WHY LANGUAGE MODELS DISAMBIGUATE CONTEXTS DESPITE SIMILAR EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models present an apparent paradox where the representations of a token across distinct contexts exhibit very high cosine similarity, and yet they cleanly disambiguate contexts. Here, we analyze the tangent space dynamics of the embedding and unembedding spaces, to understand the influence of context and layer depth on the output. We introduce a two-fold decomposition that separates representations into (i) a shared component common across contexts and its context-specific orthogonal complement, and (ii) readout-sensitive and readout-insensitive subspaces determined by their direct linear projection through the unembedding matrix. From large scale empirical studies on 16 models of 160M to 9B parameters, we observe that the shared context-invariant subspace is high-dimensional, occupying a large fraction of the ambient embedding space. The readout is dominated by the component shared across contexts, which is why representations across contexts look nearly identical. Separation survives because the shared components cancel and the contrast component is enriched tenfold in the subspace the readout acts on.

1 INTRODUCTION

A token's representation must adapt to its context, but how does the context influence the token embedding across layers of a language model? Further, how is the context information present in these token embeddings read? Multiple lines of work have studied the representation geometry of large language models. Sparse autoencoders have been used to decompose activations into interpretable features Cunningham et al. (2023); Bricken et al. (2023), the local manifold structure of hidden states has been studies Ansuini et al. (2019); Valeriani et al. (2023), and concepts have been found to arrange as linear directions or convex regions Subramani et al. (2022); Park et al. (2025). Context is central to how transformers work, yet its role remains under-studied with respect to its impact on the underlying geometry. Prior work has investigated dynamics across transformer layers via probing task hierarchies Tenney et al. (2019), vocabulary projections lenses nostalgebraist (2020); Belrose et al. (2025), concept promotion in feed-forward networks Geva et al. (2021), and causal tracing of the information flow Meng et al. (2023). These approaches primarily focus on decoding the output to measure how intermediate states map to the output predictions. Representations are also known to become increasingly context-specific with depth Ethayarajh (2019). What remains unclear is the question of how the surrounding context is integrated into the representation space across layers and read during terminal layer readout.

Polysemous words serve as a test bed for understanding representation dynamics in transformers. A polysemous word has one spelling and two or more meanings that are only distinguished by context. For example, the word *bank* after *loan* has a meaning distinct from *bank* after *river*. Naturally, in a spoken conversation, this distinction is revealed through context. We seek to understand how language models transform embeddings and use context to disambiguate the sense behind a polysemous word. For a fixed token, difference in the representation is attributable to the context, with the ground truth sense as the label that the model must distinguish. Our first question is simply how much the representation changes. Across sixteen models, we measure the two senses' final-layer centroid to have cosine similarity above 0.99 (Appendix E, consistent with the narrow cone that contextualized embeddings are known to occupy Ethayarajh (2019); Gao et al. (2019); Timkey & van Schijndel (2021). Yet, a linear probe recovers the sense from these at above 0.80 accuracy.

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

The high cosine similarity reveals that the representations differ by a small variation around a large shared component. We find that this variation is not arbitrary. For each word, we write the two sense centroids as $\bar{h}_{a,b} = \mu \pm \frac{1}{2}\delta$, splitting them into what the sense share and what distinguishes them. Separately, and independently of any word, we split the embedding space into a readout-sensitive subspace S_{out} and its complement depending on how strongly the output responds to each direction. We draw a distinction between these two decompositions. μ and δ are directions, estimated per word from the model's activations, while S_{out} is a subspace that is fixed per model and read from the weights. The measurement that crosses them asks what fraction of each direction lies in which subspace. Neither direction sits entirely in S_{out}, nor is either excluded from it. δ places 13.4% of its energy there ($10\times$ enrichment over a random direction), and μ places 34.7%. The distinguishing part of the representation is thus shown to concentrate where the output is most sensitive, but it is not confined there entirely. We measure this across layers, observing that this distinction is present from early in the network and that the final layer is where the representation aligns with the read directions.

We pose the following central questions: (1) How does context change the difference between a token's representations under its two senses? (2) How is internal state information mapped into output logits? (3) What space does the sense distinguishing component occupy relative to the readout subspace? (4) What determines the trajectory of sense separation across layers? We contribute a systematic empirical study of these questions across sixteen open-weight models from 160M to 9B parameters spanning six architectural families.

1. *Context drives sense disambiguation.* The separation between a word's two senses grows with the amount of disambiguating context in every model, by a median factor of 4.3, and more examples land on the correct side of the boundary as context accumulates (Section 4).
2. *Proportional routing into the readout subspace.* The sense distinguishing component is concentrated where the readout is sensitive, and contributes proportionally. Sense contrast δ is enriched $10\times$ in S_{out} relative to chance in all models. Causal ablation confirms that removing its S_{out} component degrades sense distinction in proportion to the share δ holds there ($R = 0.926$; Section 5.1, Section 5).
3. *Sense–readout alignment appears in one step at the terminal layer.* Across all sixteen models, the largest single-layer surge in readout subspace energy occurs at the final layer, exceeding mid-network variability by a median factor of $763\times$. Through the first two-thirds of depth, S_{out} carries no more energy than an arbitrary subspace (Section 6).
4. *Depth-wise sense separation is determined by a single ratio.* The entire layer-by-layer profile of sense separation collapses to the scalar ratio $r = \|\delta\|/\|\mu\|$ between sense distinguishing component δ and shared component μ, demonstrating that terminal cosine collapse is a norm artifact rather than signal loss (Section 6, Section 6.1).

2 Related Work

Geometry, anisotropy, and representation collapse. Anisotropy occurs in language models as contextualized representations tend to occupy a narrow cone in the vector space Ethayarajh (2019); Gao et al. (2019); Timkey & van Schijndel (2021). This phenomenon is closely tied to rank collapse and representation collapse in deep transformers, where layer dynamics cause feature representations to lose variance and compress into lower-dimensional subspaces Dong et al. (2021); Wang et al. (2022). Furthermore, the intrinsic dimensionality of these representations fluctuates across network depth, typically expanding in early layers before compressing heavily toward the final layers Ansuini et al. (2019); Valeriani et al. (2023). In Sections 5.3 and 6, we prove that high cosine similarity (the ambient condition of late-layer embeddings) is an artifact of final-layer norm growth rather than a loss of linearly separable sense information.

Context-dependent dynamics and activation caching. Moving beyond static geometric properties, context dynamically conditions representation trajectories across network depth. Recent work demonstrates that contextual reasoning paths exhibit structures, predictable activation dynamics across layers. This enables mechanisms such as ReasonCache Gupta et al. (2026) to cache and reuse context-dependent state trajectories during LLM inference. Such methods leverage temporal and contextual redundancies in the intermediate representations, however, it remains open how

the surrounding context is integrated into a token's representation across depth and context. In this work, we formally characterize how context shifts token representations within the residual stream and we evaluate its impact on downstream readout operations.

Concept directions and polytope geometries. Extracting interpretable concepts from embeddings often relies on identifying linear directions in the representation space. The baseline method defines these concept directions as the difference-of-means between contrasting sets Turner et al. (2024); Zou et al. (2025); Subramani et al. (2022); Panickssery et al. (2024). Recent work extends this linear intuition by demonstrating that categorical concepts form bounded convex polytopes in the hidden space of language models Park et al. (2025). This motivates our view of context as the parameter responsible for shifting representations into distinct regions of embedding space (Section 4).

Interpreting the residual stream and readout projections. Mechanistic interpretability frequently treats the residual stream as a central communication channel Elhage et al. (2021). To decode intermediate states, tools such as the logit lens nostalgebraist (2020) and tuned lens Belrose et al. (2025) project hidden representations directly into the vocabulary space using the unembedding matrix. However, these methods often require arbitrary affine translations or learned affine adapters to stabilize intermediate predictions. In Section 5.1, we depart from this convention by defining a folded readout operator that rigorously integrates the final layer norm's gain and mean-centering projection, making layer-wise geometric measurements well-defined without intermediate approximations. Furthermore, the topography of the unembedding matrix itself is known to suffer from structural artifacts, such as outlier or "rogue" dimensions that heavily distort the singular value spectrum Timkey & van Schijndel (2021); Puccetti et al. (2022). We encounter this phenomenon directly in Section 5.2, where we diagnose near-rank-one readouts in a subset of evaluated models.

Activation steering and causal interventions. Concept directions can be causally validated through activation steering, that is, intervening on the forward pass by adding or subtracting steering vectors to shift model generation Subramani et al. (2022); Turner et al. (2024); Panickssery et al. (2024). Prior work establishes that steering becomes ineffective at the final layers as representations fix their predictions Meng et al. (2023). We apply targeted vector ablation within the specific subspace defined by the folded readout operator to confirm that sense information is not only proportionally routed to the readout subspace but also causally governs the final vocabulary projection (Section 5).

3 Setup

Let a decoder-only transformer have L blocks and residual-stream dimension d. For a token at a fixed position let $h_l \in \mathbb{R}^d$ be its hidden state after block l, with h_0 the embedding. Let c be the number of preceding tokens of disambiguating context before the target word, drawn from $\mathcal{C} = \{0, 1, 2, 3, 5, 8, 13, \text{full}\}$, where "full" keeps the entire preceding sentence. A token's trajectory is then a map $h_{(\cdot)}(c) : \{0, \dots, L\} \times \mathcal{C} \to \mathbb{R}^d$. We study how the representations evolve over the space $\{0, \dots, L\} \times \mathcal{C}$. Figure 1 illustrates this space across context and layers, along with our two-fold decomposition of the embedding space to provide visual intuition for this paper. Appendix P provides all experimental setup information concerning models, word stimuli, and context.

4 Context and the Sense-Specific Decomposition

This section addresses the question: *How does context change the difference between a token's representations under its two senses?* A polysemous token begins ambiguous in sense and is committed at some location in layer $\times$ context space. We determine where across layers and context this happens, and how.

Fix a word with two senses, a layer l and a context c, and let $\bar{h}_a$ and $\bar{h}_b$ be the mean hidden states (centroids) for that sense. Both depend on l and c, which we suppress throughout since every comparison below holds them fixed. Write $\mu = \frac{1}{2}(\bar{h}_a + \bar{h}_b)$, and $\delta = \bar{h}_a - \bar{h}_b$, so μ is the component the two senses share and δ the direction along which they differ. Every measurement in this paper is a measurement of μ, of δ, or of how one of them sits relative to the readout subspace of Section 5.1. The reason to work with this split rather than with the angle between the centroids is that the angle conflates two things the rest of the paper needs to separate. When the two centroids have equal

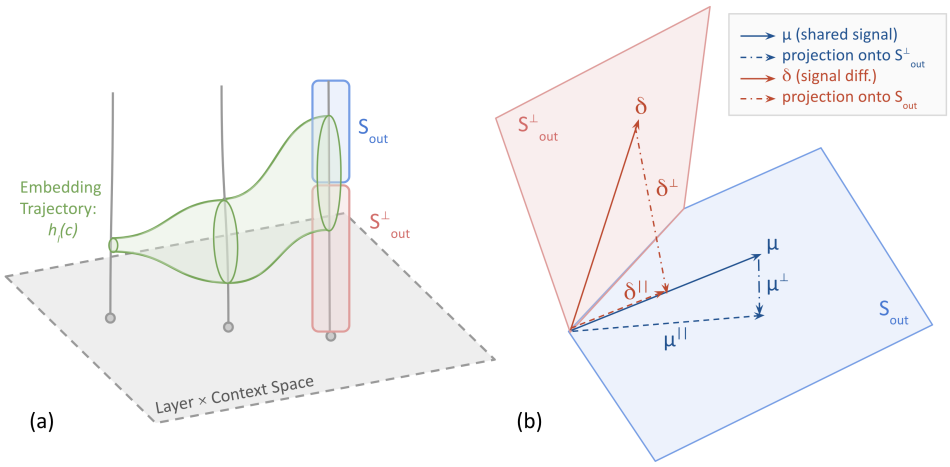

Figure 1: Overview of the two decompositions. **(a)** *Layer–context space.* The base plane is $\{0, \ldots, L\} \times \mathcal{C}$, with layer l and context c as its coordinates. Over each base point sits the d-dimensional residual stream, drawn as a vertical line since d cannot be shown directly; height along it stands for position within the stream, not for any single coordinate. The green curves are one word's two sense trajectories, coinciding at low context and separating as context accumulates. S_{out} and $S^{\perp}_{out}$ are fixed subspaces of $\mathbb{R}^d$, the same at every layer, marked at the final fiber where the readout is applied. **(b)** *The two decompositions at the final layer.* μ is the direction the two senses share and δ the direction distinguishing them, both vectors in $\mathbb{R}^d$ estimated per word from activations. Separately, $\mathbb{R}^d$ splits by how strongly the readout responds to each direction, into S_{out} and $S^{\perp}_{out}$, fixed per model and computed from the weights alone. The splits are not nested, so μ and δ each have a nonzero projection in both subspaces ($\mu^{\parallel}, \mu^{\perp}$ and $\delta^{\parallel}, \delta^{\perp}$). Most of both lies outside S_{out} (65% of μ, 87% of δ), but δ places about $10\times$ more of its length there than chance allows. Removing $\delta^{\parallel}$ reduces the sense separation by roughly the fraction of δ that S_{out} holds.

norm,

$$1 - \cos(\bar{h}_a, \bar{h}_b) = \frac{r^2/2}{1 + r^2/4}, \qquad r = \frac{\|\delta\|}{\|\mu\|}, \tag{1}$$

so the angle is a function of the single ratio r and of nothing else. The angle can shrink either because δ shrinks or because μ grows, and these are different events with different implications; r tells them apart while the cosine does not. And Equation 1 does not involve the direction of δ at all, so however small the angle becomes, it places no constraint on where the distinguishing information sits. This is what makes Section 5 a question worth asking rather than one already settled by the collapse. Since $\mu \cdot \delta = \frac{1}{2}(\|\bar{h}_a\|^2 - \|\bar{h}_b\|^2)$, the centroids have equal norm exactly when $\delta \perp \mu$. Section 5.3 verifies that Equation 1 is accurate to within 3% in every model.

Every measurement in this section compares two context conditions at a fixed layer. This is deliberate because a depth-wise comparison requires a baseline layer, and at layer 0 every stimulus of a word shares one token embedding for a model with rotary position encoding, so both sense centroids coincide and the sense-contrast direction is determined by tokenization variants alone. Comparing context conditions at one layer does not require such a baseline.

We ask whether individual stimuli follow the centroids. For a stimulus x of a sense a, the signed margin is $m(x) = d_{\cos}\big(x, \bar{h}_b^{-x}\big) - d_{\cos}\big(x, \bar{h}_a^{-x}\big)$, the amount by which x lies closer to its own sense's centroid than to the other's in cosine distance, with $\bar{h}^{-x}$ denoting leave-one-out defined centroids. We report the mean of $m(x)$ and, separately, $\mathbb{E}[m(x)]/\mathbb{E}[|m(x)|]$, which we call *coherence*. Coherence is 1.0 when every stimulus sits on the correct side and falls toward zero as the margins become directionless. The margin asks how far apart the senses are, and the coherence asks what fraction of the mass is on the right side of the boundary.

The sense centroids move apart as context accumulates. $\|\delta\|$ is smallest at low context and grows with it in every model. The locus at which the two senses are least distinguishable is therefore a position in context, not a layer. Figure 2 shows this trend directly.

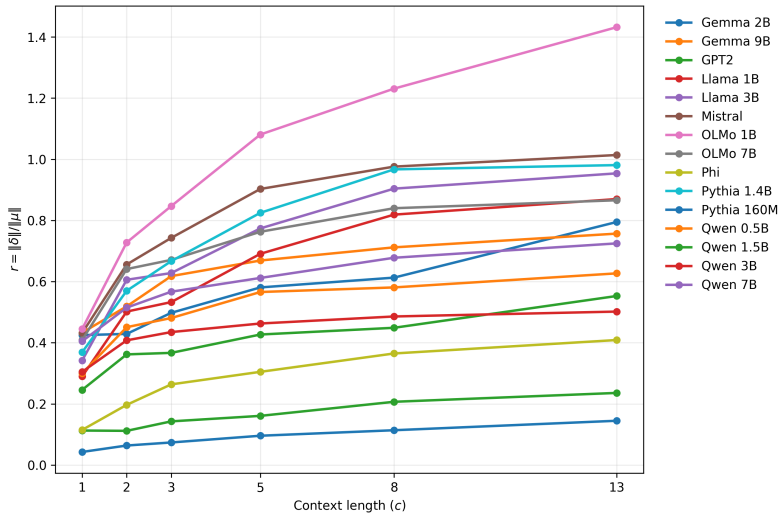

Figure 2: Centroid separation against context at the final layer. $r = \|\delta\|/\|\mu\|$ increases with context.

Table 1: All quantities compare context conditions at the final layer. The *margin* is the signed amount a stimulus lies nearer its own sense's centroid than the other's, in cosine distance. *coherence* is the signed margin divided by its absolute counterpart, so 1.0 would mean every stimulus sits on the correct side. *context ratio* is the corresponding ratio for the contrast alignment.

	signed margin			coherence		
model	short ctx	full ctx	ratio	short ctx	full ctx	context ratio
Gemma 2B	0.014	0.018	1.3	0.43	0.37	1.41
Gemma 9B	0.014	0.028	2.0	0.48	0.48	0.85
GPT-2	0.002	0.006	2.3	0.94	0.91	1.00
Llama 1B	0.014	0.077	5.5	0.71	0.80	1.84
Llama 3B	0.019	0.105	5.6	0.68	0.78	1.80
Llama 8B	0.027	0.138	5.1	0.70	0.75	1.25
Mistral 7B	0.032	0.135	4.2	0.78	0.81	1.08
OLMo 1B	0.030	0.148	5.0	0.72	0.77	1.49
OLMo 7B	0.026	0.107	4.2	0.71	0.76	1.29
Phi 3.5	0.004	0.021	5.7	0.75	0.80	1.30
Pythia 160M	0.000	0.002	4.9	0.69	0.80	1.07
Pythia 1.4B	0.024	0.119	5.0	0.71	0.80	1.52
Qwen 0.5B	0.014	0.062	4.4	0.65	0.76	1.24
Qwen 1.5B	0.009	0.037	4.2	0.64	0.72	1.21
Qwen 3B	0.015	0.054	3.6	0.65	0.74	1.20
Qwen 7B	0.021	0.070	3.4	0.62	0.70	1.04
mean	0.017	0.071	4.3	0.68	0.74	1.29

Individual stimuli follow, and context moves more stimuli to the correct side. The signed margin grows with context in all sixteen models, by a factor of 1.3 to 5.7 (median 4.3), and is positive at full context everywhere. Coherence averages 0.68 at short context and 0.74 at full, rising in thirteen of sixteen models (Table 1). Context therefore does more than merely move the centroids apart. It moves a larger fraction of stimuli into the correct sense component.

As an independent check we measure how strongly each stimulus is aligned with its word's contrast direction, $|\hat{\delta} \cdot \hat{h}|$, where $\hat{\delta} = \delta/\|\delta\|$ and $\hat{h} = h/\|h\|$. This is larger at full context than at 0–1 tokens in eleven of sixteen models, mean ratio 1.29. Directions are cross-fitted, estimated on half of a word's stimuli and evaluated on the other half. The five models below the 1.1× threshold (GPT-2 (1.00), Gemma-9B (0.85), Qwen-7B (1.04), Pythia-160M (1.07) and Mistral (1.08)) include the two with

by far the smallest margins in the roster, 0.006 and 0.002 against a mean of 0.071. They barely disambiguate on these stimuli, so there is little to detect. Across models the $p = 0.083$).

Sense separation is smallest at low context. The quantity $||\delta||$ is smallest at low context and grows as context accumulates. Where the two senses are least distinguishable is therefore a position in context, not a layer, which is why every comparison in this section is made at a fixed layer. We show that the context reorganizes the representation geometry of a polysemous token, yet this raises the next question concerning when and how this information is read downstream.

5 Where the Sense Information is Represented

Section 4 established that context separated the two senses. However, a direction the output never consults could be reorganized freely with no effect on readout behavior. This section asks whether the contrast δ lies where the output is sensitive. We measure three things. First, what fraction of $\|\delta\|^2$ lies inside S_{out}, compared against the k/d a direction unrelated to the readout geometry would place there. Second, how strongly M amplifies δ itself, as $\mathrm{amp}(\delta) = \|M\hat{\delta}\|^2/(\|M\|_F^2/d)$ with $\hat{\delta} = \delta/\|\delta\|$, so that 1.00 means the readout treats δ like an average direction. Third, what happens to the model's logits when δ's S_{out} component is removed from the final-layer state.

5.1 What the Output Reads

Next, we ask: *How is internal state information mapped into output logits?* A natural answer would be the unembedding matrix W_U, however, W_U does not act on the residual stream directly, rather, a normalization layer sits between them. With logits $W_U \cdot \mathrm{LN}_f(h_L) + b$, we write the final norm as $\mathrm{LN}_f(h) = g \odot (Ph/\sigma(h)) + \beta$, with g a learned per-coordinate gain, β a bias, $\sigma(h)$ a scalar, and $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ the mean-centering projection ($P = I$ for RMSNorm). Collecting terms gives

$$W_U \cdot \mathrm{LN}_f(h_L) + b = M\frac{h_L}{\sigma(h_L)} + b_0, \qquad M := W_U \,\mathrm{diag}(g)P, \tag{2}$$

with $b_0 = W_U\beta + b$ constant in h_L. So M maps the residual stream to the logits. Appendix G grounds on our choice of M rather than W_U throughout. Now, write the SVD as $M = U\Sigma V^\top$ with singular values $s_1 \geq s_2 \geq \ldots$ in the usual descending order and right singular vectors v_j, and define $S_{out} = \mathrm{Span}\{v_1, \ldots, v_k\} \subset \mathbb{R}^d$, the k directions the readout responds to most strongly. $\|Mv_j\| = s_j$, so a unit step along v_j moves the logits by s_j, and for an arbitrary unit direction u, $\|Mu\|^2 = \sum_j s_j^2\,(v_j \cdot u)^2$, where $\mathbb{E}_{u\sim\mathrm{Unif}(S^{d-1})}\left[\|Mu\|^2\right] = \frac{\|M\|_F^2}{d}$. We compare every measurement against the expectation because it is exact, computed in closed form from the weights rather than estimated by sampling random directions, making claims across models are directly comparable.

5.2 The readout spectrum

We study the size of S_{out}. At a 90%-mass criterion the required rank is 67–81% of d. The participation ratio $(\sum_j s_j^2)^2/\sum_j s_j^4$, the standard effective-rank measure, gives 13–145, which against d between 768 and 4096 is 0.4–8% of the ambient space. Both of these are correct because the readout spectrum has a heavy head and a long, nearly flat tail. A few dozen directions dominate the response and the remaining mass is spread thinly enough that accumulating 90% of it takes most of the space. We therefore fix $k = 32$ and report the chance level k/d for meaningful comparisons. Matching k to a mass criterion instead would set it anywhere from 1 to 1400 across the roster, so k/d would differ by three orders of magnitude and an enrichment metric would mean something different in each model. Appendix O considers $k = 8$ and $k = 128$ and finds that the sense contrast is enriched above chance at every rank in every model, so the finding is a property of the readout subspace. This also bounds how far S_{out} can be attributed as the part used directly by the readout. The $k = 32$ directions capture 7.6–49% of $\|M\|_F^2$, so the complement holds 51–92%, averaging 80%. Also, GPT-2 and Pythia-160M have a participation ratio of 1, meaning that a single direction of M accounts for almost all of its squared response, and 32 directions capture 87–93% of the mass, against 8–49% for every other model. For these two, S_{out} is effectively one-dimensional and the chance level k/d is a poor description of what an arbitrary direction would achieve. These are also

Table 2: Where the sense information goes. *frac* is the fraction of $\|\delta\|^2$ inside S_{out} and *enrich* divides it by chance. *amp* is $\|M\hat{u}\|^2$ divided by its exact expectation under a uniformly random unit direction, $\|M\|_F^2/d$, so 1.00 means the readout treats that direction as any other. The *collapse* columns give the fractional reduction in sense-centroid separation when a component is removed at the final layer, where the readout is the entire remaining computation; *random* is a unit direction drawn from within S_{out}. Logit-movement ablations are in Appendix L.

	concentration		readout response			collapse on removing		
model	frac δ	enrich δ	amp δ	amp μ	$\|\delta\|/\|\mu\|$	δ_S	$\delta_\perp$	random
Gemma 2B	0.130	9.4	15.13	1.81	0.285	0.154	0.826	+0.001
Gemma 9B	0.047	5.3	5.03	10.16	0.349	0.038	0.945	+0.000
GPT-2	0.141	3.4	43.10	261.67	0.052	0.031	0.868	−0.000
Llama 1B	0.077	5.0	1.55	6.04	0.446	0.050	0.911	+0.002
Llama 3B	0.074	7.1	1.49	2.99	0.544	0.055	0.908	+0.003
Llama 8B	0.026	3.3	1.04	1.77	0.708	0.019	0.965	+0.000
Mistral 7B	0.272	34.8	3.45	4.43	0.552	0.287	0.612	+0.001
OLMo 1B	0.132	8.4	2.38	6.28	0.722	0.110	0.832	+0.001
OLMo 7B	0.124	15.9	1.79	74.26	0.508	0.119	0.843	−0.006
Phi 3.5	0.070	6.7	2.24	94.58	0.220	0.014	0.925	−0.017
Pythia 160M	0.205	4.9	0.69	693.70	0.041	0.233	0.765	+0.003
Pythia 1.4B	0.099	6.3	1.27	1.39	0.550	0.068	0.877	−0.009
Qwen 0.5B	0.143	4.0	1.81	1.09	0.305	0.110	0.855	−0.014
Qwen 1.5B	0.234	11.3	2.00	2.83	0.194	0.228	0.753	+0.003
Qwen 3B	0.206	13.2	2.18	2.46	0.297	0.221	0.747	−0.001
Qwen 7B	0.170	19.0	14.94	19.89	0.362	0.129	0.816	+0.001
mean	0.134	9.9	6.26	74.08	0.383	0.117	0.841	−0.002

the two models with the smallest sense margins in Section 4. A current limitation of this work is that we do not yet have a mechanism connecting these.

5.3 The cosine collapse does not constrain where sense lives

Equation 1 relates the angle between the sense centroids to the single ratio $r = \|\delta\|/\|\mu\|$, and Appendix H verifies that across all sixteen models the measured $1 - \cos(\bar{h}_a, \bar{h}_b)$ agrees with $(r^2/2)/(1 + r^2/4)$ to within 3.2%, with a maximum absolute error of 0.004. We can now draw two conclusions. Equation 1 is an identity, rather than an empirical prediction. The same two vectors are used to compute μ, δ and the cosine between the centroids, so the relation is algebra. What the verification establishes is that the decomposition is clean and that δ is close to orthogonal to μ, so the angle is governed by the magnitude ratio alone. Since $\mu \cdot \delta = \frac{1}{2}(\|\bar{h}_a\|^2 - \|\bar{h}_b\|^2)$, orthogonality holds exactly when the two centroids have equal norm, and hidden-state norms at a given layer are tightly concentrated. The small residual is the second-order contribution of $\mu \cdot \delta \neq 0$. Second, this places no constraint on the direction of δ. Mean-centering the sense centroids leaves them anti-aligned rather than nearly parallel ($+\delta/2$ and $-\delta/2$). A cosine similarity of 0.99 is therefore entirely a statement about how large μ is relative to δ.

The sense direction is concentrated in the readout subspace. In all sixteen models $\|\delta\|^2$ places more of its energy inside a 32-dimensional S_{out} than chance allows: 13.4% against $k/d \approx 1.6\%$, an enrichment of 9.9× on average (median 6.9×, range 3.3–34.8×). Because δ is a difference of centroids, the shared component cancels from it exactly, so this cannot be inherited from the alignment of the state as a whole. It is a property of the sense contrast.

The decomposition is complete and proportional. Removing δ's S_{out} component collapses the separation by 0.117 on average; removing its complement collapses it by 0.841. Across models, the collapse produced by removing δ_S tracks the fraction of δ that S_{out} holds with $R = 0.926$ ($p = 2.6 \times 10^{-7}$). S_{out} carries the sense distinction in exact proportion to how much of δ it contains. Sense information is concentrated in (albeit not confined to) the readout subspace.

The readout responds to the sense direction and the state. Recall $\mathrm{amp}(u) = \|M\hat{u}\|^2/(\|M\|_F^2/d)$ compares the squared logit movement produced by a unit step along u against the exact expectation

of that quantity for a uniformly random unit direction, so $\mathrm{amp} = 1$ means the readout treats u like an average direction. The sense distinction clears that bar in twelve of sixteen models, with a mean $\mathrm{amp}(\delta)$ of 6.3. It amplifies the shared component more strongly still in fifteen of sixteen, in several cases by orders of magnitude: Phi-3.5 reads μ at $95\times$ chance and δ at $2.2\times$, and Pythia-160M reads μ at $694\times$ and δ below chance at $0.69\times$. This is the terminal concentration of Section 6 seen from the readout's side. At layer L most of the state is the prediction, making alignment expected. Table 2 adds that the sense contrast is enriched there as well, as one structured component within a subspace the whole state occupies. GPT2 and Pythia-160M have readout operators that are near-rank-one (Section 5.2). For these, almost all of $\|M\|_F^2$ sits in one direction, so both μ and δ are scored against a null dominated by that direction causing the ratios to become large and unstable.

We have uncovered that the context is preferentially routed into the most readout-responsive directions. We now analyze when the concentration of sense information into S_{out} takes place.

6 Terminal Alignment with the Readout Subspace

When does the representation become readable? For most of its depth the residual stream is a workspace rather than an answer. Directions outside S_{out} have little effect on the logits, so for most of its depth the residual stream has no need to align with M, and it can use whichever directions suit intermediate computation. Alignment is forced only where the readout is applied. We quantify this with the concentration ratio,

$$\rho_l = \frac{\mathbb{E}\big[\|\mathrm{proj}_{S_{out}} h_l\|^2\big]}{\mathbb{E}\big[\|h_l\|^2\big]}, \tag{3}$$

whose chance level is k/d (for a state with no preferential relationship to the readout geometry). The prediction is not that ρ_L is large, but rather that the increase is abrupt and terminal. Gradual convergence would have predicted ρ_l to climb smoothly with depth, but this does not apply here. Every model in Table 3 shows that the largest single-layer increase in ρ is the final layer. The size of the rise varies by a factor of 45 across models, which is why the criterion concerns the position of the step and its size relative to each model's own mid-network variability rather than an absolute threshold. On that criterion the terminal step exceeds the 95th percentile of mid-network steps by a median factor of 763, and by more than $100\times$ in thirteen of sixteen models. Notably, these models span LayerNorm and RMSNorm, tied and untied embeddings, 160M to 9B parameters and six architecture families, and none of these move the layer at which concentration occurs.

The flat region is at chance. Figure 7 in Appendix J shows that through roughly the first two thirds of depth ρ_l sits at k/d with no measurable relationship to the readout geometry at all. Then, the concentration abruptly emerges during readout at the final layer. Again, this compares to a network gradually deciding on an answer which would show ρ_l separating from chance early and rising.

The event occurs at a single layer. We observe that in eleven of sixteen models the terminal step and the total rise agree to within 0.01, meaning the entire concentration occurs in one layer; Llama-1B has a rise of 0.330 of which 0.331 occurs in the last step. OLMo-2-7B is the exception, its final layer accounting for 0.071 of a rise of 0.165, spread over the last three to five layers. We have determined thus far that sense separation grows with context, that it is routed to the subspace S_{out}, and that it is readable only at layer L. Our final question is how these three observations are tied together.

6.1 Equation 1 tracks sense separation across layers

We conclude this work with our final question which ties everything together: *What determines the trajectory of sense separation across layers?* Section 5.3 evaluated Equation 1 at the final layer only. Here we compute $r = \|\delta\|/\|\mu\|$ at every layer, predict the separation from Equation 1 with nothing fitted, and compare against the measured separation at the same layer. The claim under test is that separation at each layer is a function of the ratio at that layer.

The measured trajectory is tracked by r. Rather than an arbitrary byproduct of deep optimization, the rise and fall of semantic separation follow a tightly constrained geometric curve dictated by the magnitude ratio, r. The prediction tracks the measurement across depth in fifteen of sixteen models, significantly in fourteen, with a median Spearman correlation of $+0.926$ (Appendix I, Table 10). Nine models exceed $+0.92$ and four exceed $+0.96$. Most of the models' profiles are not simple or

Table 3: Terminal concentration of energy into S_{out}. *rise* is ρ_L minus its early-to-mid-network mean; *step* is the layer with the largest ρ increase; *step/p95* divides the increase by the 95th percentile of mid-network increases for the same model, a scale-free measure of how exceptional the terminal event is. The criterion is that the largest increase falls in the final two layers and exceeds the mid-network 95th percentile. All sixteen models satisfy it, with the largest increase at layer L exactly. The same quantities under two alternative normalization conventions are given in Appendix K.

model	L	rise	step	step / p95	confirmed
Gemma 2B	26	0.041	26	6	✓
Gemma 9B	42	0.022	42	10	✓
GPT-2	12	0.551	12	1,100	✓
Llama 1B	16	0.330	16	8,780	✓
Llama 3B	28	0.243	28	2,096	✓
Llama 8B	32	0.076	32	1,566	✓
Mistral 7B	32	0.343	32	426	✓
OLMo 1B	16	0.218	16	1,795	✓
OLMo 7B	32	0.165	32	3	✓
Phi 3.5	32	0.781	32	2,354	✓
Pythia 160M	12	0.968	12	393	✓
Pythia 1.4B	24	0.129	24	241	✓
Qwen 0.5B	24	0.149	24	127	✓
Qwen 1.5B	28	0.273	28	3,569	✓
Qwen 3B	36	0.142	36	290	✓
Qwen 7B	28	0.128	28	3,894	✓

predictable either, they have different signatures with no apparent pattern, yet a parameter-free curve follows each of them. The prediction at layer l uses only r at layer l without any fitting involved.

The ratio is not $\|\delta\|$ rescaled. μ is not constant across depth. Within the mid-network band alone, $\|\mu\|$ varies by a median factor of 1.96. Across that band r and $\|\delta\|$ are anti-correlated, at a median of -0.57 and negatively in eleven of sixteen models. Through the middle of the network $\|\mu\|$ grows faster than $\|\delta\|$, so the contrast becomes larger in absolute terms while the ratio falls. Were r a rescaling of $\|\delta\|$ the correlation would be $+1$. The cosine collapse of Section 5.3 is r at the final layer; the depth profile of sense separation is the profile of r; and the terminal contraction is a fall in r. Sense separation in these models is not an unconstrained byproduct of optimization but tracks a single scalar, from the early layers through to the terminal step.

7 Limitations

We identify the following limitations on our work. (1) Eleven polysemous words makes it such that every across-word correlation has $n = 10$, giving power 0.20 against $R = 0.3$. (2) Results are reported at fixed k with chance levels alongside, but the readout spectrum's long tail means no single k is canonical (Section 5.2). (3) GPT-2 and Pythia-160M have sense margins one to two orders of magnitude below the rest and fail several secondary tests for that reason rather than because the tested claim is false of them.

8 Conclusion

We have addressed the apparent paradox that polysemous tokens simultaneously exhibit cosine collapse and high linear separability. Polysemous disambiguation is context-driven, yet the final layer alone formats representations for terminal readout. Energy enters the readout subspace in a single terminal step, rising from a chance baseline. Sense information is preferentially routed into this subspace, contributing in exact proportion to its contextual prevalence. Context, not network depth, is the governing coordinate for sense separation. Intersecting both geometric decompositions is essential to isolate these mechanics. We have studied the tangent space dynamics and influence of context and gained a clear geometric picture via our two fold decomposition. We study this through the lens of context, but this geometric perspective applies to further the understanding of any parameter's influence on representation geometry and model behavior.

AI USE STATEMENT

We used generative AI tools minimally to help with basic tasks which do not involve intellectual contributions. We wrote all original text ourselves and occasionally used AI to edit prose for concision and clarity. We also used AI to help with the implementation of some supporting analysis and figure code. We did not use AI for any implementation of our main method. All experiments were designed and run by the authors, and their outputs were inspected by the authors, All numerical claims were checked against the saved outputs. The authors take full responsibility for the content, claims, code, figures, and artifacts of this work.

ETHICS STATEMENT

We believe that the ethical and societal implications of this work are consistent with those commonly encountered in machine learning research for scientific discovery, and we do not identify any specific risks that warrant additional discussion beyond standard considerations (i.e. responsible use).

REPRODUCIBILITY STATEMENT

All code needed to construct the set of polysemous word stimuli and contexts, to reproduce experiments, and to generate all figures within the paper is available in the supplementary material, and experimental setup information is outlined in the README. Follow the guidelines on the supplementary material README pages to reproduce the performance results, as well as the plots presented throughout the paper.

REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks, 2019. URL https://arxiv.org/abs/1905.12784.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

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.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, 2019. URL https://arxiv.org/abs/1909.00512.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models, 2019. URL https://arxiv.org/abs/1907.12009.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories, 2021. URL https://arxiv.org/abs/2012.14913.

Sharut Gupta, Phillip Isola, Stefanie Jegelka, David Lopez-Paz, Kartik Ahuja, Mark Ibrahim, and Mohammad Pezeshki. Reasoncache: Teaching llms to reason without weight updates, 2026. URL https://arxiv.org/abs/2602.02366.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

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

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models, 2025. URL https://arxiv.org/abs/2406.01506.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1286–1304. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.93. URL http://dx.doi.org/10.18653/v1/2022.findings-emnlp.93.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019. URL https://arxiv.org/abs/1905.05950.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality, 2021. URL https://arxiv.org/abs/2109.04404.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models, 2023. URL https://arxiv.org/abs/2302.00294.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice, 2022. URL https://arxiv.org/abs/2203.05962.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A TABLE OF CONTENTS

B A Note on Methodology 12
C List of Symbols 12
D Model Roster 12
E Cosine Similarity and Linear Probe Accuracy 13
F δ versus Context 14
G M versus W_U as readout Operator 14
H Where the Sense Information is Represented – Full Results 17
I Equation 1 Tracks across Layers 17
J ρ Concentrates at the Final Layer 17
K Normalization Conventions for the Concentration Ratio 18
L Logit Movement Ablations 19
M Depth-wise Isotropy 20
N Ratio r is Not $||\delta||$ Rescaled 21
O k-Sweep 22
P Experimental setup 23
Q Per Word Commitment Depth 24

B A NOTE ON METHODOLOGY

Every quantity in this paper is reported against an exact analytic null, not a convention or a threshold. The fraction of a state's energy in a k-dimensional readout subspace is compared against k/d. The readout's response to a direction is compared against $\|M\|_F^2/d$, its exact expectation under a uniformly random unit direction, and the alignment of a state with the contrast direction against $1/\sqrt{d}$. This discipline is what allows the results below to be stated as universal across a heterogeneous roster of open models rather than merely as trends: a claim that holds at $6\times$ chance in one model and $9\times$ in another is one claim, whereas two raw values differing by an order of magnitude are not obviously comparable at all.

C LIST OF SYMBOLS

D MODEL ROSTER

Sixteen open-weight decoder-only models, listed in Table 5 with the architectural and spectral properties the analyses depend on. Models were selected to vary the factors that could plausibly confound a claim stated across models, rather than to sample any particular family. The roster spans 124M to 9.24B parameters, hidden dimensions 768 to 4096, depths of 12 to 42 blocks, six architecture families, both normalization schemes (four LayerNorm, twelve RMSNorm), and both weight-tying regimes (eight tied, eight untied). Weight tying is reported because it determines whether W_U also fixes the input embedding, and therefore what S_{out} means relative to h_0; no analysis in this paper conditions on it. All models are used in evaluation mode with no sampling, and the spectral quantities are computed from the weights alone.

The spectra differ enormously across the roster. Participation ratio spans 1 to 588 and the mass captured by $k = 32$ spans 7.6% to 93.2%. This is the variation that motivates reporting every k-dependent quantity beside its chance level k/d rather than comparing raw values.

Table 4: Symbols used throughout. Chance levels are the value each quantity takes under an explicit null rather than a chosen threshold; see Appendix B.

symbol	first use	meaning
Model and states		
L	Sec. 3	number of transformer blocks
d	Sec. 3	residual-stream dimension
l	Sec. 3	layer index, $l \in \{0, \dots, L\}$, with $l = 0$ the embedding
h_l	Sec. 3	hidden state of the target token after block l
c	Sec. 3	context length: number of preceding disambiguating tokens
$\mathcal{C}$	Sec. 3	set of context conditions, $\{0, 1, 2, 3, 5, 8, 13, \text{full}\}$
W_E, W_U	Sec. 5.1	input embedding and unembedding matrices
LN_f	Sec. 5.1	final normalization layer
g, β	Sec. 5.1	learned gain and bias of LN_f
$\sigma(h)$	Sec. 5.1	scalar normalizer of LN_f
P	Sec. 5.1	mean-centering projection $I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$; the identity for RMSNorm
b, b_0	Sec. 5.1	output bias, and the constant $b_0 = W_U\beta + b$
Decomposition 1: shared and contrast		
$\bar{h}_a, \bar{h}_b$	Sec. 4	centroids of the two senses at a given layer and context
μ	Sec. 4	shared component, $\mu = (\bar{h}_a + \bar{h}_b)/2$
δ	Sec. 4	sense contrast, $\delta = \bar{h}_a - \bar{h}_b$
r	Eq. equation 1	magnitude ratio $\|\delta\|/\|\mu\|$
Decomposition 2: readout-sensitive and insensitive		
M	Eq. equation 2	folded readout operator $W_U \operatorname{diag}(g) P$
U, Σ, V	Sec. 5.1	SVD factors of M, with singular values s_j and right singular vectors v_j
k	Eq. S_{out}	rank of the readout subspace; $k = 32$ throughout
S_{out}	Eq. S_{out}	$\mathrm{Span}\{v_1, \dots, v_k\}$, the directions the readout responds to most
$S_{out}^{\perp}$	Sec. 5.1	its orthogonal complement; not in general the kernel of M
$\delta_S, \delta_\perp$	Sec. 5	components of δ inside and outside S_{out}
PR	Sec. 5.2	participation ratio $(\sum_j s_j^2)^2 / \sum_j s_j^4$ of the spectrum of M
k_{50}, k_{90}, k_{99}	App. D	rank needed for that fraction of $\|M\|_F^2$
Measured quantities and their nulls		
ρ_l	Eq. equation 3	concentration ratio: fraction of state energy in S_{out} at layer l. Chance level k/d
enrichment	Sec. 5	fraction of $\|\delta\|^2$ inside S_{out}, divided by k/d. Chance level 1
signed margin	Sec. 4	amount by which a stimulus lies nearer its own sense's leave-one-out centroid than the other's, in cosine distance
coherence	Sec. 4	signed margin divided by its absolute counterpart; 1 if every stimulus is correctly sided, $1/\sqrt{n}$ if directionless
collapse	Table 2	fractional reduction in sense-centroid separation when a component is removed at the final layer
Statistics		
R	Sec. 5	Pearson correlation across models
R_s	Sec. 4	Spearman rank correlation
p	—	p-value; permutation tests also report the smallest value the design can attain

E Cosine Similarity and Linear Probe Accuracy

We motivate this paper with the observation that at the terminal layer of all sixteen models, the cosine angle between two last-layer embeddings for polysemous words of differing senses collapses, yet the embeddings remain separable with a linear probe with high accuracy. Figure 4 visualizes the

Table 5: The full roster. L is the number of blocks and d the residual-stream dimension. *tied* indicates $W_U = W_E^\top$, determined at runtime by testing whether the two tensors are the same object; this agreed with the `tie_word_embeddings` configuration flag for all sixteen models. PR is the participation ratio $(\sum_j s_j^2)^2 / \sum_j s_j^4$ of the spectrum of the folded readout operator M; k_{50}, k_{90} and k_{99} are the ranks needed for those fractions of $\|M\|_F^2$; *mass at* k is the fraction captured by the $k = 32$ used throughout.

model	checkpoint	params	d	L	norm	tied	PR	k_{50}	k_{90}	k_{99}	mass at k
GPT-2	`gpt2`	124M	768	12	Layer	✓	1	1	120	636	86.6%
Pythia 160M	`EleutherAI/pythia-160m`	160M	768	12	Layer	×	1	1	1	579	93.2%
Qwen 0.5B	`Qwen/Qwen2.5-0.5B`	494M	896	24	RMS	✓	19	213	713	868	30.1%
OLMo 1B	`allenai/OLMo-1B-0724-hf`	1.18B	2048	16	Layer	×	588	411	1459	1958	14.2%
Llama 1B	`meta-llama/Llama-3.2-1B`	1.24B	2048	16	RMS	✓	104	626	1621	1972	13.7%
Pythia 1.4B	`EleutherAI/pythia-1.4b`	1.42B	2048	24	Layer	×	138	632	1644	1982	12.4%
Qwen 1.5B	`Qwen/Qwen2.5-1.5B`	1.54B	1536	28	RMS	✓	62	432	1233	1484	17.8%
Gemma 2B	`google/gemma-2-2b`	2.61B	2304	26	RMS	✓	31	412	1732	2202	26.0%
Qwen 3B	`Qwen/Qwen2.5-3B`	3.09B	2048	36	RMS	✓	101	604	1662	1992	14.4%
Llama 3B	`meta-llama/Llama-3.2-3B`	3.21B	3072	28	RMS	✓	103	905	2398	2953	12.9%
Phi 3.5	`microsoft/Phi-3.5-mini-instruct`	3.82B	3072	32	RMS	×	246	737	2318	2952	14.4%
Mistral 7B	`mistralai/Mistral-7B-v0.3`	7.25B	4096	32	RMS	×	52	909	2967	3876	17.6%
OLMo 7B	`allenai/OLMo-2-1124-7B`	7.30B	4096	32	RMS	×	5	63	2813	3906	49.0%
Qwen 7B	`Qwen/Qwen2.5-7B`	7.62B	3584	28	RMS	×	71	1083	2892	3472	16.4%
Llama 8B	`meta-llama/Llama-3.1-8B`	8.03B	4096	32	RMS	×	328	1403	3302	3958	7.6%
Gemma 9B	`google/gemma-2-9b`	9.24B	3584	42	RMS	✓	13	2597	2597	3423	36.6%

cosine similarity trajectory of the embeddings of two senses across layers of GPT2, along with the linear probe accuracy at every layer as a visual representation of this phenomenon. Table 6 gives the cosine similarities between sense centroids for each model across depth, and Figure 3 visualizes the linear probe accuracy remaining high in all models as well, despite the cosine collapse. We observe the same terminal behavior on all sixteen models we study.

Table 6: Inter-Sense Centroid Cosine Similarity across Normalized Model Depth

Model	**0% (Emb)**	**20%**	**40%**	**60%**	**80%**	**100% (Final)**
gemma-2-2b	1.000	0.887	0.868	0.876	0.900	0.926
gemma-2-9b	1.000	0.880	0.874	0.944	0.936	0.873
gpt2	0.955	0.981	0.968	0.975	0.982	0.998
Llama-3.2-1B	1.000	0.850	0.784	0.857	0.879	0.867
Llama-3.2-3B	1.000	0.792	0.757	0.837	0.854	0.839
Llama-3.1-8B	1.000	0.771	0.765	0.844	0.870	0.828
Mistral-7B-v0.3	1.000	0.831	0.811	0.853	0.822	0.809
OLMo-1B-0724-hf	1.000	0.887	0.821	0.841	0.843	0.781
OLMo-2-1124-7B	1.000	0.804	0.764	0.815	0.777	0.819
Phi-3.5-mini-instruct	1.000	0.938	0.910	0.934	0.942	0.963
pythia-160m	1.000	0.966	0.933	0.963	0.965	0.997
pythia-1.4b	1.000	0.943	0.927	0.962	0.974	0.861
Qwen2.5-0.5B	1.000	0.927	0.914	0.912	0.923	0.919
Qwen2.5-1.5B	1.000	0.900	0.854	0.872	0.905	0.936
Qwen2.5-3B	1.000	0.914	0.836	0.837	0.862	0.868
Qwen2.5-7B	1.000	1.000	1.000	1.000	0.999	0.834

F δ VERSUS CONTEXT

As an example of the sense distributions and separation over context, we include Figure 5 as additional visual evidence supporting Table 7, and Figure 2.

G M VERSUS W_U AS READOUT OPERATOR

We provide two main reasons behind our use of M in place of W_U. First, the gain is not a uniform rescale that would cancel out of any ratio. It re-weights individual coordinates by factors varying

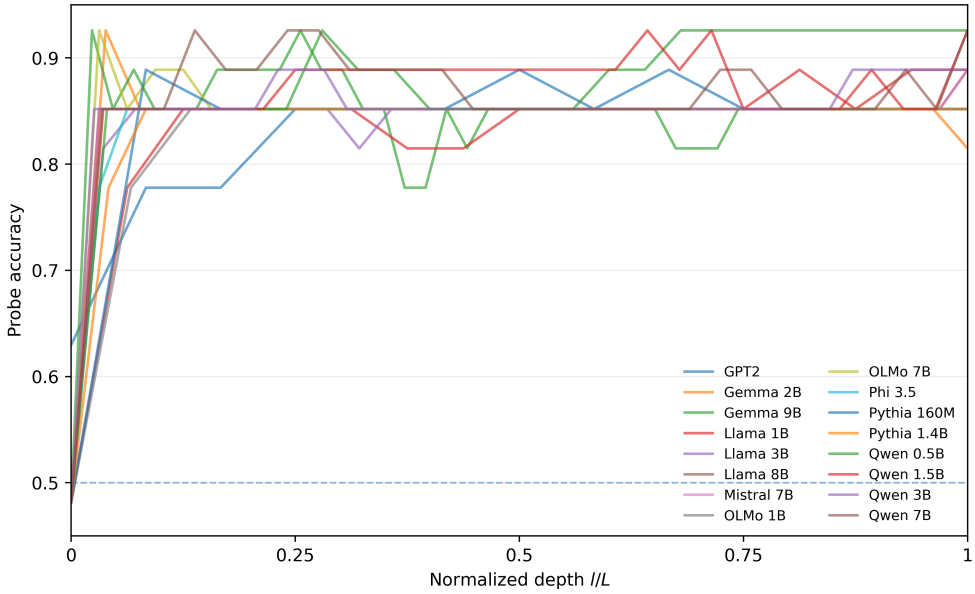

Figure 3: Linear probe accuracy trained to classify sense from the hidden state at a given layer. Across all models, the average linear probe accuracy of our word bank remains high. Layer depth is normalized such that we can visualize all models together, so the x-axis is the percentage of depth through the total number of layers for a specific model.

Table 7: Centroid separation against context at the final layer. $r = \|\delta\|/\|\mu\|$.

model	$c = 1$	$c = 2$	$c = 3$	$c = 5$	$c = 8$	$c = 13$
Gemma 2B	0.425	0.429	0.498	0.581	0.613	0.795
Gemma 9B	0.433	0.519	0.618	0.669	0.712	0.757
GPT2	0.113	0.112	0.143	0.161	0.207	0.236
Llama 1B	0.290	0.502	0.533	0.691	0.819	0.870
Llama 3B	0.342	0.606	0.628	0.774	0.904	0.954
Mistral	0.432	0.656	0.743	0.903	0.976	1.014
OLMo 1B	0.445	0.728	0.847	1.081	1.231	1.432
OLMo 7B	0.409	0.641	0.671	0.763	0.840	0.866
Phi	0.115	0.197	0.264	0.305	0.365	0.409
Pythia 1.4B	0.369	0.570	0.667	0.825	0.967	0.981
Pythia 160M	0.043	0.064	0.074	0.096	0.114	0.145
Qwen 0.5B	0.300	0.451	0.481	0.566	0.581	0.627
Qwen 1.5B	0.246	0.362	0.367	0.427	0.449	0.553
Qwen 3B	0.305	0.408	0.435	0.463	0.486	0.502
Qwen 7B	0.405	0.515	0.567	0.612	0.678	0.725

over an order of magnitude within a single model, so the directions W_U responds to most and the directions the model as a whole responds to most are different sets. This is not only a matter of correctness. Building S_{out} from W_U instead of M gives a measurably different subspace. The mean squared cosine between the two top-32 right singular subspaces is 0.77, against a chance level of 0.019, with a median principal angle of $19°$. The enrichment of δ reported in Section 5 averages $9.6\times$ against M and $5.7\times$ against W_U. Full results are in Table 8. Second, LN_f is applied at exactly one layer, so measuring anything at intermediate layers otherwise requires choosing a convention for the rest between using the raw states, or the norm applied everywhere as in the logit lens nostalgebraist (2020). Folding the norm into M removes the choice since σ is a scalar, every quantity below is a ratio of squared norms, so all of them may be computed on raw hidden states at any layer and still mean the fraction of the state lying in directions the readout responds to.

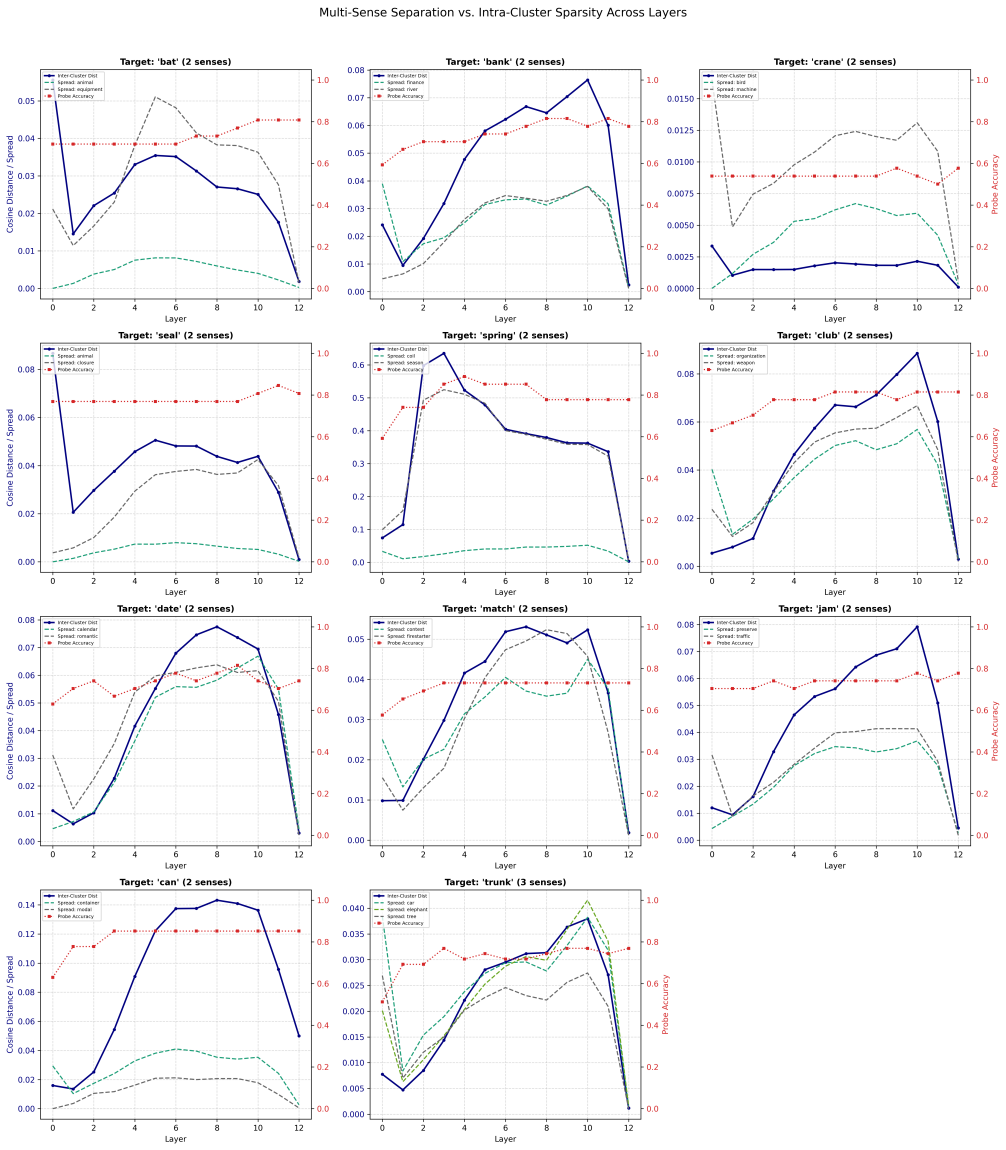

Figure 4: The cosine collapse and linear separability, measured side by side on GPT-2, one panel per word. *Blue, left axis:* cosine distance between the sense centroids. *Dashed, left axis:* the within-sense spread of stimuli about each centroid, one line per sense. *Red dotted, right axis:* accuracy of a linear probe trained to classify sense from the hidden state at that layer, against a chance level of 0.5 for two-sense words and 0.33 for *trunk*. The figure shows the two halves of the motivating observation in one place. Centroid separation rises through the network, peaks in the upper-middle layers, and then falls to near zero at layer 12 in every word (cosine collapse). The within-sense spread collapses with it, and in several words tracks the separation closely throughout. Probe accuracy is flat or slightly rising across the same range and does not fall at the final layer. Terminal collapse does not affect what the probe reads. Section 5.3 accounts for this as we find that the collapse is governed by the magnitude ratio $r = \|\delta\|/\|\mu\|$ and is insensitive to the direction of δ, which is what a linear probe recovers.

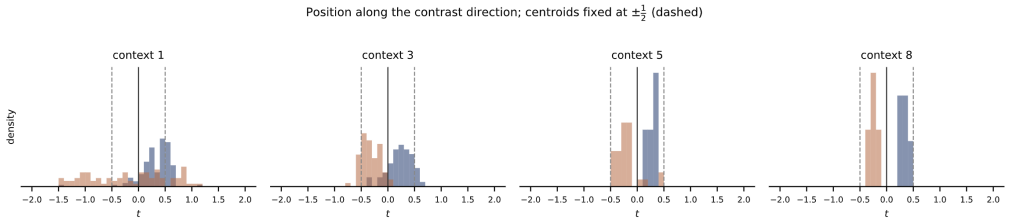

Figure 5: Senses separate at increasing contexts.

Table 8: What changes if the readout subspace is built from W_U instead of the folded operator $M = W_U \operatorname{diag}(g)P$. *overlap* is the mean squared cosine of the principal angles between the two top-32 right singular subspaces: 1 if they coincide, k/d (the *chance* column) for two random 32-dimensional subspaces of $\mathbb{R}^d$. *median angle* is the median principal angle in degrees. The last three columns give δ's enrichment at the final layer measured against each subspace, and their ratio.

model	overlap	chance	median angle	enrich (M)	enrich (W_U)	ratio
gemma$_2b$	0.715	0.0139	22°	6.7	2.8	2.38
gpt2	0.711	0.0417	25°	9.8	1.8	5.35
llama$_1B$	0.764	0.0156	20°	5.7	6.1	0.93
llama$_3B$	0.760	0.0104	22°	8.9	9.4	0.94
minstral	0.716	0.0078	24°	26.2	11.9	2.20
olmo$_7B$	0.870	0.0078	13°	13.5	10.7	1.26
phi	0.955	0.0104	4°	7.6	7.6	1.00
pythia$_1.4B$	0.942	0.0156	6°	6.6	6.6	0.99
pythia$_160M$	0.957	0.0417	6°	5.3	5.1	1.04
qwen$_0.5B$	0.687	0.0357	26°	2.9	0.9	3.22
qwen$_1.5B$	0.635	0.0208	28°	8.2	2.4	3.46
qwen$_3B$	0.694	0.0156	24°	9.0	2.4	3.75
qwen$_7B$	0.642	0.0089	32°	14.7	7.0	2.11
mean	0.773	0.0189	19°	9.6	5.7	2.20

We provide the full table comparing M to W_U, as referenced in Section 5.1.

H Where the Sense Information is Represented – Full Results

Table 9 gives the full results introduced and analyzed in Section 5.3. Figure 6 visualizes these same results.

I Equation 1 Tracks across Layers

Table I evaluates Equation 1 at every layer and compares it with the sense separation measured at that layer. The Spearman is reported over the models depth (layers) with median 0.926 across all models.

J ρ Concentrates at the Final Layer

Across all models, ρ_l increases sharply and abruptly only at the final layer after staying at chance throughout all previous layers, with two exceptions to this patters. First, OLMo 7B experiences the increase in ρ over the final 3–4 layers (the first 28–29 layers remain at chance level). Second, some models have ρ above chance at the 0th layer, the embedding layer. We attribute this to weight tying. For these models, the embedding matrix W_E and the unembedding matrix W_U are tied such that $W_E^\top = W_U$, causing some amount of concentration to be seen at the initial embedding step. In all of these models ρ decreases to chance level at the first layer.

Table 9: Cosine collapse accounted for by the magnitude ratio alone. $r = \|\delta\|/\|\mu\|$ is measured at the final layer and the predicted column is $(r^2/2)/(1 + r^2/4)$, which has no fitted parameters. Both columns are averaged over words after the per-word value is formed, since the expression is nonlinear in r. *ratio* is measured over predicted; values at or just below 1 indicate that δ is close to orthogonal to μ, the condition under which the expression is exact.

model	r	predicted	measured	ratio
Gemma 2B	0.297	0.0465	0.0450	0.97
Gemma 9B	0.343	0.0608	0.0600	0.99
GPT-2	0.057	0.0040	0.0040	1.00
Llama 1B	0.465	0.1087	0.1076	0.99
Llama 3B	0.568	0.1569	0.1556	0.99
Llama 8B	0.740	0.2507	0.2469	0.99
Mistral 7B	0.592	0.1764	0.1727	0.98
OLMo 1B	0.752	0.2540	0.2508	0.99
OLMo 7B	0.534	0.1400	0.1393	1.00
Phi 3.5	0.230	0.0277	0.0276	0.99
Pythia 160M	0.043	0.0010	0.0010	1.00
Pythia 1.4B	0.569	0.1590	0.1576	0.99
Qwen 0.5B	0.328	0.0585	0.0577	0.99
Qwen 1.5B	0.209	0.0243	0.0241	0.99
Qwen 3B	0.324	0.0596	0.0585	0.98
Qwen 7B	0.385	0.0781	0.0764	0.98
mean	0.402	0.0991	0.0965	0.99

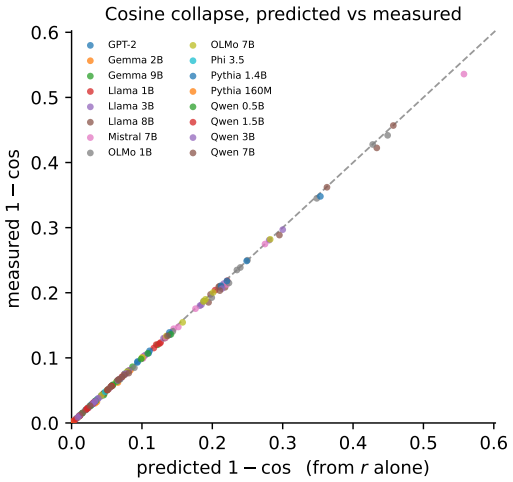

Figure 6: Measured against predicted centroid cosine distance, one point per (model, word) pair. The dashed line is equality. The prediction uses only $r = \|\delta\|/\|\mu\|$ and has no fitted parameters.

K Normalization Conventions for the Concentration Ratio

Section 5.1 argues that folding the final norm into the readout operator removes the need to choose a normalization convention for intermediate layers, because the remaining factor is a scalar and ρ_l is a ratio of squared norms. Table 3 therefore reports ρ_l on raw hidden states. For completeness, Table 11 repeats the measurement under two alternatives: *all*, in which LN_f is applied at every layer, as in the logit lens; and *final*, in which it is applied only at the last layer. The latter puts a preprocessing discontinuity at exactly the layer under test and is reported only for comparison.

The three conventions agree on the location of the terminal step for every model. They disagree in one case on whether that step is an outlier against the model's own mid-network variability: GPT-2 under the *all* convention, where applying LN_f at every layer inflates mid-network steps enough that

Table 10: Equation 1 evaluated at every layer. r is computed per (word, layer) and the prediction $(r^2/2)/(1 + r^2/4)$ is compared against the measured sense separation at the same layer, with no fitted quantity. *Spearman* is over the model's own depth profile.

model	layers	Spearman
Qwen 1.5B	24	$+0.996$
Llama 3B	24	$+0.982$
Pythia 1.4B	21	$+0.982$
Qwen 0.5B	21	$+0.977$
Gemma 9B	36	$+0.963$
Llama 8B	28	$+0.962$
Phi 3.5	28	$+0.961$
Pythia 160M	11	$+0.927$
Llama 1B	14	$+0.925$
OLMo 7B	28	$+0.920$
Qwen 3B	31	$+0.882$
Mistral 7B	28	$+0.831$
Gemma 2B	23	$+0.817$
OLMo 1B	14	$+0.648$
GPT-2	11	$+0.545$
Qwen 7B	24	-0.304
median		$+0.926$

the terminal one no longer clears the 95th percentile. This is the disagreement the folded operator is designed to avoid having to adjudicate.

Table 11: Concentration ratio under three normalization conventions. *rise* is ρ_L minus its early-to-mid-network mean and *step* the layer of the largest single-layer increase. *outlier* indicates that the terminal step exceeds the 95th percentile of that model's mid-network steps. The main text reports the *none* columns.

	none			all			final		
model	rise	step	outlier	rise	step	outlier	rise	step	outlier
Gemma 2B	0.041	26	✓	0.044	25	✓	0.052	26	✓
Gemma 9B	0.022	42	✓	0.036	41	✓	0.035	42	✓
GPT-2	0.551	12	✓	0.196	10	×	0.507	12	✓
Llama 1B	0.330	16	✓	0.281	16	✓	0.305	16	✓
Llama 3B	0.243	28	✓	0.211	28	✓	0.224	28	✓
Llama 8B	0.076	32	✓	0.061	32	✓	0.062	32	✓
Mistral 7B	0.343	32	✓	0.406	32	✓	0.442	32	✓
OLMo 1B	0.218	16	✓	0.200	16	✓	0.218	16	✓
OLMo 7B	0.165	32	✓	0.244	32	✓	0.245	32	✓
Phi 3.5	0.781	32	✓	0.757	32	✓	0.767	32	✓
Pythia 160M	0.968	12	✓	0.740	12	✓	0.971	12	✓
Pythia 1.4B	0.129	24	✓	0.109	24	✓	0.135	24	✓
Qwen 0.5B	0.149	24	✓	0.124	24	✓	0.175	24	✓
Qwen 1.5B	0.273	28	✓	0.261	28	✓	0.354	28	✓
Qwen 3B	0.142	36	✓	0.123	36	✓	0.191	36	✓
Qwen 7B	0.128	28	✓	0.141	28	✓	0.181	28	✓

L Logit Movement Ablations

Section 5 measures the causal role of δ's S_{out} component by how much removing it collapses the separation between the two sense centroids. An alternative is to measure how much it moves the logits, per unit of state movement. Table 12 reports that version.

We prefer the collapse measure in the main text for two reasons. The first is that logit movement is not sense-specific: it registers any change to the prediction, whether or not what changed is the sense

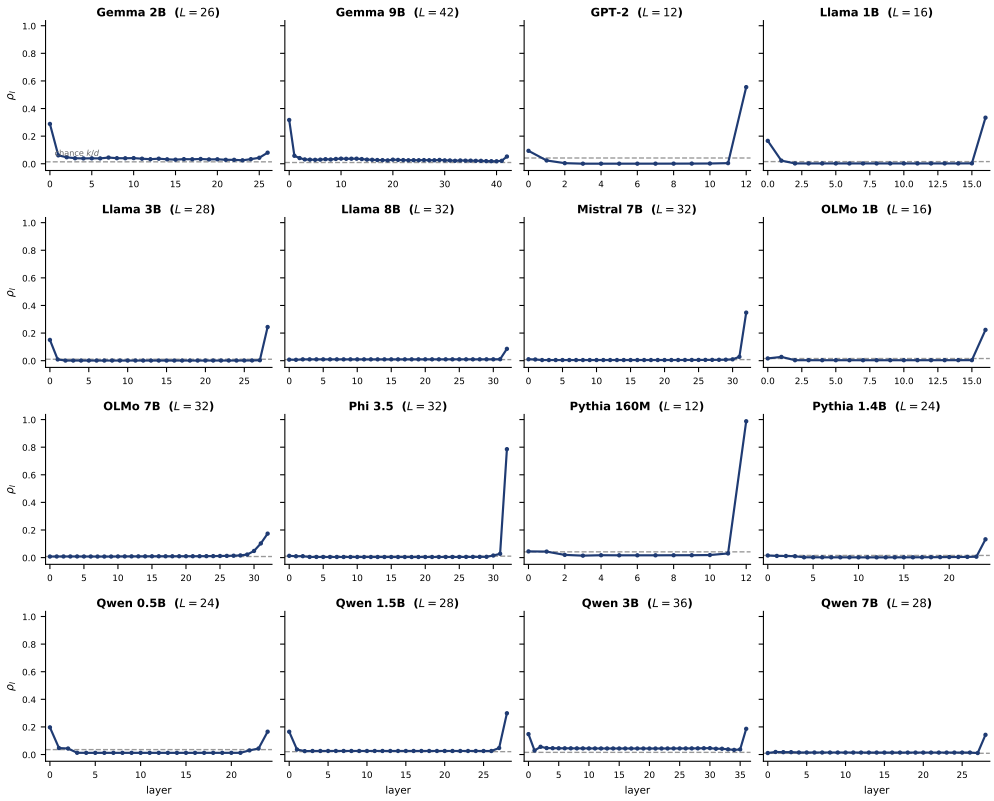

Figure 7: Concentration ratio ρ_l against layer, one panel per model, measured on raw hidden states against S_{out} built from the folded readout operator M. The dashed line is the chance level k/d, the value ρ_l takes for a state with no preferential relationship to the readout geometry. The x-axis is the percent of total layer depth. The curves track that line for most of the network and leave it once, at the end.

distinction. The second is that the natural-looking comparison is close to tautological. A direction matched to δ on its S_{out} *fraction* is only about 13% inside S_{out}, whereas δ's S_{out} component is entirely inside it, and every direction in S_{out} has a large singular value by construction.

The *rank half* column bears on Section 5.2. Half of δ's energy requires a median of 742 singular directions of M. The sense direction is therefore enriched at the head of the spectrum, as Table 2 shows, and spread thinly across hundreds of directions in the tail. Both facts hold at once, and neither on its own characterizes where δ lives.

M Depth-wise Isotropy

Section 4 measures symmetry breaking by comparing two context conditions at a fixed layer, and makes no claim about how the contrast alignment varies with depth. Table 13 reports the depth-wise measurement, which we do not regard as interpretable, for the following reason.

The natural baseline for a depth-wise comparison is layer 0. But for a model with rotary position encoding, every stimulus of a word shares one token embedding at layer 0, so the two sense centroids coincide and the sense-contrast direction is determined entirely by which tokenization variants happen to appear in the stimuli. The layer-0 values in the table range over a factor of fifteen across models, from 0.023 to 0.347, while the final-layer values all lie between 0.03 and 0.66 and, excluding GPT-2, between 0.11 and 0.25. The models agree about where they end and disagree about where the measurement says they start, which is the signature of a degenerate baseline rather than an architectural difference.

Table 12: Logit movement per unit of state movement when a component is removed at the final layer. δ_S and $\delta_\perp$ are δ's components inside and outside S_{out}; *rand (matched)* shares δ's S_{out} fraction; *rand in S_{out}* is a unit direction drawn from within S_{out} and is the appropriate control. *rank half* is the number of singular directions of M needed to contain half of δ.

model	δ_S	$\delta_\perp$	rand (matched)	rand in S_{out}	rank half
Gemma 2B	4.513	0.382	0.671	1.352	2,116
Gemma 9B	3.388	0.327	0.423	1.190	1,846
GPT-2	0.981	0.014	0.035	0.079	754
Llama 1B	1.250	0.490	0.560	1.056	730
Llama 3B	1.727	0.696	0.805	1.711	940
Llama 8B	2.942	0.825	1.039	2.177	2,916
Mistral 7B	1.730	0.618	0.973	1.781	429
OLMo 1B	1.672	0.574	0.752	1.635	483
OLMo 7B	0.507	0.119	0.211	0.511	1,170
Phi 3.5	0.449	0.124	0.127	0.320	1,159
Pythia 160M	0.060	0.013	0.021	0.043	219
Pythia 1.4B	1.811	0.956	1.104	1.894	649
Qwen 0.5B	2.945	0.862	1.016	1.715	499
Qwen 1.5B	1.570	0.713	0.775	1.097	228
Qwen 3B	1.804	0.770	0.894	1.484	400
Qwen 7B	1.973	0.232	0.324	0.622	2,184
mean	1.833	0.482	0.608	1.167	1,045

Table 13: Depth-wise contrast alignment. *chance* is $1/\sqrt{d}$, the value $|\hat{\delta} \cdot \hat{h}|$ takes for a direction unrelated to the state.

model	chance	layer 0	min	min depth	final	min / chance
Gemma 2B	0.021	0.052	0.142	0.15	0.206	6.8
Gemma 9B	0.017	0.044	0.133	0.43	0.211	8.0
GPT-2	0.036	0.223	0.141	0.17	0.655	3.9
Llama 1B	0.022	0.212	0.190	0.19	0.195	8.6
Llama 3B	0.018	0.278	0.215	0.18	0.217	11.9
Llama 8B	0.016	0.347	0.218	0.50	0.246	14.0
Mistral 7B	0.016	0.227	0.223	0.59	0.227	14.3
OLMo 1B	0.022	0.066	0.207	0.19	0.239	9.4
OLMo 7B	0.016	0.023	0.218	0.16	0.205	14.0
Phi 3.5	0.018	0.260	0.158	0.56	0.114	8.8
Pythia 160M	0.036	0.129	0.137	0.25	0.034	3.8
Pythia 1.4B	0.022	0.095	0.176	0.17	0.216	8.0
Qwen 0.5B	0.033	0.150	0.174	0.17	0.213	5.2
Qwen 1.5B	0.026	0.139	0.202	0.57	0.191	7.9
Qwen 3B	0.022	0.118	0.161	0.28	0.209	7.3
Qwen 7B	0.017	0.237	0.752	0.18	0.206	45.0
mean	0.023	0.162	0.213	0.30	0.224	11.0

N RATIO r IS NOT $||\delta||$ RESCALED

Table 14 confirms that the magnitude ratio r is not merely measurement of δ, but that it does indeed track its own measurement. We also visualize this information in Figure 8.

We also note that the terminal fall in r has more than one cause. r falls at the final layer in eleven of sixteen models, but not by a single route. In seven (GPT-2, both small Llamas, both OLMos, Pythia-160M and Qwen-0.5B) $\|\mu\|$ grows faster than $\|\delta\|$, which is the straightforward reading. In three (Phi-3.5, Qwen-3B and Qwen-7B) both norms shrink and $\|\delta\|$ shrinks more. In Qwen-1.5B $\|\delta\|$ falls while $\|\mu\|$ holds. And in four models $\|\mu\|$ contracts at the final layer rather than growing. What is common across the roster is the behavior of the ratio; the norms that produce it are not. We therefore state the terminal result as a fact about r, and do not attribute it to norm growth in the shared component.

Table 14: Whether the magnitude ratio adds anything over $\|\delta\|$. The first three columns describe how each quantity moves from its mid-network median (depth 0.2–0.8) to the final layer. The last three give the Spearman correlation between the variance in the measured separation explained by the parameter-free prediction from r, and by a one-parameter fit to $\|\delta\|^2$. The $\|\delta\|$ account is given a free scale and the r account none, so it can only lose to the extent that $\|\mu\|$ varies across depth.

model	$\|\mu\|$ at L	$\|\delta\|$ at L	r at L	corr$(r, \|\delta\|)$	mid-band	$\|\mu\|$ var (mid)
Gemma 2B	1.06	1.09	1.03	0.246	-0.527	2.26
Gemma 9B	0.56	0.79	1.42	-0.068	-0.689	2.97
GPT2	2.04	0.45	0.22	0.844	0.812	1.78
Llama 1B	8.02	2.88	0.36	0.365	-0.828	1.52
Llama 3B	4.56	2.15	0.47	0.428	-0.897	2.23
Mistral	60.36	61.90	1.03	0.036	-0.550	10.32
OLMo 1B	3.72	2.79	0.75	0.612	0.193	1.41
OLMo 7B	6.29	5.14	0.82	0.371	-0.566	3.32
Phi	0.63	0.18	0.28	0.457	-0.865	4.76
Pythia 1.4B	1.21	1.26	1.05	0.429	-0.909	2.97
Pythia 160M	13.88	1.76	0.13	0.215	0.113	2.12
Qwen 0.5B	8.79	3.28	0.37	0.433	-0.512	1.80
Qwen 1.5B	1.15	0.27	0.24	0.824	-0.744	1.37
Qwen 3B	0.61	0.21	0.35	0.949	0.112	1.27
Qwen 7B	0.07	0.04	0.54	0.943	-0.698	1.09
median	2.88	1.51	0.42	0.431	-0.558	1.96

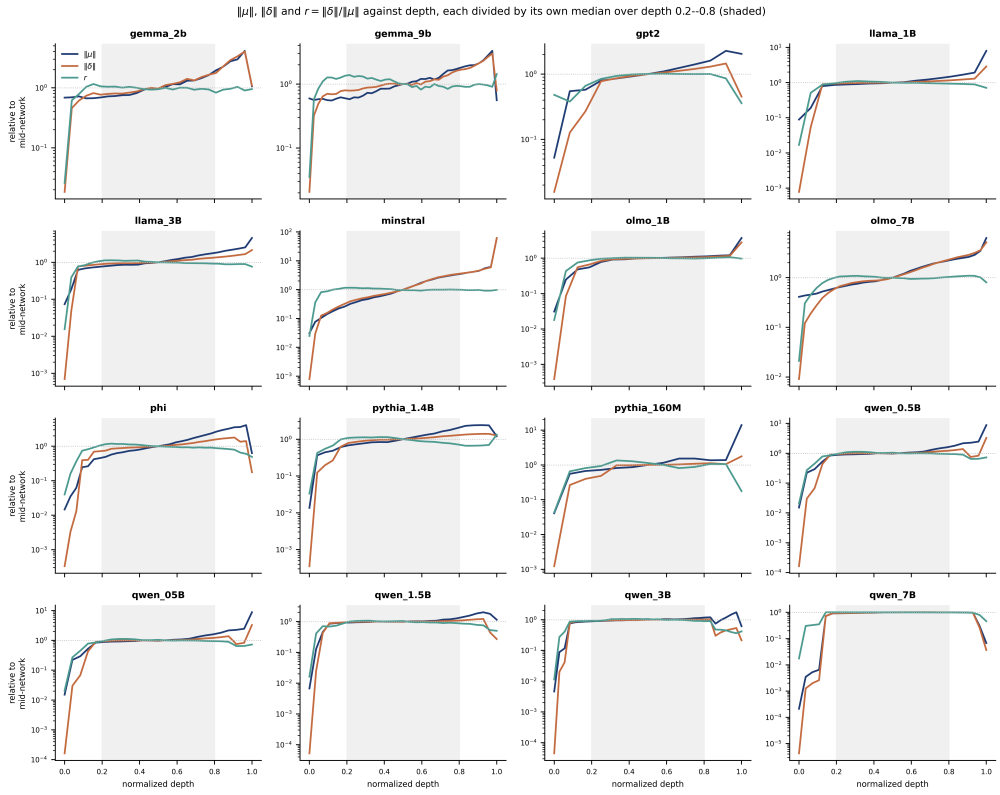

Figure 8: μ, δ, and r across depth.

O k-SWEEP

Appendix O repeats the measurement at $k = 8$ and $k = 128$. The sense contrast is enriched above chance at every rank in every model, so the finding is a property of the readout subspace and not

Table 15: Enrichment of the sense contrast in S_{out} at three subspace ranks. *frac* is the fraction of $\|\delta\|^2$ lying inside S_{out} and *enrich* divides it by the chance level k/d, which changes with k. Enrichment must approach 1 as $k \to d$, so the sweep is read for two things: whether δ stays above chance at every rank, and how steeply the enrichment falls. A fall steeper than the roughly flat profile a uniformly spread δ would give indicates that δ sits preferentially at the head of the spectrum. Three models are absent pending reruns.

	enrichment over chance		
model	$k = 8$	$k = 32$	$k = 128$
GPT-2	34.9	9.8	2.7
Pythia 160M	9.8	5.3	2.4
Qwen 0.5B	2.2	2.9	1.4
Llama 1B	7.4	5.7	3.3
Pythia 1.4B	6.3	6.6	3.6
Qwen 1.5B	9.4	8.2	3.7
Gemma 2B	15.7	6.7	2.5
Qwen 3B	10.9	9.0	3.9
Llama 3B	9.0	8.9	4.8
Phi 3.5	10.8	7.6	3.6
Mistral 7B	34.3	26.2	9.8
OLMo 7B	21.5	13.5	6.7
Qwen 7B	38.7	14.7	5.9
mean	16.2	9.6	4.2
minimum	2.2	2.9	1.4

of the rank we chose. Enrichment falls as k grows ($16.2\times$, $9.6\times$ and $4.2\times$ on average) which is expected, since enrichment must approach 1 as k approaches d, but the rate of the fall is informative because it is steeper than proportional, meaning δ sits preferentially at the head of the spectrum rather than being spread evenly through S_{out}.

P Experimental setup

P.1 Models

Sixteen open-weight decoder-only models (Appendix D), listed with their architectural properties in Table 5: two Gemma-2, GPT-2, three Llama-3.x, Mistral-7B-v0.3, two OLMo, Phi-3.5-mini-instruct, two Pythia and four Qwen 2.5. They span 160M to 9B parameters, hidden dimensions 768 to 4096, depths 12 to 42 blocks, LayerNorm and RMSNorm, and both weight-tying regimes with eight models in each group. All measurements use the model in evaluation mode with no sampling.

P.2 Stimulus bank

Our bank of sentences covers eleven polysemous English words: *bank*, *bat*, *can*, *club*, *crane*, *date*, *jam*, *match*, *seal*, *spring* and *trunk*. Senses are semantic rather than syntactic such that the two readings are not distinguishable by part of speech alone. Each sentence contains exactly one occurrence of the target word in a context that forces a single reading, with roughly ten sentences per sense, giving 226 single-sense stimuli. Sentences are written so that the disambiguating evidence precedes the target, since the models are causal and only preceding tokens are visible at the target position. We use a sentence only where the target occurs exactly once, so its index is unambiguous.

Context. For a target at index t, the condition c runs the model on $\texttt{tokens}[t - c : t + 1]$ and reads the final position. The target is therefore present in every condition and only the quantity of preceding context differs. The conditions are nested, meaning that the evidence available at $c = 3$ is a subset of that at $c = 8$, isolating the effect of context from any difference between sentences.

Sense centroids. Within each context condition, a sense centroid is the mean hidden state over that sense's stimuli, computed leave-one-out so that no stimulus is ever scored against a centroid

it helped define. The centroids are recomputed at every layer and every context condition, so all comparisons are position-matched.

P.3 THE READOUT SUBSPACE

S_{out} is the top-k right singular subspace of the folded readout operator $M = W_U \operatorname{diag}(g) P$ of Section 5.1, computed once per model from the weights alone. We use $k = 32$ throughout, except for the transport fits of Section **??**, where identifiability requires $k = 16$, below the smallest design rank in the roster. Every quantity that depends on k is reported beside its chance level, k/d for energy fractions and $\|M\|_F^2/d$ for readout response, so that no claim rests on the choice; Section 5.2 gives the spectrum and explains why no single k is canonical.

Two implementation details matter for replication. The final norm's gain is read from the module and its convention detected from the module class, because Gemma's RMSNorm computes $x \odot (1 + w)$ with w centered near zero, so using the stored weight directly applies a gain of approximately zero to every coordinate. And P is the mean-centering projection for LayerNorm and the identity for RMSNorm.

P.4 MEASUREMENTS

Concentration ratio (Section 6). ρ_l is computed on raw hidden states at every token position of the stimulus set, not only at the target tokens, with positions subsampled to $1{,}500$ where a model's stimuli exceed that. Measuring at eleven target token types instead conflates the concentration event with the identity of those types. The verdict rule (largest single-layer increase in the final two layers, exceeding the 95th percentile of increases over layers 3 to $L-2$) is fixed in the analysis code and the search begins at layer 3 because the embedding-to-block transition otherwise wins the argmax in several models.

Symmetry breaking (Section 4). Sense centroids are computed within each context condition using leave-one-out, so no stimulus is scored against a centroid it helped define. The contrast direction for the stabilizer measurement is cross-fitted: estimated on half of a word's stimuli and evaluated only on the other half. Without cross-fitting the context ratios average 18% higher, because the full-context condition is the one the directions are estimated from.

Alignment and ablation (Section 5). δ is the difference of sense centroids and μ their pooled mean, both at the layer under test. Ablations remove a component by projecting it out of h_L and recomputing Mh_L/σ directly, which is exact at the final layer because the readout is the entire remaining computation. The control for δ's S_{out} component is a random unit direction drawn from *within* S_{out}, not one matched on S_{out} fraction: the latter is only about 13% inside the subspace while δ_S is entirely inside it, and every direction in S_{out} has a large singular value by construction, so that comparison follows directly.

Q PER WORD COMMITMENT DEPTH

A token commits at the first layer after which its hidden state is, and remains, closer to its own sense centroid than to the other's. We report the layer as a fraction l/L of each model's depth, since the roster spans 12 to 42 blocks and raw indices are not comparable across it. Commitment depth is averaged over the stimuli that commit; the fraction that do is the rate, and is reported alongside because a low rate means the mean behind it is taken over a small and self-selected subset.

At full context, Table 16 places mean commitment between 0.13 and 0.26 of depth across the eleven words. Senses are therefore resolved in the first quarter of the network in every model and for every word. This is well before the terminal concentration of Section 6, and consistent with that event formatting a distinction that is already present rather than producing one.

Table 17 shows that as we increase context length, the percentage of stimuli committing to a sense increases accordingly.

Table 16: Mean commitment depth by word at full context, as a fraction l/L of each model's depth. Raw layer indices are not comparable across a roster spanning 12 to 42 blocks. Commitment depth is averaged over the stimuli that commit; the fraction that do is given in the *rate* column. Words are ordered by their across-model mean.

	club	*crane*	*date*	*spring*	*jam*	*can*	*trunk*	*seal*	*bat*	*bank*	*match*
Gemma 2B	0.32	0.07	0.16	0.32	0.47	0.18	0.28	0.04	0.13	0.22	0.28
Gemma 9B	0.22	0.04	0.10	0.22	0.25	0.14	0.25	0.14	0.10	0.13	0.16
GPT2	0.24	0.27	0.30	0.20	0.21	0.13	0.21	0.22	0.23	0.14	0.19
Llama 1B	0.15	0.23	0.19	0.12	0.10	0.17	0.11	0.17	0.17	0.09	0.14
Llama 3B	0.15	0.22	0.09	0.10	0.11	0.13	0.08	0.14	0.12	0.09	0.07
Llama 8B	0.21	0.20	0.13	0.14	0.10	0.10	0.08	0.06	0.12	0.09	0.05
Mistral 7B	0.26	0.19	0.13	0.09	0.11	0.15	0.07	0.17	0.12	0.12	0.08
OLMo 1B	0.27	0.25	0.10	0.06	0.15	0.24	0.14	0.11	0.15	0.09	0.07
OLMo 7B	0.25	0.18	0.09	0.09	0.08	0.06	0.08	0.07	0.05	0.08	0.07
Phi 3.5	0.30	0.31	0.11	0.08	0.10	0.12	0.10	0.13	0.11	0.08	0.08
Pythia 160M	0.35	0.40	0.25	0.16	0.19	0.09	0.14	0.27	0.26	0.16	0.13
Pythia 1.4B	0.23	0.30	0.14	0.14	0.12	0.10	0.11	0.13	0.14	0.12	0.04
Qwen 0.5B	0.21	0.14	0.21	0.06	0.13	0.07	0.09	0.10	0.13	0.14	0.07
Qwen 1.5B	0.19	0.07	0.12	0.13	0.11	0.06	0.09	0.21	0.05	0.12	0.11
Qwen 3B	0.32	0.18	0.23	0.11	0.12	0.45	0.09	0.21	0.27	0.10	0.05
Qwen 7B	0.55	0.04	0.44	0.55	0.23	0.25	0.39	0.12	0.04	0.39	0.41
mean	0.26	0.19	0.17	0.16	0.16	0.15	0.15	0.14	0.14	0.14	0.13

Table 17: Percentage of stimuli that commit to a sense, by context length. A stimulus commits when its hidden state becomes, and remains, closer to its own sense centroid than to the other's.

model	0	1	2	3	5	8	13
GPT-2	100	56	56	57	66	81	95
Gemma 2B	100	56	59	56	68	86	95
Gemma 9B	100	54	59	57	66	81	93
Llama 1B	100	51	48	48	59	78	91
Llama 3B	100	51	49	50	57	77	92
Llama 8B	100	50	48	48	59	78	92
Mistral 7B	100	53	52	48	54	72	89
OLMo 1B	100	53	47	49	56	77	91
OLMo 7B	100	49	46	46	57	76	91
Phi 3.5	100	64	49	47	55	66	86
Pythia 160M	100	55	52	51	59	78	91
Pythia 1.4B	100	54	51	48	57	76	90
Qwen 0.5B	100	52	52	50	59	78	91
Qwen 1.5B	100	51	51	49	61	77	91
Qwen 3B	100	53	50	49	60	78	91
Qwen 7B	100	53	51	48	61	79	92
mean	100	53	51	50	60	77	91