token-dependent direction and a scalar ζ_n^{d-s} that depends only on the lag between source and destination tokens. Attention must be lag-dependent but not necessarily content-dependent. The minimal implementation assigns each dependency to the one circuit that must carry it—lag to QK, the token direction to OV.

This division constrains the *aggregate* layer-level circuit, not individual heads. The remainder of this section makes the required update explicit.

3.3 Constrained belief updates

Optimal prediction requires sequential Bayesian updating: each new token refines the current predictive vector, which itself reflects all previous tokens (Eq. (1)). However, attention is parallel — at every position, it computes a weighted sum of value vectors from all source positions simultaneously. This creates a tension: the model must approximate an inherently sequential computation with a parallel architecture.

The key insight of Piotrowski et al. (2025) is that attention resolves this by treating each source token as *independent evidence*. The BOS token at position 0 provides the prior $\boldsymbol{\pi}$ through the OV circuit — the baseline prediction before any evidence is observed. Each subsequent token x_s at offset $k = d - s$ then contributes an independent correction: "given the prior, how does observing x_s exactly k steps ago change my current predictions?" This correction is the geometric embedding of the difference between the prior $\Pr(S_d) = \boldsymbol{\pi}$ and the pointwise update from the observed token $\Pr(S_d | X_s = x_s) = \boldsymbol{\pi} T^{|x_s} T^{d-s}$. $T^{|x_s}$ captures the immediate evidence from the token and T^{d-s} propagates that evidence forward through $d - s$ steps of latent dynamics.

Summing the prior from BOS with independent corrections from all source tokens gives the *constrained belief update*:

$$\mathbf{r}_1^{(x_{1:d})} = \underbrace{\boldsymbol{\pi}}_{\text{from BOS}} + \sum_{s=1}^{d} \underbrace{\left(\boldsymbol{\pi}\, T^{|x_s}\, T^{d-s} - \boldsymbol{\pi} \right)}_{\text{correction from source } s}. \tag{3}$$

This is the best approximation to Bayesian updating achievable by summing independent pairwise contributions (see Appendix A for nuances) — precisely the functional form that attention can implement. If a transformer learns this updating, we should expect to find some linear map from the residual stream vectors induced by $x_{1:d}$ to the $\mathbb{R}^3$ vectors defined in equation 3.

The spectral decomposition of the transition matrix T reveals a clean separation of roles. For Mess3 whose T has a single nontrivial decaying eigenvalue ζ, the corrections can be re-written through spectral decomposition as

$$\Delta \mathbf{r}_1^{(x_{1:d})} = \sum_{s=1}^{d} \zeta^{d-s}\, \boldsymbol{\pi}\, T^{|x_s}\, P_\zeta \,, \tag{4}$$

where $P_\zeta T = T P_\zeta = \zeta P_\zeta$ (see Appendix K.1 for additional details).

Equation (4) realizes the division of labor above: lag appears only in the scalar ζ^{d-s}, while token identity appears only in the displacement $\boldsymbol{\pi} T^{|x_s} P_\zeta$. The current-token term ($s = d$, $k = 0$) is the only exception, since it can also enter through the residual stream via the token embedding; see App. D.

Piotrowski et al. (2025) confirmed this correspondence in transformers trained on the Mess3 process (Appendix K) and further showed that when the eigenvalue ζ is negative, the alternating pattern ζ^{d-s} cannot be implemented by a single attention head, whose weights must be non-negative.

3.4 Per-factor constrained belief updates

For a factored process with transition operators $T^{(x)} = \bigotimes_{n=1}^{N} T_n^{(x)}$, each Mess3 factor n has its own transition matrix $T_n = \sum_{z^{(n)}} T_n^{(z^{(n)})}$, stationary distribution $\boldsymbol{\pi}_n$, and eigenvalue ζ_n. Since the factors are represented in orthogonal subspaces, the constrained belief update (Eq. (3)) decomposes into independent per-factor updates:

$$\Delta \mathbf{r}_n^{(x_{1:d})} = \sum_{s=1}^{d} \zeta_n^{d-s} \cdot \mathbf{g}_n\left(z_s^{(n)}\right) \,, \tag{5}$$

where $z_s^{(n)}$ is the sub-token for factor n at position s and $\mathbf{g}_n(z) = \boldsymbol{\pi}_n\, T_n^{|z} P_{\zeta_n, n}$ is a token-dependent displacement direction that is independent of offset. For a given destination d, each correction term depends only on factor n's transition dynamics and the corresponding sub-token at source s (Figure 1d).

4 From aggregate computation to head-level realizations

The division of labor of Section 3.2 leaves open how a fixed number of available heads distribute the aggregate update: heads may specialize, collaborate on one factor, or contribute polysemantically across factors. Thus orthogonal factor representations need not imply specialized heads. We first derive a predictive theory for the minimum required head capacity, then test it across configurations before examining two qualitatively different head-level realizations. Training details are in Appendix N.1.

4.1 Minimum architecture

For a factored process, an attention head supplies a non-negative lag-dependent attention pattern, while its OV circuit determines how that pattern is signed and routed into factor subspaces. Thus the relevant architecture is not one head per eigenvalue or one head per factor, but the number of rays needed for nonnegative attention coefficients to express the required lag-dependent updates after OV routing.

Theorem 1 (Conic capacity of lag-dependent attention)**.** *Under the empirically observed solution class—lag-dependent attention with token-independent OV routing into each factor's update direction—the minimum head count admits a simple conic characterization. Let* $\mathbf{p}_k = (\zeta_1^k, \ldots, \zeta_N^k)^\top$ *be the generator-prescribed update at lag* k*. An* H*-head attention layer can implement the exact per-factor update iff there exist head-specific OV routing vectors* $\mathbf{b}_h$ *and nonnegative coefficients* $A_h(k)$ *satisfying*

$$\mathbf{p}_k = \sum_{h=1}^{H} A_h(k)\mathbf{b}_h, \qquad A_h(k) \geq 0.$$

Equivalently, the spectral trajectory lies in $\mathrm{cone}(\mathbf{b}_1, \ldots, \mathbf{b}_H)$*. Therefore, the minimum head count is the minimum number of rays whose cone contains this trajectory.*

Corollary 2 (Two Mess3 Factors)**.** *This gives one head for identical positive eigenvalues; two heads for distinct positive, same-magnitude mixed-sign, or identical negative eigenvalues. Distinct-magnitude mixed-sign eigenvalues require two heads rather than the per-mode count of three: the non-negative rays implementing the negative mode can be reused for the positive mode. The assumptions and proofs are provided in Appendix J.*

Theorem 1 predicts the aggregate factor-level update and the minimum capacity required to express it. It does not predict a unique head-level decomposition, nor which equivalent decomposition gradient descent will select.

The distinction between naively counting eigenmodes and the conic argument matters in mixed-sign settings. To represent a negative eigenvalue, at least 2 heads are required, since attention weights are non-negative. Analogously, two non-negative rays are needed to compose the alternating signed update. However, the same two non-negative rays can also be coupled with the same sign through OV to produce a monotone positive decay (Appendix L.3). Therefore, when a positive eigenvalue and a negative eigenvalue are present, the positive one can sometimes be implemented "for free" by reusing the rays already required for the negative mode, which reduces the head count below a naive per-eigenmode assignment. Table 1 lists the resulting predictions, which we test in the next section.

Table 1: Minimum head count under different two-eigenvalue cases. We leave out $-\zeta_1, -\zeta_2, |\zeta_1| \neq |\zeta_2|$ since the narrow range $[0, 0.5]$ makes the dynamics too similar (they can be approximated with fewer heads) or pushes one value too close to 0 (where attention does not matter).

Configuration	Values	Identifier	$H_{\min}$	$H_{\min}^{(\text{modes})}$
$+\zeta_1, +\zeta_2,\ \|\zeta_1\| \neq \|\zeta_2\|$	+0.7, +0.4	Distinct positive ($H = 2$)	2	2
$+\zeta_1, +\zeta_1$	+0.7, +0.7	Identical positive ($H = 1$)	1	1
$+\zeta_1, -\zeta_1$	+0.5, -0.5	Same magnitude, $\pm$ ($H = 2$)	2	2
$+\zeta_1, -\zeta_2,\ \|\zeta_1\| \neq \|\zeta_2\|$	+0.7, -0.5	Distinct magnitude, $\pm$ ($H = 2\ /\ 3$)	2	3
$-\zeta_1, -\zeta_1$	-0.5, -0.5	Identical negative ($H = 2$)	2	2

4.2 Effective subspace attention

To measure the computation that is invariant across these equivalent head-level decompositions, we define *effective subspace attention*. Let f_n map the residual-stream contribution onto the belief update for factor n in Eq. (5). Then

$$\boxed{\alpha_n(d,s) = \frac{\left\langle f_n\left(\sum_h A_{d,s}^{(h)} \mathbf{v}_s^{(h)}\right), \mathbf{g}_n(z_s^{(n)})\right\rangle}{\left\|\mathbf{g}_n(z_s^{(n)})\right\|^2}}. \tag{6}$$

This projects the summed routed contribution from source s to destination d onto the token-dependent displacement direction for factor n. Under the constrained belief-update theory,

$$\alpha_n(d,s) = \zeta_n^{d-s}.$$

By linearity, α_n decomposes across heads as $\alpha_n(d,s) = \sum_h \alpha_n^{(h)}(d,s)$, so the theory constrains only the sum:

$$\sum_h \alpha_n^{(h)}(d,s) = \zeta_n^{d-s}. \tag{7}$$

Specialization is the case where one head supplies this quantity for a factor. Collaboration is the case where several heads sum to it. Polysemanticity is the case where a head contributes to multiple factors. We quantify these regimes from per-head effective subspace attention. Polysemy measures spread across subspaces per head, while collaboration measures spread across heads per subspace; see Appendix C. Effective subspace attention lets us distinguish head-level structure from the invariant computation implemented by the full attention layer.

4.3 Invariant Computation Across Configurations

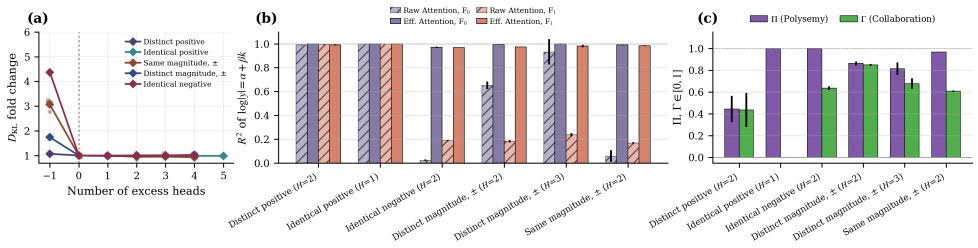

Figure 2: **(a)** D_{KL} versus the number of excess heads, , $n_{\mathrm{heads}} - H_{\mathrm{min}}$, evaluated on 10,240 sequences. The conic minimum head count aligns with the empirical loss plateau across all five configurations. **(b)** Raw attention maps are not reliable indicators of whether the correct factor-level computation has been learned: some heads match predicted decay patterns, but many do not. Effective subspace attention, which incorporates OV routing and sums across heads, consistently recovers the predicted decay rates. **(c)** Polysemy and collaboration are common whenever multiple heads are available: heads often contribute to multiple factors, and factors are often implemented by multiple heads. Thus the invariant computation is generally the routed aggregate update, not an individual attention head.

We first test these predictions across all five spectral configurations: for each configuration in Table 1, we train models with $n_{\mathrm{heads}} \in \{1, \ldots, 6\}$ across eight random seeds. Figure 2 shows that the conic head-count prediction matches the empirical loss plateau, while effective subspace attention recovers the prescribed factor dynamics even when raw attention does not. The same aggregate computation can nevertheless arise from qualitatively different head-level organizations, and collaboration and polysemy are common whenever multiple heads are available. We examine two concrete realizations next: one specialized and one collaborative.

4.4 A Specialized Realization: Distinct Positive Eigenvalues

We first examine a regime in which the head-level decomposition aligns cleanly with the factor decomposition: a two-head model with $\zeta_0 = 0.7$ and $\zeta_1 = 0.4$.

We identify the two factor subspaces $\mathbf{V}_{F_n} \subset \mathbb{R}^{d_{\mathrm{model}}}$ and measure the fraction of OV energy projected into each factor subspace (We call this "write-to energy", see Appendix G for full details). Figure 3 shows that OV learns to route most of its write-to energy into $\mathbf{V}_{F_0} \oplus \mathbf{V}_{F_1}$, with near-specialization by head; the recovered per-head residual stream geometry shows the same alignment (Figure 14). Specialization is not exact: Figure 4 shows that 11% of one factor's write-to energy comes from the other head. Early in training both heads reduce loss on the slower-decaying factor before one re-specializes to the faster-decaying factor, consistent with staged learning observed in prior work (Kunin et al., 2025; Saxe et al., 2014; Yüksel et al., 2026).

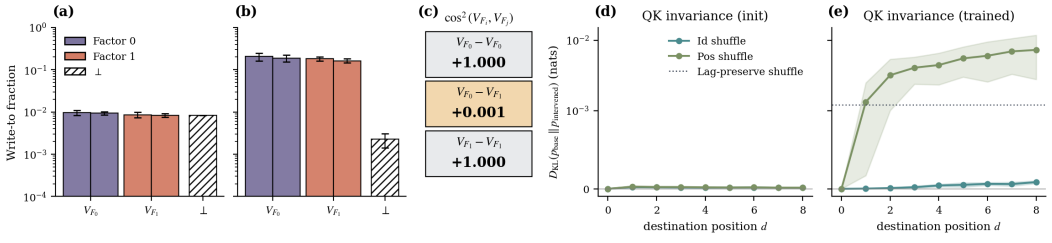

Figure 3: **OV routes factor-specific information into subspaces while QK is primarily lag-dependent rather than content-dependent. (a–c)** During training, OV concentrates its write-to energy in the two orthogonal factor subspaces, with near-specialization by head. **(d–e)** Causal QK interventions show that disrupting lag matters far more than token identity or absolute position, consistent with primarily lag-dependent attention.

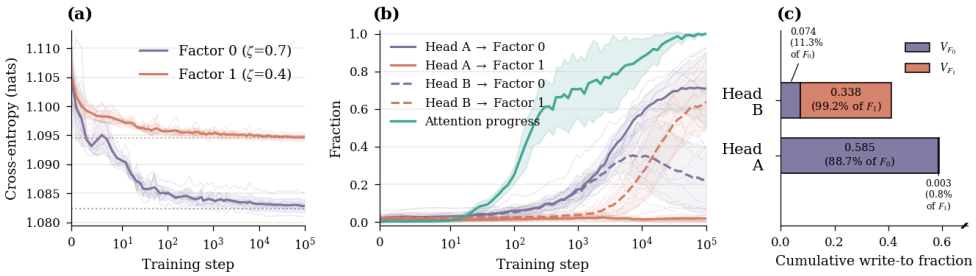

Figure 4: **Specialization emerges through a transient collaborative phase. (a)** Loss is reduced first on the slower-decaying factor. **(b)** Both heads initially contribute to that factor before one re-specializes to the other. **(c)** The final solution is largely specialized, with residual cross-factor leakage from the initial focusing. Results in **(a–b)** are averaged over 8 seeds.

QK shows the complementary structure predicted by the theory. The interventions in Figure 3 indicate that attention is primarily lag-dependent rather than content-dependent: disrupting source positions matters, while preserving lag largely removes the effect. Neither insensitivity is architecturally forced, since the same inputs carry both token identity and absolute position; the trained circuit instead tracks the dependencies of the constrained belief update. Raw per-head attention likewise matches the predicted factor-specific decay rates over the lags most strongly associated with loss reduction (Figure 2, 9); see Appendix D for the $k = 0$ degree of freedom and Appendix H for the lag-ablation criterion.

Thus, when a convenient head-level decomposition is available, raw attention heads can indeed be interpretable. Yet even here the decomposition is not fixed: training transiently passes through a collaborative solution before specialization emerges.

4.5 A DISTRIBUTED REALIZATION: POSITIVE/NEGATIVE EIGENVALUES

Specialization is not guaranteed by the aggregate constraint. We next consider a two-head model with $\zeta_0 = 0.5$ and $\zeta_1 = -0.5$, where the same factor-level computation is realized by a distributed circuit. A negative eigenvalue requires at least two heads because nonnegative attention weights cannot individually implement sign-alternating decay (Piotrowski et al., 2025).

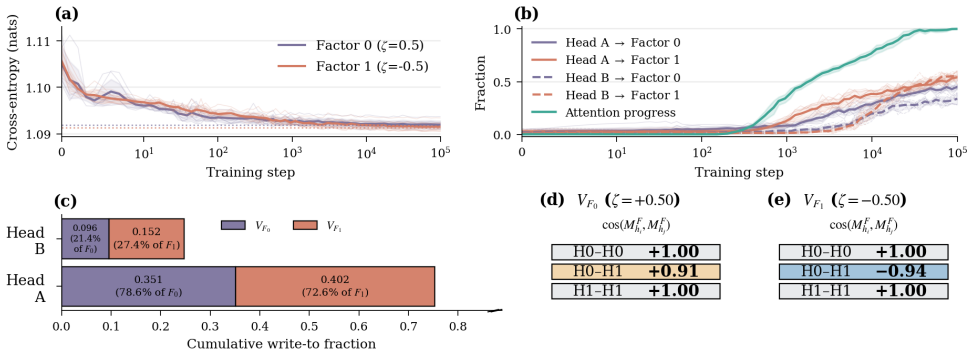

Figure 5: **Mixed-sign solutions are collaborative and polysemantic. (a–c)** Both heads contribute substantially to both factors. **(d–e)** Their OV writes align in the positive-eigenvalue subspace but are anti-parallel in the negative-eigenvalue subspace, enabling the alternating decay.

Figure 5 shows that both heads contribute substantially to both factor subspaces, so write-to energy does not reveal specialization. Instead, their signed OV directions differ: the heads write nearly in parallel in $\mathbf{V}_{F_0}$ but anti-parallel in $\mathbf{V}_{F_1}$, providing the coupling needed for the alternating update (Appendix G). Correspondingly, Figure 6 shows that neither raw per-head attention map matches

the factor dynamics, while effective subspace attention recovers both the predicted monotone and alternating decays. The recovered residual stream geometry likewise requires both heads (Figure 15).

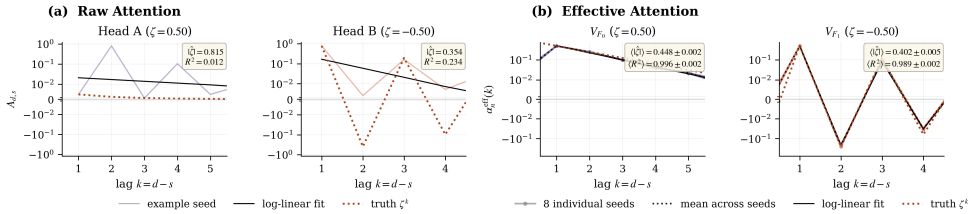

Figure 6: **Effective subspace attention recovers the mixed-sign dynamics when raw per-head attention does not.** The routed aggregate recovers both the monotone positive and alternating negative decays.

5 Application to Real Models

In real models the relevant subspace is generally unknown, but a task can provide a readout direction. For an answer token a and matched distractor b, let

$$\ell = W_U[:, a] - W_U[:, b], \qquad v_s^{(h)} = W_O^{(h)} W_V^{(h)} LN(r_s), \qquad \alpha^{(h)}(d, s) = \frac{A_{d,s}^{(h)} (v_s^{(h)})^\top \ell}{\gamma(d)}.$$

Here $\gamma(d)$ is the final-LayerNorm scalar at d, frozen from the same forward pass. We call this one-dimensional analogue "task-readout effective attention". It measures how much a head routes from source s in a task-relevant direction. This is closely related to circuit analyses combining attention patterns with OV or logit-space projections (Elhage et al., 2021; Wang et al., 2023).

On $N = 256$ IOI prompts (Wang et al., 2023), ranking all 144 heads by mean absolute task-readout effective attention recovers known circuit components: Name Movers 9.9, 9.6, 10.0; Negative Name Movers 10.7, 11.10; active Backup Name Movers 10.10, 10.6, 11.2, 10.1; and S-Inhibition heads 8.10, 7.9, 7.3.

We then apply the same procedure blind to correlative conjunctions (*neither→nor*, *either→or*) over $N = 48$ prompts. We identify seven consistently positive heads: 9.10, 10.5, 8.5, 6.11, 7.4, 4.1, and 11.8. Their contribution is routed from the trigger token, while their OV circuits point toward the partner even before it appears in context. Ablation shows the heads are causally necessary as a group despite individual redundancy. Note that this routine is a single forward pass: it can therefore suggest candidate circuits, but does not by itself establish causal necessity. See Appendix O for full details.

6 Conclusion

We asked when an attention head should be treated as a computational unit. In our setting, the answer is: the computational unit is at the level of the routed aggregate, not the individual head. In the observed solution class, the data-generating spectrum sets a minimum head count that trained models reliably reach across all five configurations we tested. Above that floor, the same predicted update admits many head-level decompositions that loss treats as equivalent and per-head inspection cannot distinguish, regardless of whether they are specialized, collaborative, or polysemantic.

The unit of computation, in our setting, is the residual-stream subspace into which a layer collectively writes rather than the heads that write into it. This opens a question for natural-data interpretability: how can we find such subspaces in scaled models, and thus uncover their natural units of computation?

Our analysis is limited to single-layer transformers, known-ground-truth processes, and lag-dependent computation. Extending to multi-layer architectures, richer spectra, and dynamics beyond full independence between factors remains future work. Our theory predicts exact head-count thresholds, but not how approximation quality varies as a function of the spectrum or head count. A natural next step is this quantification. Our real-model application is demonstrative: the real challenge is identifying the correct subspaces.

REPRODUCIBILITY STATEMENT

All the toy model parameters, model hyperparameters, architectures, and data-generator parameters are fully detailed in the appendix.

The theoretical work specifies its full set of assumptions and walks through every assumption's applicability to the models we train.

Additionally, we provide code to train all models and reproduce all figures as an uploaded archive. It is also available at: `https://anonymous.4open.science/r/headunits-F047/`.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and designing or providing feedback on research methodology or experiments. We have not used generative AI tools for generating synthetic datasets, proving mathematical claims, help developing theoretical models or conceptual frameworks, or proposing and refining hypothesis and assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analyses are not applicable to this work. Additionally, we used generative AI tools for figure generation, writing code, and drafting sections. We have reviewed all AI-assisted work. We checked the correctness of the experiments, data, and manually checked the relevant literature. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Lucius Bushnaq, Dan Braun, Oliver Clive-Griffin, Bart Bussmann, Nathan Hu, Michael Ivanitskiy, Linda Linsefors, and Lee Sharkey. Interpreting language model parameters. *Technical Report*, April 2026.

Casper L. Christensen and Logan Riggs. Decomposition of small transformer models. 2025. URL https://arxiv.org/abs/2511.08854.

Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. 2025. URL https://arxiv.org/abs/2409.14623.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

K. Fukuda and A. Prodon. Double description method revisited. In *Combinatorics and Computer Science*, volume 1120 of *Lecture Notes in Computer Science*, pp. 91–111. Springer, 1996.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, November 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc4de. URL http://dx.doi.org/10.1088/1742-5468/abc4de.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=p4PckNQR8k.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019. URL https://aclanthology.org/N19-1357/.

Jett Janiak, Chris Mathwin, and Stefan Heimersheim. Polysemantic attention head in a 4-layer transformer. *AI Alignment Forum*, 2023.

Adam Jermyn, Chris Olah, and Tom Henighan. Attention head superposition. *Transformer Circuits Thread, Circuits Updates – May 2023*, 2023.

Harish Kamath, Emmanuel Ameisen, Isaac Kauvar, Rodrigo Luger, Wes Gurnee, Adam Pearce, Sam Zimmerman, Joshua Batson, Thomas Conerly, Chris Olah, and Jack Lindsey. Tracing attention computation through feature interactions. *Transformer Circuits Thread*, 2025.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. 2020. URL https://arxiv.org/abs/2004.10102.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B Simon, Michael R DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=t7LKc0MMW6.

S. E. Marzen and J. P. Crutchfield. Nearly maximally predictive features and their dimensions. *Phys. Rev. E*, 95(5):051301(R), 2017. doi: 10.1103/PhysRevE.95.051301. SFI Working Paper 17-02-007; arxiv.org:1702.08565 [cond-mat.stat-mech].

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=5Hd6813x3U.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. 2023. URL https://arxiv.org/abs/2307.15771.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Mateusz Piotrowski, Paul M Riechers, Daniel Filan, and Adam S Shai. Constrained belief updates explain geometric structures in transformer representations. *arXiv preprint arXiv:2502.01954*, 2025.

Paul M Riechers, Thomas J Elliott, and Adam S Shai. Neural networks leverage nominally quantum and post-quantum representations. *arXiv:2507.07432*, 2025.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. 2024. URL https://arxiv.org/abs/2402.15390.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. 2014. URL https://arxiv.org/abs/1312.6120.

Adam Shai, Loren Amdahl-Culleton, Casper L. Christensen, Henry R. Bigelow, Fernando E. Rosas, Alexander B. Boyd, Eric A. Alt, Kyle J. Ray, and Paul M. Riechers. Transformers learn factored representations. *arXiv:2602.02385, ICML*, 2026.

Adam S Shai, Sarah E Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M Riechers. Transformers represent belief state geometry in their residual stream. *NeurIPS*, 2024.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019. URL https://aclanthology.org/D19-1002.

Keith Wynroe and Lee Sharkey. Decomposing the qk circuit with bilinear sparse dictionary learning. *AI Alignment Forum*, 2024.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016. URL https://aclanthology.org/N16-1174/.

Oğuz Kaan Yüksel, Rodrigo Alvarez Lucendo, and Nicolas Flammarion. Incremental learning of sparse attention patterns in transformers. 2026. URL https://arxiv.org/abs/2602.19143.

A Constrained Belief Update

We claim that the natural, single-token, parallel approximation of the full Bayesian Posterior $\Pr(S_d \mid Z_{1:d} = z_{1:d})$ is given by

$$\Pr(S_d) + \sum_{s=1}^{d} \left[\Pr(S_d \mid Z_s = z_s) - \Pr(S_d)\right]. \tag{8}$$

The natural geometric embedding of this expression, in the coordinates of a corresponding HMM with emission-labeled transition operators $T^{(z)}$, is the globally-normalized constrained belief update:

$$\mathbf{r}_{\text{global}}^{(z_{1:d})} = \boldsymbol{\pi} + \sum_{s=1}^{d} \left(\frac{\boldsymbol{\pi} T^{(z_s)} T^{d-s}}{\boldsymbol{\pi} T^{(z_s)} \mathbf{1}} - \boldsymbol{\pi} \right). \tag{9}$$

Piotrowski et al. (2025) states that Equation 9 can be rewritten as the row-normalized constrained update

$$\mathbf{r}_{\text{row}}^{(z_{1:d})} = \boldsymbol{\pi} + \sum_{s=1}^{d} \left(\boldsymbol{\pi} T^{|z_s} T^{d-s} - \boldsymbol{\pi} \right) \tag{10}$$

for processes like Mess3 (detailed in Appendix K), whose emission-labeled transition matrices have nonzero row sums. This statement is too strong. Nonzero row sums ensure that the row-normalized operator $T^{|z_s}$ in Equation 10 is well-defined but do not ensure equivalence between $T^{|z_s}$ and the globally normalized transition operator $\frac{T^{(z_s)}}{\boldsymbol{\pi} T^{(z_s)} \mathbf{1}}$ in Equation 9. To see the more general relationship between these operators, let us make their entry-wise definitions explicit. For any emitted symbol z, the emission-labeled transition matrices are given by

$$T_{ij}^{(z)} = \Pr(S_t = j, Z_t = z \mid S_{t-1} = i). \tag{11}$$

In contrast, the row-normalized operator is defined entry-wise as

$$T_{ij}^{|z} = \Pr(S_t = j \mid Z_t = z, S_{t-1} = i) = \frac{T_{ij}^{(z)}}{\sum_k T_{ik}^{(z)}}. \tag{12}$$

Equivalently, defining the emission-labeled row sums as

$$a_i^{(z)} = \sum_j T_{ij}^{(z)} = \Pr(Z_t = z | S_{t-1} = i) \tag{13}$$

and

$$D^{(z)} = \operatorname{diag}(a_1^{(z)}, \ldots, a_m^{(z)}), \tag{14}$$

we have

$$T^{|z} = \left(D^{(z)}\right)^{-1} T^{(z)}. \tag{15}$$

The corresponding stationary marginal probability of observing emission z is

$$c_{\boldsymbol{\pi}}^{(z)} = \boldsymbol{\pi} T^{(z)} \mathbf{1} = \Pr(Z_t = z \mid S_{t-1} \sim \boldsymbol{\pi}). \tag{16}$$

For the observed token z_s at source position s, when $a_i^{(z_s)} > 0$ for all i and $c_{\boldsymbol{\pi}}^{(z_s)} > 0$, we can define the *emission-likelihood re-weighting matrix* $W_{\boldsymbol{\pi}}^{(z_s)} = \frac{D^{(z_s)}}{c_{\boldsymbol{\pi}}^{(z_s)}}$ and show that the globally normalized transition operator is a token-dependent row-wise rescaling of the row-normalized operator:

$$\frac{T^{(z_s)}}{\boldsymbol{\pi} T^{(z_s)} \mathbf{1}} = W_{\boldsymbol{\pi}}^{(z_s)} T^{|z_s}. \tag{17}$$

From a Bayesian perspective, left-multiplication by $W_{\boldsymbol{\pi}}^{(z_s)}$ re-weights the stationary distribution over the pre-emission source state S_{s-1} according to the likelihood of observing emission z_s, producing the emission-conditioned posterior $\Pr_{\boldsymbol{\pi}}(S_{s-1} = i \mid Z_s = z_s)$. Thus, $W_{\boldsymbol{\pi}}^{(z_s)}$ encodes how knowledge of the emission should update our belief about the state it was emitted from. When emissions give no information about the pre-emission source state, then $W_{\boldsymbol{\pi}}^{(z_s)}$ is the identity matrix and $\mathbf{r}_{\text{global}}^{(z_{1:d})} = \mathbf{r}_{\text{row}}^{(z_{1:d})}$.

In the case of Mess3, permutation symmetries in its transition operators admit an affine relationship between the global and row-normalized forms of the constrained belief update:

$$\mathbf{r}_{\text{global}}^{(z_{1:\text{d}})} = (1 - \kappa)\boldsymbol{\pi} + \kappa \mathbf{r}_{\text{row}}^{(z_{1:\text{d}})} \tag{18}$$

where[1]

$$\kappa = \frac{q - \frac{1}{3}}{\tilde{q} - \frac{1}{3}}, \qquad \tilde{q} = \sum_i \boldsymbol{\pi}_i T_{ij}^{|0} = \frac{q}{3}\left[\frac{y}{qy + \beta(1-y)} + \frac{2p}{qp + \beta(1-p)}\right]. \tag{19}$$

This means that $\mathbf{r}_{\text{global}}^{(z_{1:\text{d}})}$ and $\mathbf{r}_{\text{row}}^{(z_{1:\text{d}})}$ make equivalent predictions for Mess3 when it comes to analyses based on linear representation. As such, the results of this work and the work of Piotrowski et al. (2025) are agnostic to the choice. In this paper, we choose to use $\mathbf{r}_{\text{row}}^{(z_{1:\text{d}})}$ to align notation with the form used in Piotrowski et al. (2025).

We want to emphasize that, in general, Equation 18 will not hold and the two updates will make linearly distinguishable predictions for candidate geometries. In principle, we could use this fact to design a data generator that would allow us to identify which version of the constrained update the transformer learns. We leave this for future work.

B Effective Attention

The constrained belief update for factor n at destination position d is:

$$\mathbf{r}_n^{(z_{1:d})} = \boldsymbol{\pi}_n + \sum_{s=1}^{d} \left(\boldsymbol{\pi}_n T_n^{|z_s^{(n)}} T_n^{d-s} - \boldsymbol{\pi}_n\right). \tag{20}$$

Each displacement can be written as:

$$\boldsymbol{\pi}_n T_n^{|z_s^{(n)}} T_n^{d-s} - \boldsymbol{\pi}_n = \zeta_n^{d-s} \cdot \mathbf{g}_n\left(z_s^{(n)}\right), \tag{21}$$

[1]The symbols (q, β, y, p) parameterize the emission-labeled transition matrices for Mess3 and are defined in Appendix K.1.

where $\mathbf{g}_n(z) = \left(\boldsymbol{\pi}_n T_n^{|z} - \boldsymbol{\pi}_n\right) P_{\zeta_n,n}$ is the token-dependent displacement direction (independent of offset), and ζ_n is the non-trivial eigenvalue of T_n. The eigenvalue decay ζ_n^{d-s} is the scalar we wish to recover.

The displacement map f_n maps the summed per-source OV contribution into factor n's constrained belief update. For a given (destination d, source s) pair:

$$f_n\left(\sum_h A_{d,s}^{(h)}\, \mathbf{v}_s^{(h)}\right) \approx \zeta_n^{d-s} \cdot \mathbf{g}_n\left(z_s^{(n)}\right) . \tag{22}$$

To isolate the decay, we project onto the known $k{=}0$ displacement direction $\mathbf{g}_n\left(z_s^{(n)}\right)$, which corresponds to offset zero where $\zeta_n^0 = 1$:

$$\alpha_n(d,s) \;=\; \frac{\left\langle f_n\left(\sum_h A_{d,s}^{(h)}\, \mathbf{v}_s^{(h)}\right),\; \mathbf{g}_n\left(z_s^{(n)}\right)\right\rangle}{\left\|\mathbf{g}_n\left(z_s^{(n)}\right)\right\|^2} . \tag{23}$$

This is a scalar-valued function of (d,s). Under the constrained belief update theory:

$$\alpha_n(d,s) = \zeta_n^{d-s} . \tag{24}$$

The denominator normalizes out the token-dependent magnitude, leaving only the offset-dependent decay. This is well-defined whenever $\mathbf{g}_n\left(z_s^{(n)}\right) \neq \mathbf{0}$, which holds for any token z that is informative about factor n.

By linearity of f_n, the effective attention decomposes additively across heads:

$$\alpha_n(d,s) = \sum_h \alpha_n^{(h)}(d,s) , \tag{25}$$

where

$$\alpha_n^{(h)}(d,s) = \frac{\left\langle f_n\left(A_{d,s}^{(h)}\, \mathbf{v}_s^{(h)}\right),\; \mathbf{g}_n\left(z_s^{(n)}\right)\right\rangle}{\left\|\mathbf{g}_n\left(z_s^{(n)}\right)\right\|^2} . \tag{26}$$

Under head specialization, $\alpha_n^{(h)}(d,s) \approx \zeta_n^{d-s}$ for exactly one head h and ≈ 0 for all others. In the compositional specialized case, multiple heads may contribute to the same factor while remaining factor-specific. Under polysemantic attention, no individual $\alpha_n^{(h)}$ matches ζ_n^{d-s}, but their sum does.

C Measuring polysemy and collaboration

The per-head effective attention (26) also defines a natural coupling matrix $C \in \mathbb{R}^{H\times N}$ that can be used to quantify how factor information is distributed across heads:

$$C_{h,n} \;\propto\; \mathbb{E}_{z,\,k}\left[\left|\bar{\alpha}_n^{(h)}(z^{(n)},k)\right|^2\right] . \tag{27}$$

where

$$\bar{\alpha}_n^{(h)}(z^{(n)},k) = \mathbb{E}_{(d,s):d-s=k,\; z_s^{(n)}=z}\left[\alpha_n^{(h)}(d,s)\right] \tag{28}$$

The average effective attention $\bar{\alpha}_n^{(h)}(z,k)$ describes the average attention head h pays to subtoken $z^{(n)}$ at a position k steps back from the query token. The average is over positions consistent with this lag, and sequences conditioned on a particular subtoken identity at this position. Since the effective attention can be negative, the coupling matrix measures the average attention power relevant to factor n within head h. From this matrix, we can define two probability distributions:

$$P(n \mid h) = \frac{C_{hn}}{\sum_{n'} C_{hn'}} \tag{29}$$

$$P(h \mid n) = \frac{C_{hn}}{\sum_{h'} C_{h'n}} \tag{30}$$

which characterize the amount a given head is responsible for a factor and the amount a given factor is handled by a particular head, respectively. We can also define relative weightings

$$w_n = \frac{\sum_h C_{hn}}{\sum_{n'h'} C_{hn'}} \tag{31}$$

$$w_h = \frac{\sum_n C_{hn}}{\sum_{n'h'} C_{hn'}} \tag{32}$$

that quantify the overall magnitude of a given head or factor's contribution to the coupling matrix. The entropy $\mathrm{H}[P]$ (not to be confused with number of heads H) of distribution P quantifies how peaked or spread a distribution is. Using the distributions in (29)- (30), we define

$$\Pi = \frac{1}{\log N} \sum_{h=1}^{H} w_h \mathrm{H}[P(n \mid h)] \quad \text{(polysemy)} \tag{33}$$

$$\Gamma = \frac{1}{\log H} \sum_{n=1}^{N} w_n \mathrm{H}[P(h \mid n)] \quad \text{(collaboration)}\,, \tag{34}$$

The polysemy score Π quantifies the extent to which a given head attends to multiple factors on average. It is zero when heads are responsible for one factor only and one when heads attend to all factors evenly. Similarly, the collaboration score Γ quantifies the fraction of total heads responsible for a particular factor on average. It is zero when each factor is handled by a single head and one when each factor is handled evenly by all heads. Finally, we define the effective number of heads for a particular factor

$$N_{\text{eff}}(n) = \exp(\mathrm{H}[P(h \mid n)]) \tag{35}$$

and its weighted average

$$\bar{N}_{\text{eff}} = \sum_n w_n N_{\text{eff}}(n) \tag{36}$$

C.1 Π, Γ, AND N_{EFF} AS A FUNCTION OF H

Figure C.1 plots the metrics Π, Γ, and N_{eff} as for a varying number of total heads.

D SKIP-CONNECTION SHARES DEGREES OF FREEDOM WITH DIAGONAL ATTENTION

The residual stream at position d after the attention layer is:

$$\mathbf{x}_{\text{mid}}^{(d)} = \underbrace{\mathbf{x}_{\text{pre}}^{(d)}}_{\text{skip connection}} + \sum_h \left[A_{d,0}^{(h)} \mathbf{v}_{\text{BOS}}^{(h)} + A_{d,d}^{(h)} \mathbf{v}_d^{(h)} + \sum_{s=1}^{d-1} A_{d,s}^{(h)} \mathbf{v}_s^{(h)} \right]. \tag{37}$$

Projecting into factor n's subspace, this must equal the constrained belief update (Eq. 4). The three terms on the right-hand side of Eq. 37 correspond to three groups of contributions:

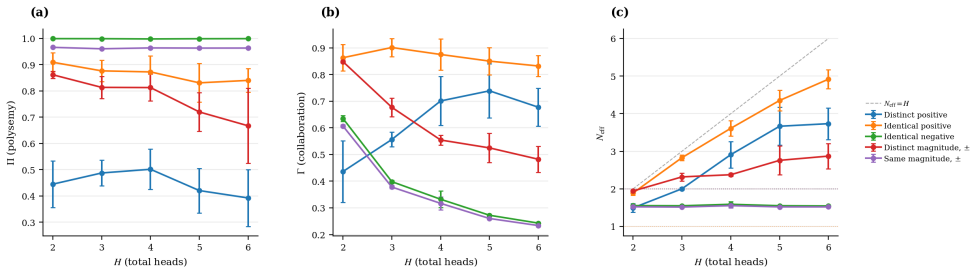

Figure 7: Polysemy Π, collaboration Γ, and effective number of heads as a function of total number of heads in the model. In each run, we fix d_{model} and vary n_{heads} which determines d_{head}. Error bars reflect the 95% confidence intervals over 8 random seeds.

Non-local sources ($1 \leq s < d$)**.** These contribute the offset $k \geq 1$ displacements and enter exclusively through attention. The spectral theory applies unambiguously:

$$\sum_h A^{(h)}_{d,s} f_n(\mathbf{v}^{(h)}_s) = \zeta_n^{d-s} \cdot \mathbf{g}_n(z^{(n)}_s)\,. \tag{38}$$

BOS token ($s = 0$)**.** The BOS token encodes a share of the prior $\boldsymbol{\pi}_n$ through attention to position 0:

$$\sum_h A^{(h)}_{d,0} f_n(\mathbf{v}^{(h)}_{\text{BOS}}) = \gamma_n \cdot \boldsymbol{\pi}_n \tag{39}$$

for some scalar γ_n that may vary with position d but not with token identity.

Current token ($s = d$) **and skip connection.** The offset $k = 0$ displacement $\mathbf{g}_n(z^{(n)}_d)$ and the remaining share of the prior $(1 - \gamma_n) \cdot \boldsymbol{\pi}_n$ must be jointly delivered by two pathways:

$$f_n(\mathbf{x}^{(d)}_{\text{pre}}) + \sum_h A^{(h)}_{d,d} f_n(\mathbf{v}^{(h)}_d) = (1 - \gamma_n) \cdot \boldsymbol{\pi}_n + \mathbf{g}_n(z^{(n)}_d)\,. \tag{40}$$

The left-hand side has two terms but only their sum is constrained. The model is free to distribute the k=0 displacement across these two pathways without affecting the total. This is a degree of freedom that does not exist at any other offset.

D.1 Experimental test: one-hot embeddings

If the token embedding carries part of $\mathbf{g}_n(z^{(n)}_d)$ through the skip connection, then removing this information pathway should force the model to compensate through diagonal attention. We test this by comparing two training conditions:

- **Learned embeddings** (standard): the embedding matrix W_E is trained jointly with all other parameters. The skip connection can carry arbitrary token-dependent information into resid_mid.
- **One-hot embeddings**: W_E is frozen as a one-hot encoding (identity matrix, zero-padded to d_{model}). The skip connection passes through a vector that identifies the token but contains no learned displacement structure. The k=0 displacement must be delivered entirely through diagonal attention.

All other parameters are trained normally in both conditions. We train 8 seeds per condition on the $\zeta_1 = 0.5$, $\zeta_2 = -0.5$ case (as we observe this setting empirically having little diagonal-attention), holding all hyperparameters fixed.

D.2 Results

Figure 8 shows the mean attention weights across lag for both conditions. With learned embeddings, $\bar{A}_{d,d}$ is lower across lag 0, but all other positions match or are higher, reflecting the degree of freedom:

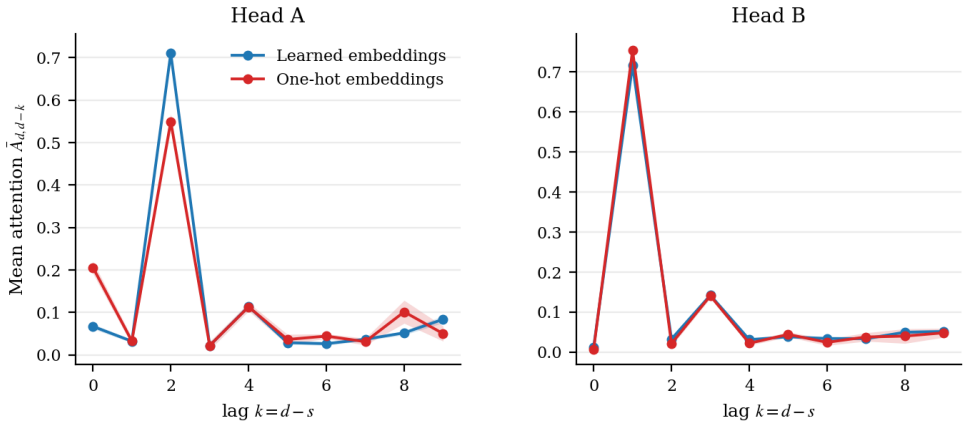

Figure 8: Mean attention-weights grouped by lag across seeds for learned vs. one-hot embeddings. Learned embeddings show lower diagonal attention across all seeds for head A, which we take to be the head implementing the $\zeta_1 = 0.5$ and the even offsets of $\zeta_2 = -0.5$. Notably, scores at all other lags are essentially an exact match. The only difference is in the diagonal attention. One-hot embeddings force higher diagonal attention because the model cannot use the embedding to store additional information. This isolates the $k = 0$ degree of freedom: the current-token displacement can be shared between the learned embedding/skip pathway and diagonal attention, while non-local attention patterns remain unchanged.

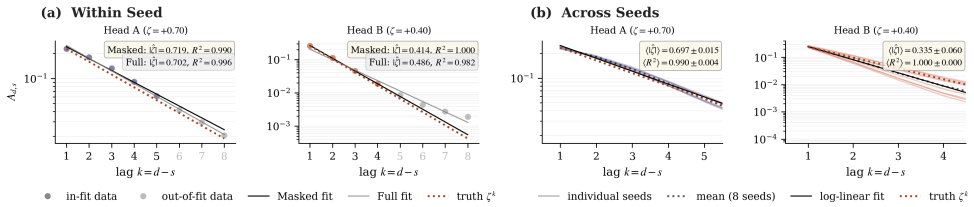

Figure 9: **Per-head attention patterns can be interpretable on their own.** In this case, for two factors with positive but different magnitude eigenvalues ($\zeta_0 = 0.70$ and $\zeta_1 = 0.40$), the decay rates of the attention profiles of each head uniquely match the temporal dynamics of one of the two factors – indicating we are in a regime of head *specialization*.

different seeds distribute the k=0 displacement differently between the embedding and diagonal attention. With one-hot embeddings, $\bar{A}_{d,d}$ is consistently higher.

These results support Eq. 40: the skip connection and diagonal attention are interchangeable pathways for the k=0 displacement, and the model may exploit this freedom when learned embeddings are available.

E 2 Positive Eigenvalues Are Interpretable At Most Seeds

We first examine a regime in which the head-level decomposition aligns cleanly with the factor decomposition: a two-head model with $\zeta_0 = 0.7$ and $\zeta_1 = 0.4$.

First, we use a behavioral method to find $\mathbf{V}_{F_n} \subset \mathbb{R}^{d_{\text{model}}}$, 2-d orthogonal subspaces that correspond to each factor. Once identified, we can define a write-to fraction that measures how much of the image of the OV circuit lands inside or outside of each subspace (see Appendix G for a detailed description of finding and evaluating these subspaces). Figure 3 shows that OV matrices initially write no more to the factor subspaces $\mathbf{V}_{F_n}$ than to other directions, but by the end of training place most of their write-to energy in $\mathbf{V}_{F_0} \oplus \mathbf{V}_{F_1}$, a 4-dimensional subspace of the 120-dimensional residual stream. Thus OV learns to route updates into the orthogonal factor subspaces. Separating

by head shows near-specialization. Recovering the intermediate activation geometry in Figure 14 also suggests strong alignment between a specific head and a factor's belief states. However, it is not exact specialization: in Figure 4, 11% of one factor's write-to energy comes from the head primarily associated with the other factor. This leakage is consistent with the training trajectory. Early in training, both heads reduce loss on the slower-decaying factor (largest eigenvalue), which has the larger available loss reduction; only later does one head re-specialize to the faster-decaying factor. This kind of staged learning is consistent with prior work on learning dynamics in neural networks and transformers (Kunin et al., 2025; Saxe et al., 2014; Yüksel et al., 2026). Thus, even in a setting where a specialized solution is available, gradient descent can transiently pass through collaborative decompositions.

QK shows the complementary structure predicted by the theory. The rightmost panel in Figure 3 shows token-shuffle interventions that indicate that attention is primarily lag-dependent rather than content-dependent: KL divergence increases when attention weights are shuffled across source positions, but not when shuffled only among tokens at the same position. When the position shuffles are constrained to preserve lag, the effect is significantly reduced. This indicates the QK circuits' sensitivity to the position shuffle is largely due to a sensitivity to lag and not the absolute position. Neither insensitivity is architecturally forced: both circuits take the same embeddings as inputs, which carry both token identity and absolute position. The trained circuit tracks the dependencies of the constrained belief update rather than the capacity of the architecture. Figure 9 shows attention weights binned by lag $k = d - s$. Since the $k = 0$ term shares a degree of freedom with the residual skip connection, we fit only off-diagonal attention weights; see Appendix D. The $\zeta_0 = 0.7$ head matches the predicted exponential decay, while the $\zeta_1 = 0.4$ head drifts at large lags, where the required corrections are small. Masking attention beyond lag τ confirms this: the slower-decaying factor remains sensitive up to $\tau = 5$, while the faster-decaying factor is sensitive only up to $\tau = 4$. We determine these thresholds using a noise floor. Lags whose ablation changes loss by less than 10^{-5} nats are deemed irrelevant. See Appendix H for details. Fitting only behaviorally relevant lags yields good agreement for both heads.

Thus, when a convenient head-level decomposition is available, raw attention heads can indeed be interpretable. Yet even here the decomposition is not fixed: training transiently passes through a collaborative solution before specialization emerges.

F A Positive and a Negative Eigenvalue are Uninterpretable

Specialization is not guaranteed by the aggregate constraint. We next consider a mixed-sign regime in which the same invariant factor-level computation is realized by a distributed head-level circuit.

We apply effective subspace attention to a two-head transformer trained on a factored process with eigenvalues $\zeta_0 = 0.5$ and $\zeta_1 = -0.5$. Piotrowski et al. (2025) showed that a negative eigenvalue requires two heads: because attention weights are nonnegative, a single head cannot implement the sign-alternating decay $(-0.5)^{d-s}$.

Let us begin by validating our predictions on the OV circuit. Just like the pattern shown in Figure 3, we find that the subspaces $\mathbf{V}_{F_0}$ and $\mathbf{V}_{F_1}$ are *nearly* orthogonal to each other and carry the lion's share of variation in the residual stream of the trained network. Unlike the previous case, however, we find no evidence of head specialization when looking at the write-to fraction. Figure 5 shows that instead, both heads contribute to both factors in approximate proportion to their total write-to magnitude. However, the *direction* in which they write differs. We establish this by looking at the cosine between the two heads' write operators into $\mathbf{V}_{F_n}$, computed in matrix-Frobenius geometry (details can be found in Appendix G).

The right panel in Figure 5 reveals that while the global write share doesn't distinguish factor 0 and 1, the signed value does. The two heads write to $\mathbf{V}_{F_0}$ in nearly the same direction, but they are almost exactly anti-parallel in how they write to $\mathbf{V}_{F_1}$, providing the mechanism to fit the alternating decay rate. The top panel of Figure 6 shows a naive attempt at fitting a decay rate for each head individually for a trained model. We can see each head implements an alternating attention pattern, with one head specializing to odd k and one specializing to even k. Fits to the expected exponential decays are very poor when we look only at a single head, but the effective subspace attention allows contributions from both heads, and recovers the predicted monotone decay for ζ_0 and the predicted alternating

decay for ζ_1. Recovering the geometry associated with each factor in Figure 15 also reveals that no single head is sufficient to represent the full geometry; both are required.

G Factor Subspace Identification and Quantification

We use the vary-one routine from Shai et al. (2026) to identify subspaces.

Because our data generator is a product of two independent processes, we can sample datasets that vary one factor's subsequence while keeping the other fixed. For each factor, we fix the other to a single realization, collect activations across many realisations of the varied factor, and mean-center them per position. The top 2 principal components then form an orthonormal basis for that factor's subspace: per-position centering cancels everything that is constant across the batch leaving only the variation induced by the varied factor, and since Mess3 beliefs are 2-dimensional, two components span the belief simplex.

We define these bases as $\mathbf{V}_{F_n} \subset \mathbb{R}^{d_{\text{model}}}$. We find that these subspaces account for the vast majority of the variance in the residual stream, explaining on average 97.2% of the variation, and the subspaces $\mathbf{V}_{F_1}$ and $\mathbf{V}_{F_2}$ are effectively orthogonal to each other, with a near-zero overlap. Once identified, we can define a write-to-fraction that measures how much of the image of the OV circuit lands inside or outside of each subspace.

Once equipped with the subspaces, $\mathbf{V}_{F_n} \subset \mathbb{R}^{d_{\text{model}}}$ we build the orthogonal-complement basis as the null space of the stacked factor bases $\mathbf{V}_{\perp} \in \mathbb{R}^{d\text{model} \times (d_{\text{model}}-4)}$ since $[\mathbf{V}_{F_0} \mid \mathbf{V}_{F_1}]$ is rank 4. Finally, we stack everything into a change-of-basis matrix:

$$\mathbf{B} = [\mathbf{V}F_0 \mid \mathbf{V}F_1 \mid \mathbf{V} \perp] \in \mathbb{R}^{d\text{model} \times d_{\text{model}}}$$

When $\mathbf{V}_{F_0}$ and $\mathbf{V}_{F_1}$ are exactly orthogonal, $\mathbf{B} \in \mathrm{O}(d_{\text{model}})$ is an orthogonal matrix and we can decompose any $x \in \mathbb{R}^{d_{\text{model}}}$ into coordinate vectors relative to those subspaces

$$y = \mathbf{B}^{\top} x = [y_0; y_1; y_{\perp}], \quad y_0 \in \mathbb{R}^2,\ y_1 \in \mathbb{R}^2, y_{\perp} \in \mathbb{R}^{d_{\text{model}}-4}\ . \tag{41}$$

In d_{model}-dim form, the orthogonal projectors are:

$$P_n = \mathbf{V}_{F_n}\mathbf{V}_{F_n}^{\top}, \quad P_{\perp} = \mathbf{V} \perp \mathbf{V} \perp^{\top} \approx I - P_0 - P_1$$

We note that the $\mathbf{V}_{F_n}$ subspaces are not completely orthogonal to each other, which is why $P_1 + P_2 + P_{\perp}$ is only approximately the identity. This allows the following approximate decomposition.

$$x \approx P_0 x + P_1 x + P_{\perp} x = \mathbf{V}_{F_0} y_0 + \mathbf{V}_{F_1} y_1 + \mathbf{V}_{\perp} y_{\perp}$$

With these definitions in hand, we can clearly measure the *global write share* which quantifies how much each head's OV circuit energy writes to the different interpretable subspaces of the residual stream.

$$E(h, j) \equiv \frac{\left|\mathrm{OV}_h \mathbf{b}_j\right|_F^2}{\sum_h \left|\mathrm{OV}_h\right|_F^2} \tag{42}$$

where $\mathrm{OV}_h \in \mathbb{R}^{d\text{model} \times d\text{model}}$ is head h's OV matrix and $\mathbf{b}j \in \mathbb{R}^{d\text{model}}$ is the j-th column of $\mathbf{B}$, lying in $\mathbf{V}_{F_0}$,$\mathbf{V}_{F_1}$, or $V_{\perp}$ depending on which block index j falls into. Summing per-direction within a block recovers the per head write share for subspace n: $E(h, n) = \sum_{j \in \mathbf{V}_{F_n}} E(h, j)$ which we interpret as the percent of head h's total write-to budget that lands in factor n's subspace.

Since we assume that only the computational subspace $\mathbf{V}_{F_0} \oplus \mathbf{V}_{F_0}$ matters, it seems like that normalizing just within this subspace would provide insight into the computational details. We call this quantity we call the *write-to fraction*, and it quantifies what percentage of the computationally relevant information coming into the residual stream came from which head and which factor:

$$\mathcal{E}(h,n) \equiv \frac{E(h,n)}{\sum_n \sum_h E(h,n)} \tag{43}$$

In panel b of figure 4, the vertical axis is a head-conditioned version of the write-to fraction $\mathcal{E}(h,n)/\sum_{n'} \mathcal{E}(h,n')$ that can be read as how each head splits its own computational write budget among the two relevant subspaces.

The treatment above assumes a uniform-input distribution, but this can be replaced with an empirical average over real data that also folds in attention. We define head h's contribution to the residual stream for batch b at position d

$$r_h(b,d) = \sum_{s \le d} A^{(h)}_{d,s}(b) v_{b,s} \mathrm{W}^{(h)}_O \in \mathbb{R}^{d\text{model}} \tag{44}$$

where $v_{b,s} \in \mathbb{R}^{d_{\text{head}}}$ are the post LayerNorm inputs from the residual stream projected through $W^{(h)}_V$. From this, we can define

$$E_{\text{realized}}(h,j) = \frac{\sum_{b,d} \left(\mathbf{b}_j^\top r_h(b,d)\right)^2}{\sum_{b,d} |r_h(b,d)|_2^2} \tag{45}$$

Realized versions of the differently normalized metrics follow from this definition. In practice, the difference between the OV only and the realized version is not qualitatively significant in any of the models we analyzed.

For a measure that is sensitive to direction, we use $\mathbf{M}^{(n)}_h = \mathrm{OV}_h \mathbf{V}_{F_n} \in \mathbb{R}^{d\text{model}} \times 2$ which is the OV map restricted to the rank-2 output subspace of $\mathbf{V}_{F_n}$. Its squared Frobenius norm is the unnormalized write-to-$\mathbf{V}_{F_n}$ energy: $|\mathbf{M}^{(n)}_h|_F^2 = |\mathrm{OV}_h \mathbf{V}_{F_n}|F^2$ (the numerator of $E(h,n)$). We then look at the cosine between two heads' write operators, computed in matrix-Frobenius geometry:

$$s(h,h',n) = \frac{\left\langle \mathbf{M}^{(n)}_h \mathbf{M}^{(n)}_{h'} \right\rangle_F}{|\mathbf{M}^{(n)}_h|F \cdot |\mathbf{M}^{(n)}_{h'}|_F}; \tag{46}$$

where $\langle \mathbf{A}, \mathbf{B} \rangle_F$ is the inner product of the matrices viewed as vectors in $\mathbb{R}^{d_{\text{model}} \cdot 2}$. This is the cosine displayed in figure 5.

H Additional Experiment Details

See Fig. 10 for details.

I Weight initialization and attention structure

The experiments in the main text use an initialization range of 0.02 (uniform $\mathcal{U}(-0.02, 0.02)$ for all weight matrices). Here we investigate how initialization scale affects both the final loss and the structure of the learned attention patterns.

I.1 Background: rich vs. lazy learning

The scale of weight initialization determines whether a neural network operates in the *rich* (feature learning) or *lazy* (kernel) regime Dominé et al. (2025) Geiger et al. (2020). At small initialization, the network starts with negligible structure and the optimizer builds representations from scratch — weights move far from their initial values and the learned features bear no resemblance to the random initialization. At large initialization, the initial random structure is amplified: gradients refine rather than replace the existing weight configuration, and the network stays closer to its initialization throughout training.

For attention, this distinction has a concrete consequence. At small initialization, the QK dot products are near zero, so softmax produces *near-uniform attention patterns*. The optimizer has an easier time

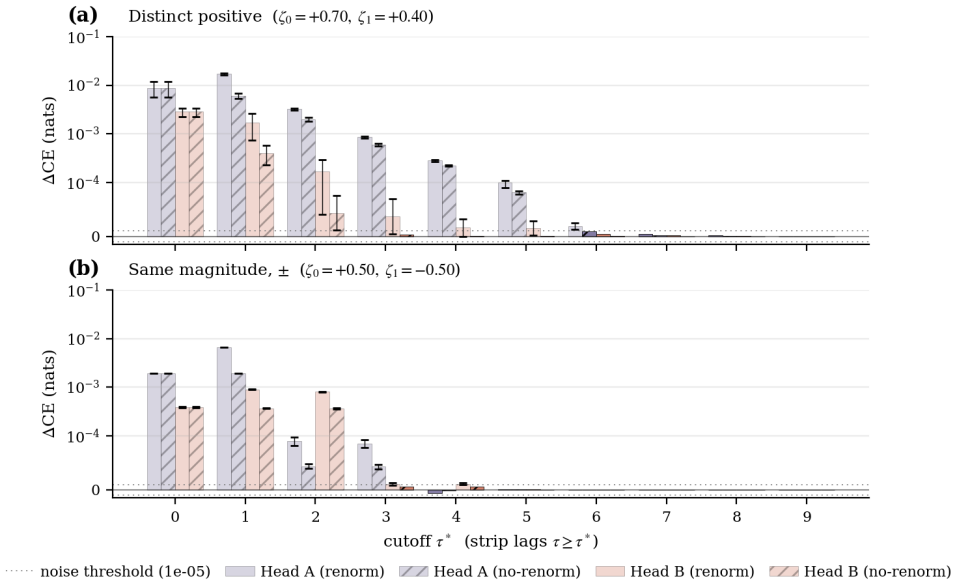

Figure 10: **The importance of attention weights to loss decay with lag.** We measure the cumulative impact on loss of ablating attention weights $A_{d,s}$ for all $\tau = d - s \geq \tau^*$. We then use this to determine the threshold τ for our attention decay fits. Bar heights are mean values over 8 model seeds.

changing these to match the spectral structure predicted by the constrained belief update framework. At large initialization, the QK dot products are large, so softmax produces *sharp, random attention patterns* from the start. The OV circuit adapts to make use of whatever temporal structure the random initialization provides, but the QK circuit — whose gradients through softmax are small when attention is already saturated — has limited ability to reshape these patterns.

I.2 Experimental setup

We compare two initialization scales across all five configurations from Table 1:

- **Small init** (init range $= 0.02$, $W_{ij} \sim \mathcal{N}\left(0, 0.02^2\right)$): the setting used throughout the main text.
- **Default init** (TransformerLens default, $W_{ij} \sim \mathcal{N}\left(0, \left(\frac{0.8}{\sqrt{d_{\text{model}}}}\right)^2\right)$): substantially larger initialization scale.

All other hyperparameters are held fixed. For each configuration and initialization scale, we train 8 models with different random seeds.

I.3 Results: loss

Table 2 compares the final KL divergence between the two initialization scales. Models trained with small initialization consistently achieve lower loss across all five configurations. The gap is most pronounced for configurations requiring compositional attention, where the QK circuit must learn structured non-monotonic patterns that the random initialization does not provide.

I.4 Results: raw and effective attention

The loss difference is explained by the quality of the learned temporal dynamics. Figure 11 shows our predictive framework applied to higher initialization.

Table 2: Final D_{KL} (nats, mean $\pm$ std across seeds) for small vs. default initialization. H: number of heads. The ratio column shows how many times worse the large initialization is. Configurations requiring compositional attention (negative eigenvalues) are most affected. Not only is loss worse on average, the variance is much more extreme.

Configuration	H	Small init (0.02)	Default init	Ratio
Identical positive	1	0.00039 ± 0.00001	0.00045 ± 0.00001	$1.14\times$
Distinct positive	2	0.00018 ± 0.00001	0.00034 ± 0.00004	$1.96\times$
Distinct mag., $\pm$ (2H)	2	0.00030 ± 0.00001	0.00040 ± 0.00001	$1.36\times$
Distinct mag., $\pm$ (3H)	3	0.00029 ± 0.00001	0.00049 ± 0.00008	$1.69\times$
Same magnitude, $\pm$	2	0.00028 ± 0.00001	0.00091 ± 0.00026	$3.29\times$
Identical negative	2	0.00016 ± 0.00000	0.00106 ± 0.00062	$6.78\times$

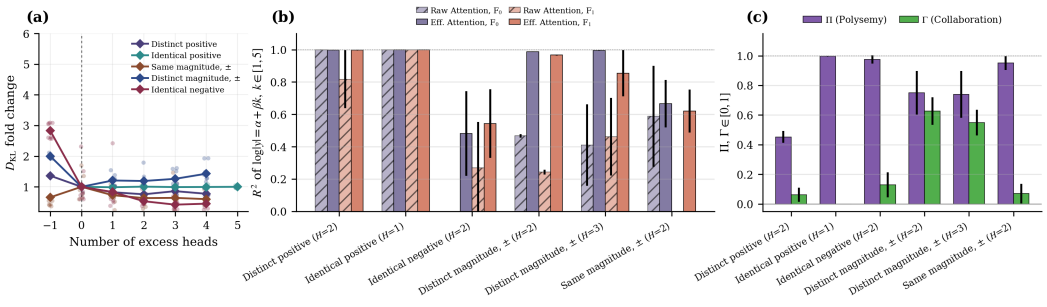

Figure 11: At larger initializations, the model fails to converge to optimal solutions. **(a)**: Compared to the main text, we here see a much more inconsistent pattern, where sometimes adding additional heads will even worsen the model, as it gets stuck in a worse setting during optimization. **(b)**: Effective-attention reveals that for some configurations, the models simply do not learn the correct patterns. We see that the cases with the worst matches are exactly the ones with the largest increase in loss from table 2. **(c)**: Our metrics show evidence of much more degenerate solutions. For the 2 worst cases, we have high polysemy yet low collaboration, indicating a dead head in the model and therefore effectively an underparameterized architecture.

With small initialization, attention patterns achieve high R^2 to the predicted candidate patterns across all configurations (consistent with the main text results). With default initialization, the R^2 is lower — the patterns show exponential-like decay but at rates that deviate from the eigenvalues of the data generator. This is consistent with the rich/lazy distinction: small initialization allows the QK circuit to learn the precise spectral structure from scratch, while large initialization locks in approximate patterns early in training that the optimizer cannot fully correct.

I.5 Interpretation

These results connect to the gauge freedom framework in two ways. First, they show that the degrees of freedom in how attention distributes computation are not merely theoretical — the initialization scale shifts which point in the solution space the optimizer finds, with small initialization favoring specialized solutions and large initialization favoring more entangled configurations where no head cleanly matches a single predicted pattern.

Second, the loss degradation at large initialization suggests that not all points in the solution space are equally accessible to gradient descent. The spectral theory predicts a family of equivalent solutions, but the QK parameterization's inductive bias — shaped by the softmax saturation at large initial weights — restricts which solutions the optimizer can reach. The gauge freedom is real in the loss landscape but constrained in practice by the training dynamics.

J CONIC VS. EIGENMODE DECOMPOSITION

The conic analysis (Appendix L) establishes that the distinct magnitude $\pm$ configuration ($\zeta_1 = +0.70$, $\zeta_2 = -0.50$) admits solutions at both $H = 2$ ($H_{\min}$, the conic minimum) and $H = 3$ ($H_{\min}(\text{modes})$, the eigenmode prediction). Both achieve equivalent loss (Figure 2a), but they implement the computation in qualitatively different ways.

J.1 TWO HEADS: CONIC DECOMPOSITION

At $H = 2$, the model cannot assign one head per eigenvalue. There are two eigenvalues but one is negative, requiring sign-alternating contributions that a single non-negative pattern cannot produce. Instead, the model finds a *conic* solution: both heads learn non-standard attention patterns that are individually difficult to interpret, but whose linear combination through the OV routing recovers the correct per-factor decay rates.

Figure 2b shows that for the 2-head case (Distinct magnitude, $\pm$ $(H = 2)$), the raw attention R^2 is low for both factors. No head cleanly implements either 0.7^k or the even/odd pattern for $(-0.5)^k$. However, the effective subspace attention R^2 is high: the OV routing successfully unmixes the heads' contributions to recover both ζ_1^{d-s} and ζ_2^{d-s}.

This is the conic decomposition in action. The two heads' patterns are non-negative elements in the feasibility cone (Definition 58), and the coupling matrix decomposes them into the target eigenvalue decays. This requires that no head is fully specialized; all heads must collaborate on all factors.

J.2 THREE HEADS: EIGENMODE DECOMPOSITION

At $H = 3$, the model has enough capacity to assign one head per spectral role: one head implementing the positive decay, one implementing the even-offset pattern and one implementing the odd-offset pattern.

Figure 2b confirms this: for the 3-head case (Distinct magnitude, $\pm$ $(H = 3)$), the raw attention R^2 for the positive 0.7 factor is much higher, despite loss being equivalent. The model finds it useful to allocate a head primarily to serving exponential decay. There is, however, also higher variance in this metric than for the other cases. The model effectively has a head more than it needs, meaning there is a larger space of equivalent solutions it can find. It has less pressure to implement the exact attention maps. The hard requirement that 2 heads must implement the negative eigenvalue remains, and as such we need effective attention to recover the contribution to the 2nd factor.

J.3 IMPLICATIONS

The crucial observation is that both solutions implement the same per-factor belief update. The residual stream at resid_mid contains the same constrained predictive geometry in both cases, and both achieve equivalent loss. The difference is purely in *legibility*: the 3-head eigenmode solution admits a slightly more directly interpretable solution, while the 2-head conic solution is completely opaque just from the attention maps.

These findings reveal that the opaqueness of raw attention maps can come from multiple sources: in some cases, the model may be taking advantage of degrees of freedom and in others, there may simply exist a minimal decomposition of the mechanism that admits no clearly interpretable map.

K THE STRUCTURE OF OUR TRAINING DATA: A PROCESS WITH TWO NON-MARKOVIAN FACTORS

Throughout this paper, we use non-Markovian training data sampled from an ergodic source composed of two independent factors to assess how the transformer would learn to represent and predict tokens from a simple "world" made of parts. Here each part is a particular three-state HMM from the parametrized Mess3 family.

K.1 A MESS3 FACTOR

The Mess3 process Marzen & Crutchfield (2017); Shai et al. (2024); Piotrowski et al. (2025) has three hidden states and three observable tokens $\mathcal{Z} = \{0, 1, 2\}$. Since Mess3 is a 3-state HMM, its local predictive vectors live in a 2-simplex. This 2-simplex corresponds to the set of probability distributions over the three hidden states.

The Mess3 process is defined by two parameters, which can be interpreted as a latent-transition probability $p \in (0, 1/2]$ and observation fidelity $q \in [0, 1]$, with dependent quantities $\beta = (1 - q)/2$ and $y = 1 - 2p$.[2]

The labeled transition matrices are:

$$T^{(0)} = \begin{bmatrix} qy & \beta p & \beta p \\ qp & \beta y & \beta p \\ qp & \beta p & \beta y \end{bmatrix} \tag{47}$$

$$T^{(1)} = \begin{bmatrix} \beta y & qp & \beta p \\ \beta p & qy & \beta p \\ \beta p & qp & \beta y \end{bmatrix} \tag{48}$$

$$T^{(2)} = \begin{bmatrix} \beta y & \beta p & qp \\ \beta p & \beta y & qp \\ \beta p & \beta p & qy \end{bmatrix} . \tag{49}$$

The net transition matrix

$$T = T^{(0)} + T^{(1)} + T^{(2)} = (q + 2\beta) \begin{bmatrix} 1 - 2p & p & p \\ p & 1 - 2p & p \\ p & p & 1 - 2p \end{bmatrix} \tag{50}$$

has a uniform stationary distribution $\boldsymbol{\pi} = \boldsymbol{\pi}T = \frac{1}{3}[1 \quad 1 \quad 1]$, and a stationary right eigenstate of $\mathbf{1} = T\mathbf{1} = [1 \quad 1 \quad 1]^\top$.

Besides the stationary eigenvalue of 1, the net transition matrix T has eigenvalue $\zeta = 1 - 3p$, which has (algebraic and geometric) multiplicity of 2 for $p \in (0, 1/2] \setminus \{1/3\}$. (Notice that Mess3 degenerates into a memoryless uniform three-symbol process at $p = 1/3$.)

The spectral projection operator associated with stationarity is $P_1 = \mathbf{1}\boldsymbol{\pi}$. The spectral projection operator associated with ζ is $P_\zeta = I - P_1 = I - \mathbf{1}\boldsymbol{\pi}$. Note that these satisfy the eigen-relation $P_\lambda T = TP_\lambda = \lambda P_\lambda$, are orthonormal $P_\lambda P_\xi = P_\lambda \delta_{\lambda,\xi}$, and form a decomposition of the identity $\sum_{\lambda \in \Lambda_T} P_\lambda = I$. Accordingly, they form not only a decomposition of the net latent transition operator $T = \sum_{\lambda \in \Lambda_T} \lambda P_\lambda$, but also an eigen-decomposition of powers of T, such that $T^n = \sum_{\lambda \in \Lambda_T} \lambda^n P_\lambda = \mathbf{1}\boldsymbol{\pi} + \zeta^n P_\zeta$. Note that T^n corresponds to the cumulative latent transition dynamic when marginalizing over observed symbols for n time steps.

K.2 THE JOINT DYNAMIC

The minimal HMM capable of generating the probabilistic structure of the training sequences is a 9-state HMM obtained via the tensor product of two constituent Mess3 transition matrices:

$$T^{(x)} = T_1^{(z^{(1)})} \otimes T_2^{(z^{(2)})} \qquad \text{for } x = z^{(1)} + 3z^{(2)} . \tag{51}$$

Practically, training sequences are generated by running two HMMs in parallel and, based on the outputs of the two constituent HMMs, producing a token $x \in \mathcal{X} = \{m\}_{m=0}^{8}$ at each timestep, where each token corresponds to a unique $(z^{(1)}, z^{(2)})$ pair.

L MINIMUM ARCHITECTURAL REQUIREMENT

For each factor n, the marginal transition operator T_n has a two-dimensional eigenspace $\mathbb{R}^2_{\text{belief}}$ on which T_n^k acts as the scalar ζ_n^k. The model embeds the constrained belief updates in $\mathbb{R}^2_{\text{belief}}$ in a

[2] In previous publications, p and q were denoted by x and α respectively. We use p and q here to avoid a serious internal conflict of notation.

two-dimensional subspace of the residual stream V_{F_n}:

$$\mathbf{g}_n(z) \;=\; E_n\big(\boldsymbol{\pi}_n\, T_n^{|z} - \boldsymbol{\pi}_n\big)\, P_{\zeta_n,\, n} \;\in\; V_{F_n}. \tag{52}$$

$\mathbf{g}_n(z) \in V_{F_n}$ is the residual-stream image of the belief-space displacement direction determined from the generator alone, under the model's learned embedding $E_n : \mathbb{R}^2_{\text{belief}} \to V_{F_n}$.

For each head h and source position s carrying observed token $z = z_s$ in factor n, the OV-projected contribution to V_{F_n} is

$$\mathbf{w}_n^{(h)}(z, s) \;=\; P_{V_{F_n}}\Big(\mathrm{OV}_h \cdot \mathrm{LN}(W_E[z] + W_{\mathrm{pos}}[s])\Big). \tag{53}$$

The model's per-head contribution to factor n at (d, s) is $A_{d,s}^{(h)}\, \mathbf{w}_n^{(h)}(z, s)$, and the constrained belief update target requires

$$\sum_h A_{d,s}^{(h)}\, \mathbf{w}_n^{(h)}(z, s) \;=\; \zeta_n^{d-s}\, \mathbf{g}_n(z) \tag{54}$$

at every (d, s, n, z). Note that the map from the observed tokens and the per-factor sub-tokens is deterministic – so the observed token z unambiguously defines $\mathbf{g}_n$. Equation (54) is the fundamental constraint; everything below is a decomposition of it.

Without loss of generality, split each head's contribution into a part along $\mathbf{g}_n(z)$ and a part orthogonal to it (both still in the two-dimensional V_{F_n}). The vectors $g_n(z)$ are found by using a linear probe from activations in V_{F_n} to the generator-defined belief displacements $(\boldsymbol{\pi}_n T_n^{|z} - \boldsymbol{\pi}_n) P_{\zeta_n, n}$, effectively absorbing the model's embedding freedom.

$$\mathbf{w}_n^{(h)}(z, s) \;=\; \tilde{\mathbf{g}}_n^{(h)}(z, s) \;+\; \boldsymbol{\epsilon}_n^{(h)}(z, s) \tag{55}$$

Equation (54) splits into two equations:

$$\text{(parallel)} \qquad \sum_h A_{d,s}^{(h)}\, \tilde{\mathbf{g}}_n^{(h)}(z, s) \;=\; \zeta_n^{d-s}\, \mathbf{g}_n(z), \tag{56}$$

$$\text{(perpendicular)} \qquad \sum_h A_{d,s}^{(h)}\, \boldsymbol{\epsilon}_n^{(h)}(z, s) \;=\; \mathbf{0}. \tag{57}$$

L.1 Architectural flexibility

The architecture admits more freedom than the spectral target alone requires:

1. Attention is non-negative, so it can rescale magnitudes but *cannot* flip signs. Wherever sign-alternation is needed in (54), it must come from the OV maps.
2. Attention need not be lag-only. $A_{d,s}^{(h)}$ can depend on the source token through the QK circuit, so token-dependence in (54) can in principle live in attention rather than in OV.
3. Per-head perpendicular components $\boldsymbol{\epsilon}_n^{(h)}(z, s)$ can be non-zero, provided they cancel under the attention-weighted sum.
4. The OV projection $\mathbf{w}_n^{(h)}(z, s)$ may depend non-trivially on the source position s, through positional embeddings and LayerNorm.
5. The parallel coefficient can be token-dependent. Writing $\tilde{\mathbf{g}}_n^{(h)}(z, s) = b_n^{(h)}(z, s)\, \mathbf{g}_n(z)$, the model is free to vary $b_n^{(h)}$ with z (and with s).

The minimum head count is gauge-dependent: different combinations of (1)–(5) admit different solutions. We therefore first specify the algorithmic regime our trained models actually implement, derive the head count under that regime.

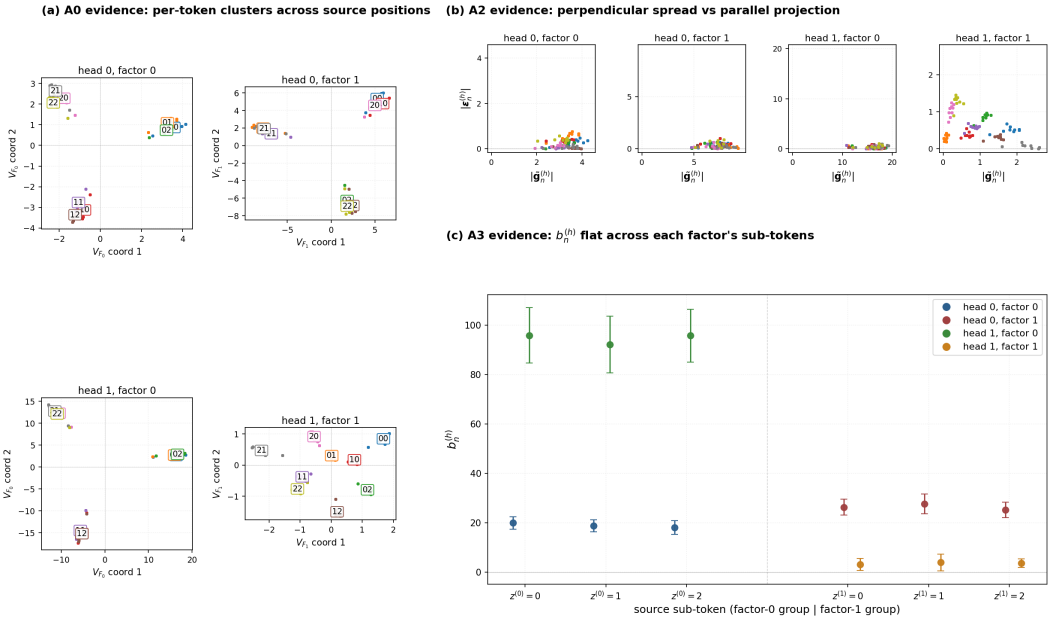

Figure 12: **(a)** the residual stream contribution from each head projected onto $\mathbf{V}_{F_n}$. The points cluster into groups according to the observed token without a clear signature of absolute position. Note that head 1 does not contribute to factor 1, and so its projections are simply noise. **(b)** Parallel components have significantly larger projections in the computational space than the perpendicular ones. **(c)** When we group our observed tokens by their factor-specific sub-token, our estimate of $|b_n^{(h)}|$ is roughly the same across different sub-tokens given a head and factor.

L.2 The observed-algorithm ansatz

Trained models satisfy (approximately) three properties across all configurations analyzed. Evidence for the first one is provided in the main text. Figure 12 provides detailed evidence of the others.

1. Lag-only attention. $A_{d,s}^{(h)} \approx A^{(h)}(k)$ where $k = d - s$. $\mathbf{w}_n^{(h)}(z, s) = \mathbf{w}_n^{(h)}(z)$, independent of s.
2. Negligible perpendicular component. $\|\boldsymbol{\epsilon}_n^{(h)}(z)\| \ll \|\tilde{\mathbf{g}}_n^{(h)}(z)\|$ per head.
3. Token-independent parallel coefficient. $\tilde{\mathbf{g}}_n^{(h)}(z) = b_n^{(h)}\, \mathbf{g}_n(z)$ with $b_n^{(h)} \in \mathbb{R}$ independent of z.

Under these assumptions, the constraints in (54) reduce to a z- and s-independent scalar equation:

$$\sum_h A^{(h)}(k)\, b_n^{(h)} \;=\; \zeta_n^k \tag{58}$$

Equation (58) is a conic problem.

L.3 The conic minimum head count

Each head contributes a fixed ray $\mathbf{b}^{(h)} = (b_1^{(h)}, b_2^{(h)}, ..., b_F^{(h)}) \in \mathbb{R}^F$ where F is the number of factors. Attention supplies non-negative lag-dependent scaling $A^{(h)}(k) \geq 0$. The target curve is $\mathbf{p}_k = (\zeta_1^k, \zeta_2^k, ..., \zeta_F^k) \in \mathbb{R}^F$. Equation (58) demands that each $\mathbf{p}_k$ lie in the non-negative conic hull of the rays:

$$\mathbf{p}_k \;\in\; \mathrm{cone}\big(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(H)}\big) \;\equiv\; \Big\{ \textstyle\sum_h \alpha^{(h)}\, \mathbf{b}^{(h)} : \alpha^{(h)} \geq 0 \Big\}. \tag{59}$$

The minimum head count for the two-factor case is therefore

$$H_{\min} \;=\; \min\Big\{\, H \;:\; \exists\, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(H)} \in \mathbb{R}^2 \text{ with } \{\mathbf{p}_k : k \geq 0\} \subseteq \mathrm{cone}\big(\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(H)}\big) \,\Big\}. \tag{60}$$

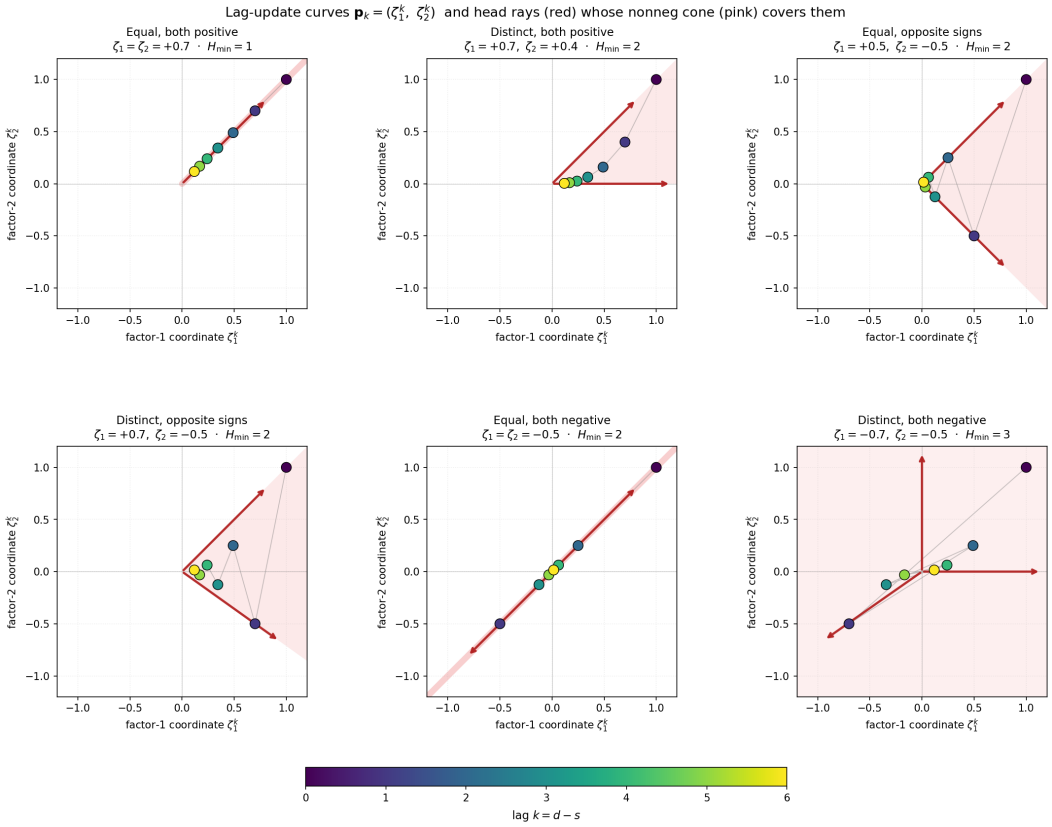

Figure 13: Geometric argument for the conic minimum for our 2-factor test cases.

We can solve this case with a simple geometric argument. The non-negative conic hull of H rays in $\mathbb{R}^2$ has angular extent strictly less than 180° when $H \leq 2$: two rays produce either a 2-D wedge of angle $\theta \in (0^\circ, 180^\circ)$, a 1-D line (if antiparallel), or a single ray (if parallel). Spanning an angular sector of 180° or more requires $H \geq 3$. To see several worked-out cases, see Figure 13. More generally, $N+1$ suitably chosen rays can positively span $\mathbb{R}^N$, giving the upper bound $H_{\min} \leq N+1$ for N factors.

Table 3: Minimum head count under different two-eigenvalue configurations.

Configuration	Lag-update geometry	$H_{\min}$
$+\zeta_1, +\zeta_2,\ \lvert\zeta_1\rvert \neq \lvert\zeta_2\rvert$	Curve in positive quadrant	2
$+\zeta_1, +\zeta_1$	Single ray	1
$+\zeta_1, -\zeta_1$	Two parity rays in half-plane	2
$+\zeta_1, -\zeta_2,\ \lvert\zeta_1\rvert \neq \lvert\zeta_2\rvert$	Two parity branches in half-plane	2
$-\zeta_1, -\zeta_1$	One line through origin	2
$-\zeta_1, -\zeta_2,\ \lvert\zeta_1\rvert \neq \lvert\zeta_2\rvert$	Curve in alternating quadrants	3

For a generic number of factors, we can find $H_{\min}$ using the following approach: Treat each lag point $(\zeta_1^k, \ldots, \zeta_N^k)$ as a candidate ray and reduce the resulting set to its minimum-size generating set via the double description method (Fukuda & Prodon, 1996). The reduced set has two parts:

- r one-way rays v that lie in the cone but whose opposite $-v$ does not.
- An ℓ-dimensional two-way subspace inside the cone. Every v and $-v$ inside of this subspace are in the cone.

If the cone is pointed ($\ell = 0$, no two-way piece), the one-way rays are all that's needed and $H_{\min} = r$. An additional $\ell + 1$ rays in general position are required to cover the ℓ-dimensional two-way subspace. Our intuition from $\mathbb{R}^2$ can help explain this: two antiparallel rays cover a line ($\ell = 1$ requires 2 rays); three rays cover the whole plane ($\ell = 2$ requires 3 rays).

L.4 Relaxing the assumptions

The architectural freedoms discussed above each correspond to dropping one of our assumptions about the solutions. The following introduces some, but not all, considerations introduced by relaxing our assumptions.

L.4.1 Token-dependent attention

Writing $A_{d,s}^{(h)} = A^{(h)}(k)\,\beta^{(h)}(z)$ with $\beta^{(h)} \geq 0$, the token-dependence in (54) can equivalently be absorbed into attention rather than OV:

$$\sum_h A^{(h)}(k)\left[\beta^{(h)}(z)\,\mathbf{w}_n^{(h)}(z)\right] \;=\; \zeta_n^k\,\mathbf{g}_n(z). \tag{61}$$

This is the same equation with the OV writes reweighted. Token-dependent attention and token-dependent OV are therefore gauge-equivalent: any solution in one gauge can be transported to the other without affecting the minimum head count. The token-shuffle interventions in the main text show that trained models tend towards the gauge where attention is primarily lag-only and the z-dependence sits largely in OV. Here, we see that this did not have to be the case.

L.4.2 Token-dependent parallel coefficient

If $b_n^{(h)}(z)$ depends on z, each head contributes a different ray per source sub-token. The conic argument then operates on $(H \cdot Z)$ rays in $\mathbb{R}^2$ rather than H, and the head-count grows with vocabulary. For Mess3, the three displacement vectors $\{\mathbf{g}_n(z)\}$ have equal magnitude, which makes a z-independent $b_n^{(h)}$ natural. For arbitrary generators where $\|\mathbf{g}_n(z)\|$ varies with z, the OV map would have to encode the magnitude variation itself or rely on the gauge freedom between token-dependent attention weights and OV magnitudes as described above.

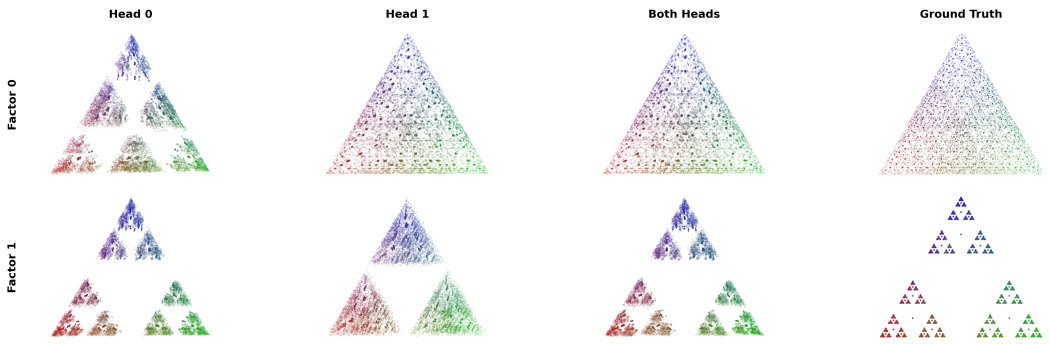

Figure 14: For $\zeta_1 = 0.7$, $\zeta_2 = 0.4$, regressing from the individual heads + the sum of both heads to the constrained belief simplex shows strong alignment between each head and a specific geometry.

L.4.3 Position-dependent OV projection

If the OV's projection into V_{F_n} depends non-trivially on the source position s, the constraint (54) no longer reduces to a z-only equation. The position-dependent component must either (i) be absorbed by position-dependent attention, or (ii) contribute residual error that the MLP must clean up.

L.4.4 Non-negligible perpendicular component

If $\boldsymbol{\epsilon}_n^{(h)}(z)$ is large in magnitude when compared to $\tilde{\mathbf{g}}_n^{(h)}(z)$, the cancellation condition (57) must hold at every (k, z). This forces an additional strong coupling between the heads' OV maps. The conic count (60) is then a lower bound. Empirically, trained models keep $\|\boldsymbol{\epsilon}\|$ small per head rather than relying on cross-head cancellation. This is a real freedom the model has but chooses not to use.

M Belief State Geometry

To visualize how individual heads contribute to the constrained belief geometry, we regress from partial attention-layer outputs to the ground-truth factored constrained belief state. Let x_d^{pre} denote the residual stream before the attention layer at destination position d, and let $o_d^{(h)}$ denote the OV-routed output of head h at the same position. For a two-head model, we construct three activation sets:

$$x_d^{(1)} = x_d^{\text{pre}} + o_d^{(1)}, \qquad x_d^{(2)} = x_d^{\text{pre}} + o_d^{(2)}, \qquad x_d^{(1,2)} = x_d^{\text{pre}} + o_d^{(1)} + o_d^{(2)}.$$

These correspond to adding only head 1, only head 2, or both heads to the pre-attention residual stream.

For each activation set $S \in \{1, 2, (1, 2)\}$, we fit a linear map

$$\widehat{r}_d^S = W_S x_d^S + b_S$$

to predict the factored constrained belief state. In the two-factor Mess3 setting, each factor has a three-dimensional constrained belief vector, so the regression target is

$$r_d^{\text{fact}} = [r_d^{(0)}; r_d^{(1)}] \in \mathbb{R}^6.$$

For visualization, we plot the recovered belief for each factor separately. Each three-dimensional factor belief is mapped to RGB coordinates, so points with similar inferred beliefs have similar colors. Comparing the regressions from $x_d^{(1)}$, $x_d^{(2)}$, and $x_d^{(1,2)}$ lets us see whether the constrained belief geometry is linearly recoverable from individual head contributions or only from their sum.

For the 2 cases we walked through in the main body, $\zeta_1 = 0.7$, $\zeta_2 = 0.4$ and $\zeta_1 = 0.5$, $\zeta_2 = -0.5$, we show here the per-head and combined regressions to the belief state geometry in figures 14 and 15.

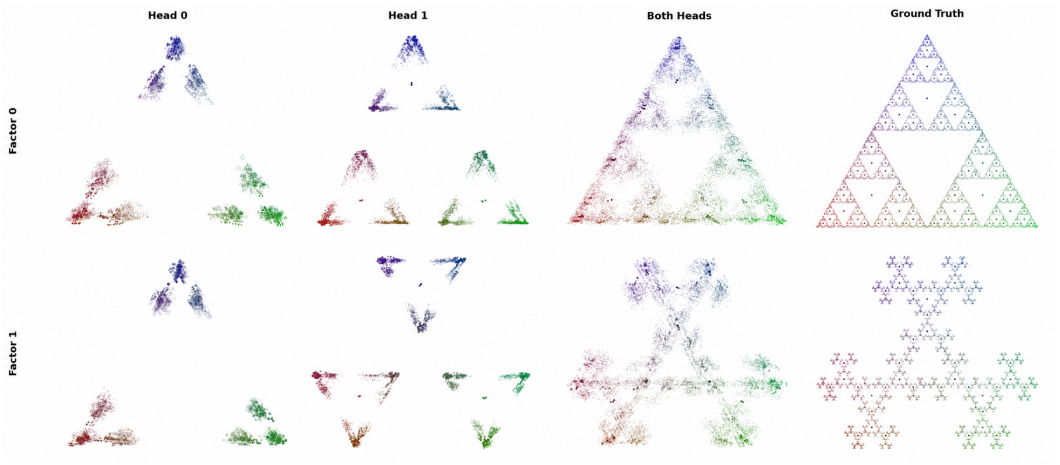

Figure 15: For $\zeta_1 = 0.5$, $\zeta_2 = -0.5$, regressing from the individual heads + the sum of both heads to the constrained belief simplex shows that no head is strongly responsible for specific ground truth geometry. In fact, both geometries require both heads.

N Experiment Details

N.1 Model Hyperparameters

We train pre-norm, decoder-only transformers on next-token prediction with cross-entropy loss. We use a single attention layer, sequence length of 11 (including BOS token), batch size of 128, $d_{\text{model}} = 120$, $d_{\text{head}} = d_{\text{model}} \,//\, n_{\text{heads}}$, and $d_{\text{MLP}} = 4 \cdot d_{\text{model}}$. We use the Adam optimizer with a learning rate of $5e-4$ and train for 100,000 steps.

N.2 Compute

We trained on Nvidia H100 GPUs, with each model training being limited to a single instance (no distributed training). No model at any point during training took up more than 10GB of total GPU memory, and every model required less than 20 minutes of training.

N.3 Mess3 Parameters

Table 4: Mess3 parameters for each two-factor configuration. For each factor, the transition parameter is determined by $p = (1-\zeta)/3$, and the emission parameter is fixed to $q = 0.6$.

Configuration	ζ_1	ζ_2	p_1	p_2	q_1	q_2
Distinct positive	$+0.7$	$+0.4$	0.1	0.2	0.6	0.6
Identical positive	$+0.7$	$+0.7$	0.1	0.1	0.6	0.6
Same magnitude, mixed sign	$+0.5$	-0.5	$1/6$	0.5	0.6	0.6
Distinct magnitude, mixed sign	$+0.7$	-0.5	0.1	0.5	0.6	0.6
Identical negative	-0.5	-0.5	0.5	0.5	0.6	0.6

O Effective Subspace Attention in a Real Model

In the main text our data-generator gives us both ground-truth subspaces and ground-truth decay rates. This however opens up the question of whether effective subspace attention is a reasonable way to think about disentangling computations in real language models. To do this, we study GPT-2 small on two tasks: indirect object identification (IOI), whose circuit is known (Wang et al., 2023), and

correlative-conjunction pairing (*neither* → *nor*, *either* → *or*), for which we could not find a published circuit. This gives us one existing circuit to validate the method against and unknown circuit to see if it has any utility as circuit discovery.

O.1 Setup

We use GPT-2 small in TransformerLens and, following its convention, fold each LayerNorm's learned scale and bias into the weights that read from it, subtract the d_{model}-mean from every weight that writes to the residual stream, and subtract the vocabulary mean from the unembedding. All three are exact reparameterizations that leave the output distribution unchanged, but cause each head's write to contain only what downstream layers can read. We write head h in layer l as $l.h$. For a task with a correct token a and an incorrect token b, we take the subspace map as the projection onto the logit-difference direction $\ell = W_U[:, a] - W_U[:, b]$. This is a one-dimensional subspace of the residual stream that the unembedding reads linearly.

When the final layer-norm scale is held frozen, the logit difference decomposes exactly into per-(head, source) terms, MLP terms, and a direct-path remainder. On IOI the terms sum to the behavioral logit difference to within 3×10^{-5}. All effects below are in logit-difference units. The per-head total is $\alpha^{(h)}(d) = \sum_s \alpha^{(h)}(d, s)$.

O.2 Experiment

We use 256 IOI prompts which consists of 4 templates, ABBA and BABA ordering, and 32 name pairs. All filler is single token. Letting $\ell = u_{\text{IO}} - u_{\text{S}}$, the model's logit difference is 3.48 ± 1.29, with the correct sign on 100% of prompts. Wang et al. (2023) reported ≈ 3.3.

Per-head attention patterns are misleading. The negative name movers 10.7 and 11.10 attend to the indirect object about as strongly as the name movers do: 0.79 and 0.62, compared with 0.64 for name mover 9.6. However, they they contribute -2.2 and -1.2 from that position. If you read the attention maps one head at a time, it would seem like these heads work *against* the task. Their write directions within $\text{span}\{u_{\text{IO}}, u_{\text{S}}\}$ show the same structure as Figure 5. The positive movers are mutually aligned (pairwise cosines $+0.78$ to $+0.95$), while 9.9 and 10.7 are anti-parallel (cosine -0.98). The routed aggregate is necessary to see the circuit.

Self-repair re-selects the head-level decomposition. Mean-ablating 9.9 (with per-template, per-position means) removes its $+3.02$ direct contribution. The routed aggregate drops by only 0.42, because other heads re-supply 85% of what was removed, and the behavioral logit difference recovers by 82%. The re-supply is signed: the largest single compensator is 10.7 weakening its suppression ($+1.31$), ahead of the backup name movers ($+1.03$ combined). Ablating 10.7 instead makes 11.10 nearly double its suppression ($-0.89 \rightarrow -1.49$). These patterns hold under both mean and zero ablation and across four ablation targets. When one head is removed, the model re-selects the head-level decomposition and keeps the aggregate. This is the head-level non-identifiability of Section L, observed under intervention. Note that heads that do not originally write to the relevant subspace but do so when other heads are ablated are undetectable by this intervention, since their OV circuit changes in response. They require ablation to verify.

O.3 A new circuit: correlative-conjunction pairing

Choosing a task by attention share. The theory applies where the attention layer's routed update is read out linearly. To pick a task where that holds, we first screened candidate tasks by the share of the behavioral logit difference carried by the attention-head aggregate, computed with the exact decomposition above (Table 5). All five tasks are solved on 100% of prompts. We chose *neither/nor* pairing: attention carries more than the full effect (share 1.08), while the MLPs slightly oppose the correct answer.

Data. The prompts have the form "{opener} {neither|either} {subject}", with 3 openers, 8 subject phrases, and both trigger words (48 prompts, 24 per trigger), for example "She liked neither the music". The readout is $\ell = u_{\text{nor}} - u_{\text{or}}$, with the sign flipped on *either* prompts so that positive always means the correct partner. The destination d is the final (subject) token.

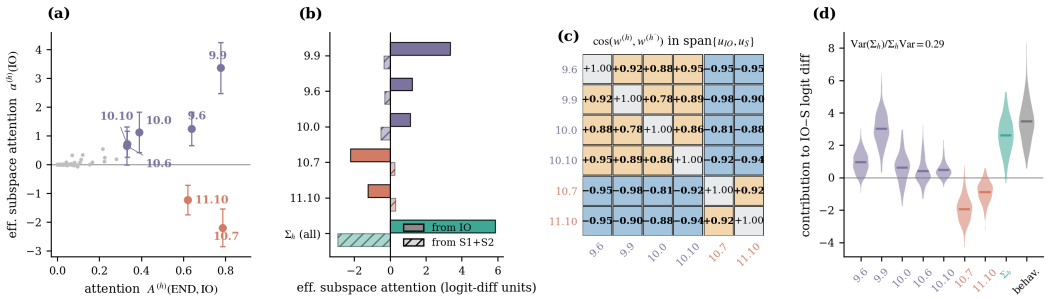

Figure 16: IOI in GPT-2 small. Negative name movers attend to the indirect object as strongly as the name movers but contribute with the opposite sign (anti-parallel writes, cosine -0.98). Head contributions anti-covary across prompts, while the routed aggregate tracks behavior.

Table 5: Task screen by attention share in GPT-2 small. Columns are the behavioral logit difference and its exact split into the attention-head aggregate and the MLPs.

task	behavioral	attention	MLPs	attn. share
quote closing	$+5.87$	$+3.28$	$+2.60$	0.56
bracket closing	$+8.39$	$+5.56$	$+2.54$	0.66
neither/nor	$\mathbf{+4.65}$	$\mathbf{+5.04}$	$\mathbf{-0.39}$	**1.08**
list separator	$+4.16$	$+4.16$	$+0.01$	1.00
pronoun gender	$+2.77$	$+2.84$	-0.07	1.03

The circuit. Across prompts, we screen heads and choose ones that satisfy $|\overline{a^{(h)}}| > 0.1$ consistently. These, with their mean contributions, are: 9.10 ($+1.45$), 10.5 ($+1.19$), 8.5 ($+0.67$), 6.11 ($+0.33$), 7.4 ($+0.24$), 4.1 ($+0.18$), and 11.8 ($+0.15$). Unlike IOI, this circuit has no suppressor heads; all contributions are positive. Checking contributions by their source position shows that they come from the trigger word. For example, 9.10 attends to *neither/either* with weight 0.63 and draws $+1.37$ of its $+1.45$ from that position. The exception is 7.4, which contributes mostly from other positions.

Mechanism. We project each head's OV image of the trigger onto the vocabulary. For every top head, the trigger's *partner* ranks at about 0 of 50,257: *neither* maps to "nor" at rank 0, and *either* maps to "or" at rank 0–2. The trigger itself ranks between 56 and 1935. Since the partner never appears within the context, this is neither copying nor is it induction. Instead, the heads must store the lexical pairing in their OV weights: each attends to the trigger and emits its partner. Within the ensemble, 9.10 and 10.5 discriminate *nor* from *or* (on *neither* prompts the wrong partner ranks 200 and 112 respectively), whereas 8.5 promotes conjunction continuations generically.

Geometry: a fully parallel ensemble. Within $\mathrm{span}\{u_{\mathrm{nor}}, u_{\mathrm{or}}\}$, all seven heads write in nearly the same direction (pairwise cosines $+0.90$ to $+1.00$). The heads here are fully collaborative and redundant.

O.4 Causal validation

Head ablation. Zero-ablating 9.10 removes $+1.45$ of direct write but costs only -0.65 in behavior, because 10.5 re-supplies $+0.53$. Zero-ablating 8.5 (-0.67 direct) costs -0.21, and again 10.5 compensates most ($+0.26$). Zero-ablating the top three heads together costs -2.71 and drops accuracy from 100% to 50%. As in IOI, the circuit repairs itself when a single head is removed, but the heads are necessary as a group. Mean-ablating all seven heads at once, with means taken separately over *neither* and *either* prompts, leaves behavior unchanged ($\Delta\mathrm{ld} = +0.04$). The circuit's entire routed update is the mean vector for each trigger: it carries the trigger's identity and nothing specific to the prompt.

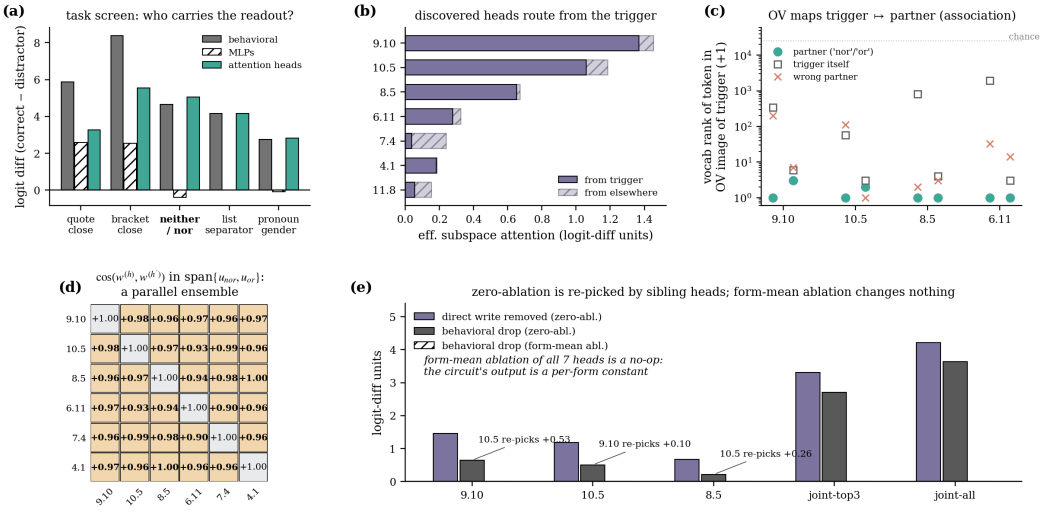

Figure 17: The correlative-conjunction circuit in GPT-2 small. (a) Attention-share task screen. (b) Per-source routing: the contribution comes from the trigger word. (c) OV association: each head's OV image of the trigger ranks the partner near the top of the vocabulary. (d) Parallel write geometry in $\mathrm{span}\{u_{\mathrm{nor}}, u_{\mathrm{or}}\}$. (e) Self-repair under zero-ablation and a null effect under group-mean ablation.