BAG OF DIMS: TRAINING-FREE MECHANISTIC INTERPRETABILITY VIA DIMENSION-LEVEL SIGN PATTERNS

Varun Reddy Nalagatla
Amazon Web Services
nalavaru@amazon.com

ABSTRACT

We show that the standard basis of transformer hidden states already provides a training-free, architecture-general feature basis. Individual dimensions encode semantic content via their signs (± 1) and confidence via their magnitudes, functioning as independent binary registers. A feature is simply a subset of dimensions with a consistent sign pattern, readable by counting sign agreements with no learned rotation. We validate this Bag of Dims framework across seven models spanning language (Qwen 3.5-4B, Gemma 3-4B, Mistral 7B, Qwen3-32B), vision (DINOv2, ViT-Base), and audio (AST).

Sign patterns alone carry predictive content: replacing all magnitudes with unity preserves 60–93% top-5 next-token accuracy through the LM head, and pure Hamming scoring with no decoder on either side reaches 80–90% top-4096. These patterns organize into semantic features: from a single-token cache (one forward pass per vocabulary token, no context, no labels), we detect 175 categories at AUC 0.97–0.99 by counting sign agreements. A trained probe adds only $+0.018$ AUC and converges to axis-aligned weights: the rotation that autoencoders and probes learn buys almost nothing. These features are causally operative, not merely readable: they survive the K and V attention projections, trace back to the FFN neuron coalitions that write them (random-weight controls never reproduce this), and are already present—axis-aligned and sign-readable at residual parity—in the FFN's own activation space, surviving the nonlinear SwiGLU gate. Flipping a feature's sign pattern during the live forward pass suppresses its concept across four language models, magnitude-matched and concept-specific. Dimensions stay independent throughout (pairwise MI < 0.006 bits).

Given only random seeds and no labels, discovery scales to 1500 features per language model. And the structure is not specific to language: the same per-dimension signs appear in self-supervised vision (DINOv2, 9/12 ImageNet superclasses), supervised vision (ViT-Base, 11/12), and audio (AST, 50/50 ESC-50 categories), indicating it reflects transformer training in general rather than the language-modeling objective. The standard basis already suffices for feature reading: the only cost is one forward pass, with no optimization and no GPU-days. The open problem shifts from finding the right rotation to cataloging what each dimension encodes.

1 INTRODUCTION

Reading features from transformer hidden states currently requires training a separate model: sparse autoencoders need millions of contextual activations and GPU-hours; probes need labeled datasets per property. This paper presents evidence for a simpler alternative: individual dimensions already encode semantic features, readable by counting sign agreements.

Figure 1 plots all 2560 dimensions of a language model's hidden state across 32 layers for 20 different prompts overlaid. The same structured envelope appears every time, content-independent. The repeating diamond lattice and banding are a population effect of overlaying many oscillating dimensions: they appear in both trained and randomly initialized models, so the lattice itself is not what training produces. What training produces is the *internal* organization within that envelope. Figure 2 makes the contrast explicit across language, vision, and audio, feeding the same real inputs

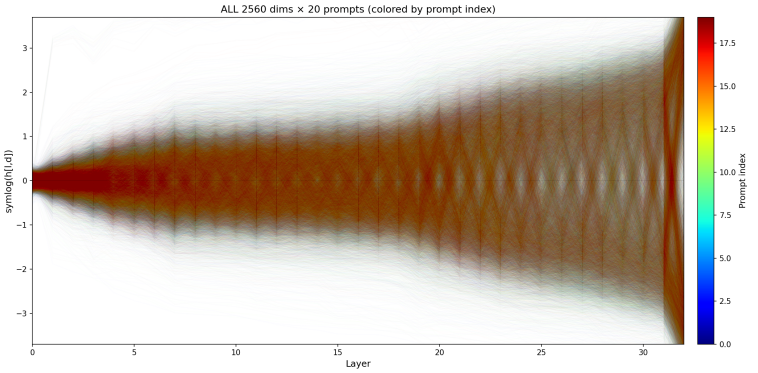

Figure 1: Twenty different prompts overlaid (Qwen 3.5-4B). All 2560 dimensions of the residual stream state $h_l[d]$ (the layer output after layer l) plotted across layers. The same expanding structure appears regardless of input content.

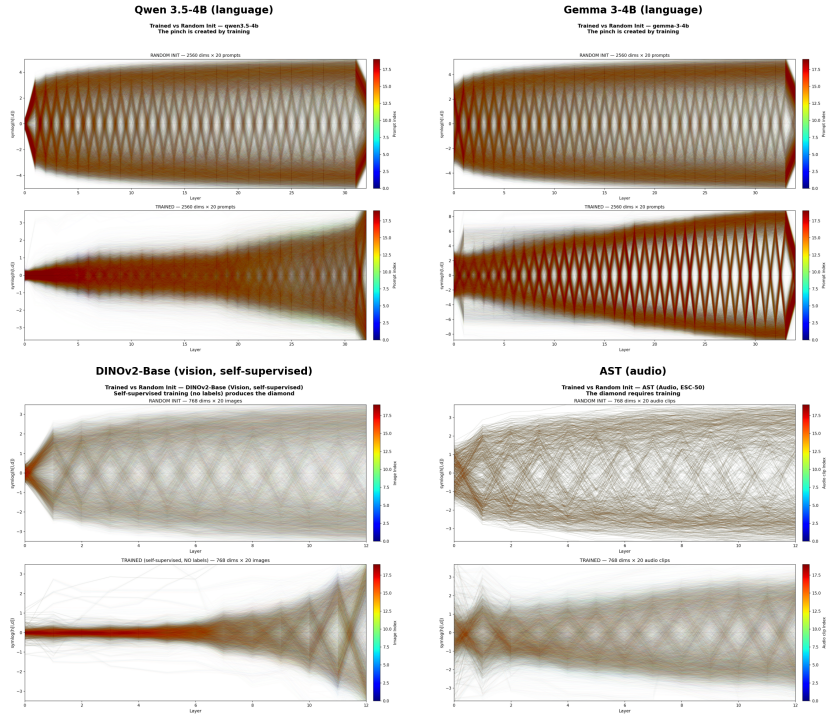

Figure 2: Trained (bottom of each panel) vs. random init (top) across language (Qwen 3.5-4B, Gemma 3-4B), vision (DINOv2-Base, self-supervised), and audio (AST). The expanding envelope is a population artifact; the internal structure (pinch, banding, non-uniform density) is created by training.

through trained and random-init models: under random weights the lattice is symmetric and its density uniform about zero, while training skews it into a characteristic pinch, asymmetric banding, and non-uniform density—regardless of modality or training objective.

Figure 3 shows the microscopic picture: individual dimensions follow independent paths, each consistent across prompts of the same domain but uncorrelated with other dimensions. The population-level envelope is a statistical property of many independent channels, not a property of any single dimension. Cross-dimension mutual information is negligible ($<$0.006 bits; §3.2), and adding context lowers it further rather than introducing entanglement.

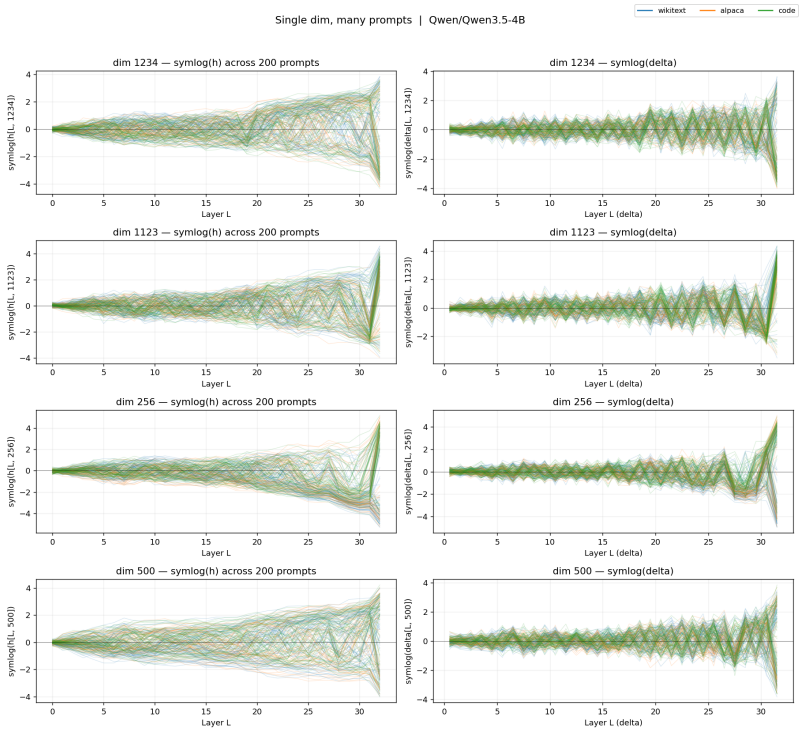

Figure 3: Individual dimension trajectories (4 dims, 200 prompts colored by domain). Each dim follows its own path; the population-level envelope emerges only when 2560 such independent trajectories are overlaid.

This motivates the **Bag of Dims** framework: treating hidden states as collections of independent binary registers, where each dimension encodes content via its sign ($+1$ or -1) and confidence via its magnitude. Features are subsets of dimensions with consistent sign patterns, readable without training. We validate this through progressive experiments:

1. **Sign encodes content, magnitude encodes confidence** (§3.1). Pure sign agreement (Hamming distance, no learned decoder) narrows a 248K vocabulary to the correct 4096 candidates 80–90% of the time across three architectures.
2. **Dimensions are independent** (§3.2). Pairwise mutual information between dimension signs is negligible ($<$0.006 bits), and context only lowers it. An MLP with full cross-dimension capacity adds zero AUC over per-dim reading.
3. **Zero-training feature discovery** (§3.3). From a single-token type cache (one forward pass per vocab token, no context), 175 semantic categories emerge at mean per-dim AUC 0.80; unsupervised discovery scales to 1500 features at 100% yield. A trained probe adds only $+0.018$ AUC and converges to axis-aligned weights.
4. **Features are readable and causally operative in context** (§3.4). Type-level prototypes detect tokens in running text, and read cross-category word sense (a category prototype scores a polysemous word higher in its category-sense context than its other-sense context, 77–80% across 77 cases on three of four models). Going beyond detection, flipping a feature's signs during the live forward pass suppresses its concept across four language models—sign, not magnitude (a magnitude-matched control does nothing), and concept-specific (a disjoint coalition does nothing).
5. **Features survive attention projections** (§3.5). All 175 categories exceed null calibration in both K and V dimensions across four architectures, confirming W_k/W_v preserve axis-aligned structure.
6. **FFN neurons write axis-aligned** (§3.6). Static weight inspection links up to 20% of features to individual writer neurons ($>$0.70 sign agreement; random controls: 0%), and

top-200 neuron coalitions reconstruct 99.9% of prototypes via majority vote. The FFN's own activation space is itself sign-readable at residual parity, surviving the nonlinear SwiGLU gate.

7. **Cross-modality universality** (§3.7). The same method works on DINOv2 (self-supervised vision, 9/12 superclasses), ViT-Base (supervised, 11/12), and AST (audio, 50/50 categories $>$ 0.70), showing the structure emerges from transformer training itself, not from classification objectives.

These results trace a complete read-write circuit in the standard basis: FFN neurons write axis-aligned sign patterns—and their activations are themselves a sign-readable per-dim space $\rightarrow$ the residual stream stores them $\rightarrow$ K/V projections preserve them for attention routing. The read side (residual detection, K/V preservation) operates identically across language, vision, and audio; the write side (FFN specificity) is present in all but sharpened by classification. The pattern is not merely correlational: flipping a feature's signs during inference causally suppresses its concept, confirming the model computes with these patterns rather than just carrying them. None of this requires gradient computation: a forward pass per input builds the cache, and everything else runs on sign matrices.

2 METHOD

2.1 THE BAG-OF-DIMS FRAMEWORK

Each of the D dimensions ($D = 2560$ for a 4B model, $D = 4096$ for Mistral 7B, $D = 5120$ for Qwen3-32B) functions as an independent channel:

- **Sign** ($+1$ or -1): the semantic content, what the dimension is encoding.
- **Magnitude** ($|$value$|$): the confidence, how committed the dimension is.

A **feature** is a subset of dimensions $\mathcal{D} \subseteq \{1, \ldots, D\}$ with a consistent sign pattern $\boldsymbol{\pi} \in \{+1, -1\}^{|\mathcal{D}|}$ across tokens of a given semantic category. For example, if tokens representing animals consistently have dims $\{47, 512, 1893\}$ positive and dims $\{203, 678\}$ negative, that sign pattern *is* the "animal" feature, readable directly from the standard basis without any learned rotation.

Notation. We write h_l for the hidden state after layer l (e.g., h_{24} is the output of layer 24), $h_l[d]$ for its d-th dimension, and $\mathbf{h}$ for the final-layer state when the layer index is clear from context.

2.2 SINGLE-TOKEN TYPE CACHE

Feature discovery requires no contextual data. Type-level semantics—what a token *is*—are already encoded in the sign pattern after a single-token forward pass. With only one position, self-attention has no cross-token context to mix in, so the resulting hidden state reflects the token's identity as processed by the full transformer stack without influence from surrounding tokens.

We construct a **type-level cache** by running every vocabulary token through the model individually (no context):

1. For each of the V tokens in the vocabulary ($\sim$248K for Qwen 3.5-4B, $\sim$262K for Gemma, $\sim$32K for Mistral, $\sim$152K for Qwen3-32B):
   - Feed the token as a single-token input
   - Extract the hidden state at target layers
   - Store the D-dimensional state
2. Result: a matrix $\mathbf{H} \in \mathbb{R}^{V \times D}$ per layer.

We focus on the optimal semantic layer per model: layer 24 for Qwen 3.5-4B (of 32), layer 24 for Mistral (of 32), layer 34 for Gemma (of 34), and layer 48 for Qwen3-32B (of 64). A per-layer sweep confirms these choices maximize category separability (Appendix D); Gemma peaks at the final layer due to its U-shaped layer profile (strong embeddings, middle-layer reorganization, late recovery). Dim assignments are layer-specific (cross-layer Jaccard $= 0.042$); discovery must be performed per-layer.

The cache requires one forward pass per vocabulary token (∼20 minutes on a single GPU) and is computed once per model. Since all subsequent analysis operates only on sign patterns, the cache can be stored as packed bits (1 bit per dimension), reducing storage 32× relative to float32: 93 MB for the full Qwen3-32B vocabulary at one layer. All feature discovery, prototype building, and evaluation in this paper operate on this cache, requiring no sentences, no prompts, and no gradient computation.

2.3 Feature Discovery via Per-Dimension AUC

Given the type-level cache $\mathbf{H} \in \mathbb{R}^{V \times D}$, we discover features through three steps.

Step 1: Anchor tokens. For a category c (*e.g.*, "animal"), select a set of $n_a = 50$ single-token exemplars $\mathcal{A}_c \subset \{1, \ldots, V\}$.

Step 2: Per-dimension AUC. For each dimension d, we compute how well its sign separates category members from the full vocabulary. Let $s_{t,d} = \text{sign}(H_{t,d})$ denote the sign of token t at dimension d. Define:

$$p_d^+ = \frac{1}{|\mathcal{A}_c|} \sum_{t \in \mathcal{A}_c} \mathbf{1}[s_{t,d} = +1] \tag{1}$$

$$p_d^- = \frac{1}{|\bar{\mathcal{A}}_c|} \sum_{t \notin \mathcal{A}_c} \mathbf{1}[s_{t,d} = +1] \tag{2}$$

where $\bar{\mathcal{A}}_c$ is the full vocabulary (negative set). The per-dimension AUC is:

$$\text{AUC}_d = \max\left(\frac{1 + p_d^+ - p_d^-}{2},\ \frac{1 - p_d^+ + p_d^-}{2} \right) \tag{3}$$

The first term measures separability assuming positive polarity ($s_{t,d} = +1$) indicates category membership; the $\max$ selects whichever polarity better separates the category, making the metric invariant to sign convention.

Step 3: Build sign prototype. Register dimensions exceeding a threshold $\tau = 0.75$:

$$\mathcal{D}_c = \{d : \text{AUC}_d \geq \tau\} \tag{4}$$

For each registered dimension, record the expected polarity:

$$\pi_d = \begin{cases} +1 & \text{if } p_d^+ > p_d^- \\ -1 & \text{otherwise} \end{cases} \tag{5}$$

The feature prototype is the pair $(\mathcal{D}_c, \boldsymbol{\pi}_c)$.

Scoring via sign agreement. To classify a new token, count the fraction of registered dimensions where the token's sign matches expected polarity: score $= 1 -$ (Hamming distance$/|\mathcal{D}_c|$). No learned weights appear; the score is a normalized count of sign agreements.

Note on metrics. We report two distinct AUC values throughout: (1) *per-dimension AUC* (Eq. 3) measures how well a single dimension separates a category from the full vocabulary, serving as a discovery metric for selecting which dims belong to a feature. (2) *Prototype-level AUC* applies the composite sign-agreement score across all tokens and computes the standard AUC of this classifier, measuring the detection performance (0.95+ in head-to-head evaluation; Appendix A).

Null calibration. We run the same procedure with 100 random anchor sets of identical size. The p_{95} of the null distribution establishes the threshold a category must exceed to be reported, and the "Exceed null p_{95}" columns throughout report this test. The margin is large enough that the choice of percentile is not load-bearing: null p_{99} ranges only from 0.642 (Qwen) to 0.692 (Mistral), while the weakest real category scores 0.70–0.76, so every reported category clears even the stricter p_{99} bar (§3.3).

2.4 Extension to Attention Projections

The same procedure applies to K and V projections. Instead of reading signs from the residual stream, we read from:

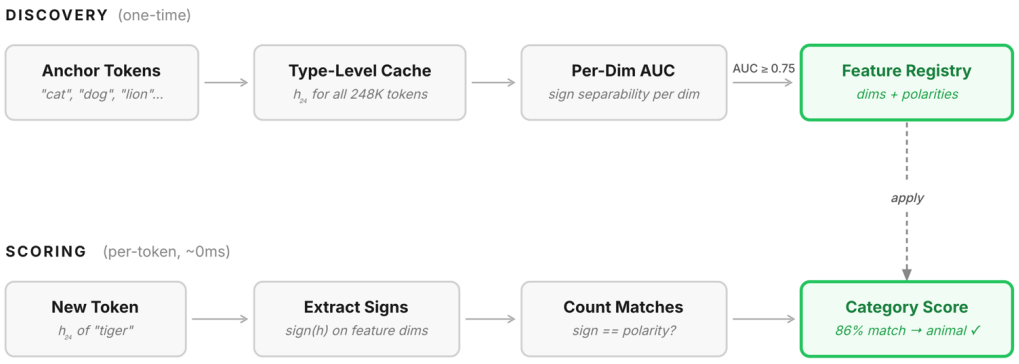

Figure 4: The Bag-of-Dims discovery and scoring pipeline. Top: one-time discovery builds a feature registry from anchor tokens and the type-level cache. Bottom: scoring any new token is instant via sign matching on registered dims.

- $\text{sign}\big(\text{RoPE}(\text{Norm}(\mathbf{h}) \cdot W_k^\top)\big)$ for K dimensions
- $\text{sign}(\text{Norm}(\mathbf{h}) \cdot W_v^\top)$ for V dimensions

where Norm is the model's pre-attention normalization (RMSNorm for all language models here) and RoPE is the rotary position embedding applied to K after projection.

We build a single-token KV cache capturing the actual K and V tensors through the full compute path (normalization, projection, RoPE for K). We apply the same discovery procedure to K and V dimensions and compare category-level AUC against residual-stream baselines (§3.5).

3 Experiments

All experiments use publicly available models. Language: Qwen 3.5-4B (Alibaba), Gemma 3-4B-pt (Google), Mistral 7B v0.3 (Mistral AI), and Qwen3-32B (Alibaba). Vision: DINOv2-Base (Meta, self-supervised) and ViT-Base (Google, supervised on ImageNet-21k). Audio: Audio Spectrogram Transformer (MIT, supervised on AudioSet/ESC-50). The single-token type cache for each model is computed once and shared across all analyses.

3.1 Sign Encodes Content, Magnitude Encodes Confidence

Sign-only prediction. We evaluate whether sign patterns alone suffice for next-token prediction. Given the final hidden state $\mathbf{h}$ for 200 prompts, we compute logits as $\text{sign}(\mathbf{h}) \cdot W$ (all magnitudes set to 1). We compare against the full representation and a random-sign control.

Table 1: Sign-only next-token prediction accuracy ($N = 200$ prompts per model).

Model	Method	Top-1	Top-5	Top-10	Top-100
Qwen 3.5-4B	Full $\mathbf{h} \cdot W$	100%	100%	100%	100%
	$\text{sign}(\mathbf{h}) \cdot W$	56.5%	84.0%	91.0%	97.0%
	Top-800 loud (sign only)	68.5%	90.0%	95.5%	99.5%
	Random sign permutation	0%	0%	0%	0%
Gemma 3-4B	Full $\mathbf{h} \cdot W$	100%	100%	100%	100%
	$\text{sign}(\mathbf{h}) \cdot W$	49.0%	72.0%	79.0%	96.5%
	Top-800 loud (sign only)	53.5%	81.5%	90.0%	100%
	Random sign permutation	0%	0%	0%	0%
Mistral 7B	Full $\mathbf{h} \cdot W$	100%	100%	100%	100%
	$\text{sign}(\mathbf{h}) \cdot W$	62.5%	93.0%	97.5%	100%
	Top-800 loud (sign only)	73.5%	96.5%	99.5%	100%
	Random sign permutation	0%	0%	0%	0%
Qwen3-32B	Full $\mathbf{h} \cdot W$	99.5%	100%	100%	100%
	$\text{sign}(\mathbf{h}) \cdot W$	37.5%	59.5%	67.0%	95.5%
	Top-800 loud (sign only)	51.5%	77.0%	84.0%	95.0%
	Random sign permutation	0%	0%	0%	0%

Sign alone preserves 60–93% top-5 accuracy across architectures (vs. 0% for random permutations). The 800 highest-magnitude dimensions achieve better accuracy than all dims across all four models (including D=5120). Low-magnitude dimensions act as noise: magnitude identifies which dimensions carry signal, sign encodes what they say.

Pure Hamming prediction (no learned decoder). A natural objection is that $\text{sign}(\mathbf{h}) \cdot W$ still relies on the LM head W. To rule this out, we test pure sign matching with no learned components on either side. We define:

- **Sign context majority:** For each dim, take the majority sign across all context token embeddings. Score each vocab token by sign agreement with this majority pattern.

No magnitudes appear on either side. Every dimension contributes exactly one vote.

Table 2: Pure Hamming prediction without the LM head ($N = 200$ prompts). No magnitudes or learned parameters are used in any sign-based method.

Model	Method	top-100	top-4096	Magnitudes?
Qwen 3.5-4B	Full dot (baseline)	62.5%	92.5%	Yes
	Sign context majority	53.0%	80.5%	None
Gemma 3-4B	Full dot (baseline)	33.5%	66.0%	Yes
	Sign context majority	62.5%	90.0%	None
Mistral 7B	Full dot (baseline)	0.5%	5.5%	Yes
	Sign context majority	58.5%	81.5%	None

The sign context majority achieves 80–90% top-4096 across all architectures with zero learned parameters. On Gemma and Mistral, it actually beats the full-dot baseline because untied embeddings make h_0 a poor projection target; the sign method avoids this entirely by reading only embedding signs without touching h_{final}. If predictive structure were encoded in magnitude-weighted combinations decodable only by W, uniform-weight Hamming matching would yield chance-level results. Instead it achieves 80–90%.

3.2 DIMENSION INDEPENDENCE

The Bag-of-Dims framework treats each dimension as an independent channel. We test this directly by measuring pairwise mutual information between dimension signs.

For dimensions d_i and d_j, let $S_i = \mathbf{1}[h_{d_i} > 0]$ and $S_j = \mathbf{1}[h_{d_j} > 0]$. The MI is:

$$I(S_i; S_j) = \sum_{a,b \in \{0,1\}} P(S_i{=}a, S_j{=}b) \log_2 \frac{P(S_i{=}a, S_j{=}b)}{P(S_i{=}a)\, P(S_j{=}b)} \tag{6}$$

Table 3: Pairwise mutual information between dimension signs (1000 random pairs per condition).

Model	Condition	Mean MI (bits)	Pairs > 0.01	Max MI
Qwen 3.5-4B	Type-level	0.0014	1.4%	0.021
	Contextual (200×128)	0.0006	0.1%	0.013
Gemma 3-4B	Type-level	0.0011	0.6%	0.015
	Contextual	0.0008	0.5%	0.015
Mistral 7B	Type-level	0.0051	13.9%	0.094
	Contextual	0.0006	0%	0.010
Qwen3-32B	Type-level	0.0014	2.9%	0.050
	Contextual (200×128)	0.0004	0%	0.005

Across all models and conditions, mean pairwise MI is below 0.006 bits (maximum possible: 1.0 bit). No pair exceeds 0.1 bits. Critically, context reduces pairwise coupling rather than creating it: mean contextual MI is lower than type-level MI for every model. This rules out the possibility that attention introduces cross-dimension entanglement.

The MLP ablation (Appendix B) provides the complementary functional test: an MLP with full cross-dimension capacity (128 hidden units) adds zero AUC over per-dim logistic regression at any training scale. Whatever residual coupling exists provides no practical benefit for feature reading.

These measurements establish a necessary condition for the Bag-of-Dims framework: dimensions carry independent information. The same independence holds for vision and audio transformers (§3.7).

3.3 ZERO-TRAINING FEATURE DISCOVERY

We apply the discovery method (§2) to 175 semantic categories spanning animals, emotions, numbers, code constructs, grammar, professions, and others (category tiers in Appendix C). Each category has 50 hand-curated anchor tokens. The negative set is the full vocabulary.

Table 4: Cross-model feature discovery (175 categories, 50 anchors each, full-vocabulary negatives). Per-dim AUC: best single dimension's separability. Prototype AUC: composite sign-agreement classifier using all dims above the 0.75 registration threshold (0.70 fallback when no dim clears 0.75).

Model	Discoverable (≥ 0.75)	Per-dim AUC	Prototype AUC	Exceed null p_{95}
Qwen 3.5-4B (h_{24})	161/175 (92%)	0.801	0.980	175/175
Gemma 3-4B (h_{34})	137/175 (78%)	0.772	0.975	175/175
Mistral 7B (h_{24})	175/175 (100%)	0.844	0.993	175/175
Qwen3-32B (h_{48})	154/175 (88%)	0.792	0.978	175/175

All 175 categories exceed the null calibration threshold on every model. The same categories emerge across four architectures despite different training data, vocabulary sizes, and dimension counts. Per-dim AUC measures discovery quality (can individual dims detect this category?); prototype

AUC measures practical detection (how well does the full composite classifier work?). The gap from per-dim to prototype (0.80 $\rightarrow$ 0.98) reflects the benefit of combining multiple dims via sign agreement, with no training involved.

Probe comparison. To test whether a trained linear combination adds value over per-dim reading, we train logistic regression on all 175 categories. The probe achieves mean AUC 0.9997 vs. our sign method's 0.9814, a difference of $+0.018$. Inspecting probe weights:

Table 5: Probe weight analysis: the probe learns axis-aligned voting.

Metric	Value
Sign agreement (probe weight sign vs. polarity, 6,962 dims with AUC > 0.7)	99.9%
Spearman ρ (signed weight vs. signed per-dim AUC)	0.72
MLP (2-layer, 128 hidden) vs. LogReg advantage	-0.001

The probe converges to magnitude-weighted axis-aligned voting, not a rotated direction. An MLP with full cross-dimension capacity adds nothing at any training scale (50–2000 positive examples; Appendix B), confirming that cross-dimension structure provides no practical benefit even when the probe has ample samples to learn arbitrary rotations.

Unsupervised discovery. Beyond the 175 curated categories, we test whether features emerge without any human-provided labels. Here we take signs relative to each dimension's vocabulary mean ($\text{sign}(h_d - \mu_d)$) rather than relative to zero: with curated anchors the per-dimension AUC already absorbs each dim's baseline firing rate, but the unlabeled nearest-neighbor search has no such reference, so dims with a strong positive or negative bias would otherwise dominate the Hamming distance. This recentering is diagonal (axis-aligned)—no rotation or cross-dimension mixing—and it changes the sign of *only* the baseline-biased dimensions: raw and mean-relative signs agree on 95% of near-balanced dims but differ on the strongly biased ones (per-dim sign-disagreement correlates with baseline bias at $\rho = 0.99$). The discriminative dimensions that define a feature keep their raw standard-basis sign; centering merely stops the uninformative "broadcast" dimensions from swamping the unlabeled distance. Starting from a random vocabulary token as anchor, we find its $K = 20$ nearest neighbors by Hamming distance on these centered signs, build a prototype from dims where all neighbors agree on sign, and verify it fires on a coherent, sparse token set. Repeating for 1500 random seeds:

Table 6: Unsupervised feature discovery (random seeds, no labels).

Model	Features found	Mean dims/feature	Fire rate ($<$0.1% vocab)
Qwen 3.5-4B	1500/1500 (100%)	$\sim$897	99.9%
Gemma 3-4B	1500/1500 (100%)	$\sim$609	99.9%
Mistral 7B	1500/1500 (100%)	$\sim$1488	98.9%
Qwen3-32B	1500/1500 (100%)	$\sim$2826	99.9%†

†At D-adjusted threshold (0.95 agreement for D=5120).

Every random seed produces a valid, sparse feature (firing on $<$0.1% of vocabulary). Manual inspection of a random sample confirms the top-activating tokens typically share semantic or phonological similarity (see examples in Appendix A). The method does not depend on carefully curated anchor sets. Sign-based detection also recovers features from Elhage et al.'s toy superposition model (23/24 features from a 20-dim bottleneck encoding 24; §5, Appendix F).

3.4 Features Generalize to Context—and Act Causally There

A natural objection to the type-level cache is that tokens in isolation may behave differently from tokens in natural text. We test this directly.

Contextual AUC. Type-level prototypes (discovered from isolated tokens) generalize to tokens in running text without modification. Scoring 25,600 tokens (200 prompts $\times$ 128 tokens) against type-level prototypes: mean contextual AUC 0.814 (Qwen), 0.722 (Gemma), 0.856 (Mistral). The sign patterns that define a category at type-level remain detectable in context. Approximately 58% of dimensions maintain their type-level sign in context; the flipping 42% are predominantly low-magnitude dimensions below the prototype threshold. This is consistent with attention carrying contextual role on the low-magnitude dims while preserving type identity on the high-confidence, feature-registered dims.

Polysemy: prototypes read contextual meaning, not just token identity. A sharper test of contextual reading: take a polysemous word whose senses straddle a category boundary (e.g. "bat" as *animal* vs. sports equipment, "train" as *vehicle* vs. the verb), and ask whether the relevant category prototype scores the word higher in its category-sense context than in its other-sense context. This is a within-method test—the same per-dim sign prototype, scored on the contextual hidden state—and it directly addresses the objection that bag-of-dims detects only token identity: the token is identical across both sentences, so any difference must come from context.

We assemble 77 such cross-category cases spanning animals, weapons, vehicles, fruit, trees, instruments, and others, and score the target word's contextual representation against its category prototype in both sentences. The target token is verified to be located in both sentences (cases that fail tokenization are excluded), and we report every case, including failures. Because the prototype's breadth is set by the registration threshold τ, we sweep τ rather than tuning a single value.

Table 7: Cross-category polysemy: fraction of cases where the category prototype scores the category-sense context higher than the other-sense context, and pooled AUC, at the τ that maximizes AUC per model (full sweep stable across $\tau \in [0.55, 0.70]$). On-thesis (per-dim sign prototype scoring, no whole-vector similarity), token-validated, all cases reported.

Model	D/V	best τ	Accuracy	Pooled AUC
Qwen 3.5-4B	0.010	0.55	79%	0.768
Mistral 7B	0.13	0.55	78%	0.746
Gemma 3-4B	0.010	0.58	77%	0.738
Qwen3-32B	0.034	—	60%	0.671

On three of the four models the category-sense context scores higher in 77–80% of cases (pooled AUC 0.72–0.77), and the effect *strengthens* as the prototype broadens (lower τ, more dimensions voting)—it is not a tuned artifact. Separations are clean on ontologically distinct pairs ("train" the vehicle scores 0.93 vs. 0.45 for "train the network"; "owl" the bird 0.94 vs. 0.56; "pine" the tree 0.94 vs. 0.61); failures cluster where both senses are concrete and category-adjacent (body parts, metals). This is contextual meaning read directly from the standard basis, using the same prototype discovered from isolated tokens—no whole-vector similarity, no training.

Qwen3-32B is the exception, remaining near chance (AUC 0.60–0.67) at every layer and threshold we tested (full layer$\times\tau$ sweep in Appendix G). This tracks its low dimension-to-vocabulary ratio ($D/V = 0.034$): as established for per-dim discovery (§5, Limitation 4), models with fewer dimensions per vocabulary token carry less sharp per-dim category structure, and the contextual read inherits this. We note the scope of the claim overall: this reads *cross-category* sense shifts (a word moving in or out of a semantic category), not fine-grained within-category word-sense disambiguation.

Type-level prototypes thus generalize to context: stable high-magnitude dims preserve type identity while context modulates the read on the remaining dims.

Features are causally operative in context, not merely readable. The results above show type-level prototypes *detect* in running text. We now test whether they are *causally active* during a live forward pass: we flip a feature's registered signs to their opposite (preserving each dimension's magnitude) at a late layer, at all token positions, while the model generates, and measure the change in the mean logit of the category's target tokens. The sign pattern is the only thing edited; magnitudes are untouched.

Table 8: Causal sign-flip during the live forward pass (mean target-logit change, 5 categories per model). **Away**: flip the feature's signs away from expected. **Toward**: force the same dims to their expected sign (same dims, same magnitudes—isolates sign from magnitude). **Disjoint**: flip a different feature's coalition with the target's shared dims removed (isolates concept-specificity from overlap). **Random**: flip an equal number of random dimensions.

Model	Away	Toward	Disjoint	Random
Qwen 3.5-4B (h_{24})	-10 to -14	≈ 0	≈ 0	≈ 0
Gemma 3-4B (h_{34})	-19 to -24	≈ 0	≈ 0	≈ 0
Mistral 7B (h_{24})	-5 to -7	≈ 0	≈ 0	≈ 0
Qwen3-32B (h_{60})†	-5 to -9	≈ 0	≈ 0	≈ 0

†The intervention requires patching near the output; on the 64-layer model the effect appears at h_{60}, not the detection-optimal h_{48} (§5).

Three findings establish that the sign pattern is the causal variable. First, **sign, not magnitude**: the *away* flip suppresses the concept while the *toward* flip—identical dimensions, identical magnitudes, only the sign direction differs—does nothing. Second, **the coalition is the causal unit**: a single dimension or a small subset is inert; sweeping the fraction of the coalition flipped shows the effect switches on only past $\sim$200–500 dimensions and grows with coalition size. This is why per-dimension interventions fail—a single sign is outvoted by the rest. Third, **concept-specificity**: flipping a *disjoint* coalition (a different feature's dims with shared dimensions removed) leaves the target untouched, indistinguishable from random. An apparent "general damage" from flipping unrelated features is entirely explained by coalition overlap—features share dimensions, so flipping one partially flips another.

This is the in-context counterpart to the static write-side analysis of §3.6: there, neuron coalitions are shown to *write* the sign pattern (from the weights alone); here, flipping that same pattern during generation *changes behavior*. Two independent angles—write-side weight inspection and read-side intervention—converge on the same unit: the sign coalition. The standard-basis sign is therefore not only a readout of the model's content but a quantity the model causally computes with. Per-category numbers and the disjoint specificity control are reported in Appendix H.

3.5 Features Survive Attention Projections

We test whether per-dim features survive the K and V projection transforms applied during attention computation.

We build a single-token KV cache for each model's full vocabulary, capturing K and V tensors through the complete compute path (layernorm, projection, RoPE). We run the same 175-category discovery on K and V dimensions independently (Table 9).

Table 9: Feature discovery in K/V projections (full-vocabulary negatives, same 50 anchors). Qwen3-4B is substituted for Qwen 3.5-4B because the latter uses hybrid attention, which does not produce standard KV caches at all layers.

Model	Space	Mean AUC	Discoverable ($\geq$0.75)	Exceed p_{95}
Gemma 3-4B (L25)	K	0.757	97/175 (55%)	175/175
	V	0.759	106/175 (61%)	175/175
Mistral 7B (L24)	K	0.850	174/175 (99%)	175/175
	V	0.810	167/175 (95%)	175/175
Qwen3-4B (L24)	K	0.757	93/175 (53%)	175/175
	V	0.757	85/175 (49%)	175/175
Qwen3-32B (L48)	K	0.744	70/175 (40%)	175/175
	V	0.734	57/175 (33%)	175/175

All 175 categories exceed null calibration on both K and V across all four architectures. Discoverability at the $\geq$0.75 threshold ranges from 33–61% (Gemma, Qwen3-4B, Qwen3-32B) to 95–99% (Mistral), tracking the same dimension-to-vocabulary ratio that governs residual-stream sharpness. Per-dim feature discovery works in K and V space across all four architectures, confirming that attention projections preserve axis-aligned structure. The same K/V preservation holds for vision and audio transformers (§3.7).

3.6 Circuit Tracing: FFN Neurons Write Axis-Aligned

If features are axis-aligned sign patterns in the residual stream, the mechanism that writes them must also be axis-aligned. Each FFN neuron writes to the residual stream via its column in the `down_proj` weight matrix. We test whether individual neurons' write patterns match discovered features.

Method. For each of 1500 unsupervised prototypes discovered at layer h_L, we examine the FFN at layer $L-1$ (whose output becomes h_L). For each neuron, we compute sign agreement between its `down_proj` column (restricted to prototype dims, filtered by magnitude threshold) and the prototype's expected signs.

Single-neuron linkage:

Table 10: Single-neuron prototype linkage (1500 unsupervised prototypes, confidence-thresholded Hamming). The writer layer is $L-1$ for prototypes at h_L; for Gemma the unsupervised prototypes (here and in Table 12) are built at h_{25}, the layer at which its released SAE and K/V analysis are pinned, so its writer is L24. The linkage holds at h_{25} just as feature discovery holds at h_{34} (Table 4); Table 13 confirms the same FFN-write mechanism at Gemma h_{34}, so the circuit is demonstrated at two Gemma layers.

Model	Writer layer	Neurons	Conf >0.60	Conf >0.70	Conf >0.80
Qwen 3.5-4B	L23	9,216	84.7%	20.1%	5.8%
Gemma 3-4B	L24	10,240	91.5%	19.7%	1.6%
Mistral 7B	L23	14,336	61.0%	1.9%	0%

Control. The identical procedure on random Gaussian weights and column-shuffled trained weights:

Table 11: Circuit tracing: trained vs. random controls (pure Hamming >0.70 threshold).

Model	Trained	Random	Shuffled
Qwen 3.5-4B	88/1500 (6%)	0/1500	0/1500
Gemma 3-4B	27/1500 (2%)	0/1500	0/1500
Mistral 7B	3/1500 (0.2%)	0/1500	0/1500

Random weights never achieve the linkage thresholds that trained models reach. The strong writer neurons are exclusively a product of training. On Qwen3-32B (25,600 neurons, 5120 dims), individual neuron agreement is limited by prototype breadth ($\sim$245 dims at AUC $\geq$ 0.65); the coalition analysis below provides the meaningful comparison.

Coalition linkage: the remaining 80%. A single neuron achieves only 0.58 median agreement because prototypes span $\sim$900 dimensions, too many for any single neuron to cover. We test whether *coalitions* of neurons collectively write the full pattern. For each prototype, we rank all neurons by individual agreement. We then take the majority vote of the top-K neurons' `down_proj` signs on the prototype's dims and measure agreement with the prototype.

Table 12: Top-K neuron coalition: majority-vote sign agreement with prototypes (median across 1500 prototypes).

K neurons	Qwen 3.5-4B	Gemma 3-4B	Mistral 7B	Qwen3-32B
1	0.583	0.587	0.565	0.533
10	0.635	0.648	0.606	0.571
50	0.704	0.740	0.672	0.638
100	0.743	0.793	0.712	0.679
200	0.786	0.845	0.758	0.728
500	0.842	0.908	0.823	0.798
1000	0.876	0.942	0.866	0.850

At $K = 200$, 99.9% of Qwen 3.5-4B prototypes exceed 0.70 agreement (vs. 20% from a single neuron). The curve is logarithmic for every model: steep improvement from $K = 1 \rightarrow 100$, then diminishing returns, all climbing monotonically toward $\sim$0.85–0.94 by K=1000–2000. Gemma reaches 0.94 at $K = 1000$; Mistral converges to the same range despite its larger FFN width (14,336 neurons). Qwen3-32B (25,600 neurons) converges along the same logarithmic path but more gradually per neuron added (0.638 at K=50, 0.728 at K=200, 0.850 at K=1000): its prototypes are the widest ($\sim$2000 dims) and its FFN the largest, so more neurons are needed to cover the pattern, but the coalition still assembles it. The write mechanism is distributed across all four models: many neurons each contribute partial axis-aligned writes that combine via majority vote.

The neuron activations themselves carry the features. The results above concern the `down_proj` *weights*—the directions neurons write. A sharper question is whether the neuron *activations* (the SwiGLU intermediate $\mathrm{SiLU}(\mathrm{gate}(x)) \cdot \mathrm{up}(x)$, of width 9,216–25,600) are themselves a readable per-dim sign space, the same way the residual stream is. This is not implied by the write-side result: the activation sign is $\mathrm{sign}(\mathrm{gate}(x)) \cdot \mathrm{sign}(\mathrm{up}(x))$, a nonlinear product of two projections, not a single linear reprojection of the residual, so axis-aligned structure could in principle be destroyed by the gate interaction. We test it directly: hook the FFN at block $L-1$ (the writer of h_L), read each neuron's activation sign for the single-token vocabulary, and run the identical per-dim AUC and prototype procedure used for the residual stream. To keep the comparison exact, we evaluate the residual stream under the same harness (same 175 categories, same 20K-token negative sample, same τ and null calibration) rather than against the full-vocabulary numbers of Table 4.

Table 13: Per-dim sign features in FFN *neuron activations* vs. the residual stream, under an identical harness (175 categories, 171 for Mistral; 20K-token negative sample; τ=0.70). The neuron read hooks block $L-1$, whose output writes h_L. Neuron activations carry per-dim sign features at parity with the residual stream.

		Per-dim max AUC		Prototype AUC		
Model	Neurons	Neuron	Residual	Neuron	Residual	Exceed null
Qwen 3.5-4B (h_{24})	9,216	0.819	0.801	0.991	0.983	174/175
Gemma 3-4B (h_{34})	10,240	0.818	0.773	0.976	0.962	174/175
Mistral 7B (h_{24})	14,336	0.861	0.843	0.994	0.992	171/171
Qwen3-32B (h_{48})	25,600	0.787	0.792	0.973	0.980	165/175

Despite the nonlinear gate, neuron activations match the residual stream within ± 0.05 on per-dim AUC and within ± 0.02 at the prototype level, exceeding null calibration on 165–174 of 175 categories. Three of four models read *slightly higher* in neuron space than in the residual, and Qwen3-32B is within 0.005. The same null control as for the residual holds: no random anchor set forms a prototype (null $p_{99} = 0.500$). The Bag-of-Dims structure is therefore not specific to the residual stream—it is already present, axis-aligned and sign-readable, in the FFN's own activation space, and survives the SwiGLU gate intact.

Mechanistic picture. This completes the circuit: FFN neuron coalitions at layer $L-1$ write axis-aligned sign patterns via `down_proj` majority vote—and the neuron activations driving that write

are themselves a sign-readable per-dim space (Table 13) → the residual stream at h_L stores them as per-dim features → W_k/W_v project them preserving axis-alignment → attention reads and routes per-dim. The entire pipeline operates in the standard basis. The weight-side analysis requires loading one matrix and takes ∼30 seconds with no training and no forward passes.

3.7 Cross-Modality Universality

The experiments above establish per-dim sign structure in language models. We now test whether the same framework applies to transformers trained on entirely different modalities and, crucially, whether it requires classification supervision. We compare a self-supervised and a supervised vision model on the same data, then extend to audio.

Models and data. We evaluate three non-language transformers:

- **DINOv2-Base** (Oquab et al., 2024) ($D = 768$, 12 layers): self-supervised via self-distillation on LVD-142M with *no classification labels*.
- **ViT-Base** (Dosovitskiy et al., 2021) ($D = 768$, 12 layers): supervised on ImageNet-1K classification (1000 categories).
- **AST** (Gong et al., 2021) ($D = 768$, 12 layers): supervised on AudioSet classification, evaluated on all 2000 ESC-50 clips (40 per category) (Piczak, 2015).

DINOv2 and ViT-Base share the same ViT-Base architecture ($D = 768$, 12 layers) and are evaluated on the same 1000 ImageNet validation images. The only difference is the training objective: self-distillation vs. cross-entropy classification. This isolates the effect of supervision on per-dim structure.

Method. We apply the identical per-dim AUC procedure from §2: extract CLS token hidden states at the final layer (h_{12}) for vision models, and mean-pool over all sequence positions for AST (which uses a distillation token rather than a standard CLS for feature extraction). We compute sign patterns and test whether individual dimensions separate semantic categories.

Vision: self-supervised vs. supervised on the same data.

Table 14: Per-dim sign detection of ImageNet superclasses: DINOv2 (self-supervised, no labels) vs. ViT-Base (supervised, 1000-class classification). Same architecture, same evaluation data, different training objective.

	DINOv2 (self-supervised)			ViT-Base (supervised)		
Category	N	Max AUC	Dims>0.70	N	Max AUC	Dims>0.70
Primate	18	0.819	30	18	0.808	107
Cat	13	0.780	26	13	0.807	48
Insect	20	0.760	6	20	0.779	36
Fish	16	0.741	10	16	0.765	39
Flower	11	0.736	9	11	0.759	29
Musical instr.	20	0.721	5	20	0.802	14
Reptile	34	0.719	1	34	0.748	12
Bird	38	0.701	1	38	0.744	8
Furniture	12	0.749	2	12	0.729	6
Vehicle	33	0.667	0	33	0.737	3
Dog	118	0.666	0	118	0.758	8
Food	46	0.697	0	46	0.679	0
Above 0.70		9/12			11/12	
Prototype AUC (K=50)		0.831			0.977	

Both models exhibit per-dim sign structure. The supervised model achieves higher per-dim AUC and engages more dims per category (median 12 vs. 5 dims above 0.70), consistent with explicit classification pressure sharpening per-dim specificity. But the structure is already present without supervision: DINOv2 detects 9/12 superclasses from per-dim signs alone, with a 50-dim prototype

achieving AUC 0.831 for binary animal detection. Classification amplifies per-dim specificity; it does not create the underlying organization.

AST: per-dim sign detects audio categories.

Table 15: Per-dim sign detection of ESC-50 audio categories (AST, h_{12}, mean-pool over sequence positions, all 2000 clips, 40 per category). Showing top-10 of 50 categories.

Category	Max dim AUC	Dims > 0.70
Train	0.970	69
Thunderstorm	0.961	79
Sea waves	0.942	73
Pouring water	0.939	78
Rooster	0.939	117
Dog	0.938	44
Rain	0.933	58
Sheep	0.931	71
Cow	0.923	68
Chainsaw	0.921	58
All 50 categories: 50/50 exceed 0.70; 47/50 exceed 0.80		

All 50 ESC-50 categories exceed AUC 0.70, with 47/50 exceeding 0.80—the highest per-dim AUC across all models and modalities tested. The combination of domain-specific training (AudioSet) and a compact category space produces near-perfect per-dim discrimination.

Sign-only nearest-neighbor retrieval.

Table 16: Sign-only 1-NN accuracy (Hamming distance on ± 1 signs, no magnitudes, no training).

Model	Task	Accuracy	Chance	Lift
DINOv2-Base	Animal vs. non-animal	93.0%	60.2%	+32.8%
ViT-Base	Animal vs. non-animal	96.0%	60.2%	+35.8%
AST	Same superclass (5-class)	97.0%	20.0%	+77.0%

Pure Hamming distance on sign patterns achieves 93–97% retrieval accuracy across all models and modalities. The 3% gap between DINOv2 and ViT-Base reflects the classification sharpening effect, but both far exceed chance with zero learned parameters.

Probe ablation: cross-dimension structure is negligible across modalities.

Table 17: Probe ablation: LogReg (per-dim, axis-aligned) vs. MLP (128 hidden, cross-dim capacity).

Model	Training	LogReg AUC	MLP AUC	Gap
DINOv2-Base	Self-supervised	0.931	0.938	$+0.007$
ViT-Base	Supervised	0.978	0.978	-0.000
AST	Supervised	1.000	1.000	$+0.000$

An MLP with full cross-dimension capacity adds at most $+0.007$ AUC over axis-aligned logistic regression, confirming negligible cross-dim structure regardless of training objective. On the supervised models (ViT-Base and AST), the MLP provides exactly zero benefit. This replicates the language model finding (§3.3) across modalities and supervision regimes.

Pairwise MI. Cross-dimension mutual information: 0.001 bits (DINOv2), 0.001 bits (ViT-Base), 0.002 bits (AST), comparable to language models (0.001–0.005 bits). Only 0.1% of DINOv2 dim pairs exceed 0.01 bits. Per-dim independence is modality-universal and does not depend on the training objective.

K/V projections preserve per-dim structure across modalities.

Table 18: Feature discovery in K/V attention projections for non-language models (same categories, layer 11). Categories discoverable = max per-dim AUC > 0.70.

Model	Space	Discoverable	Exceed null p_{95}
DINOv2-Base (self-supervised)	H (residual)	12/12	12/12
	K	12/12	11/12
	V	12/12	11/12
ViT-Base (supervised)	H (residual)	12/12	12/12
	K	12/12	11/12
	V	12/12	12/12
AST (audio, 50 categories)	H (residual)	50/50	50/50
	K	50/50	50/50
	V	50/50	50/50

The same categories discoverable from the residual stream are also discoverable from K and V projection dimensions on all three non-language models (Table 18). On AST, all 50 categories exceed null calibration in both K and V. This confirms that the W_k/W_v matrices preserve axis-aligned structure universally, extending the language model finding (§3.5) to vision and audio.

FFN write mechanism across modalities. We test whether the FFN circuit tracing (§3.6) extends to non-language models. Because vision/audio models have smaller hidden dimension ($D = 768$ vs. 2560–4096), we use two complementary analyses:

Static specificity. For each prototype's best-matching neuron, does its `fc2` column agree more with the matched prototype than with unmatched prototypes?

Activation-weighted coalition. Rank neurons by category selectivity (firing rate on category inputs minus overall firing rate), then take majority vote of the top-K selective neurons' `fc2` signs on prototype dims. Compare against a shuffled-label control where "selective" neurons are selected from randomized category assignments.

Table 19: FFN write mechanism: static specificity and activation-weighted coalition. Static: best neuron agreement with matched vs. unmatched prototypes. Coalition: top-50 category-selective neurons, mean `fc2` sign agreement (random-neuron control in parentheses).

Model	Matched	Unmatched	Gap	Coalition K=20 (control)
DINOv2-Base (self-supervised)	0.69	0.49	+0.20	0.50 (0.50)
ViT-Base (supervised)	0.79	0.49	+0.30	0.52 (0.50)

Both models contain category-specific sign patterns in their `fc2` columns: the best-matching neuron agrees with its matched prototype far more than with unmatched ones (gap $+0.20$ for DINOv2, $+0.30$ for ViT-Base), and the supervised model's static specificity is the stronger of the two (matched 0.79 vs. 0.69). The *write vocabulary* is thus present in the weights of both, sharpened by classification. The activation-weighted coalition, however, is flat for both models: category-selective neurons write the matching pattern at only 0.50 (DINOv2) and 0.52 (ViT-Base), barely above the random-neuron control (0.50). **Neither training objective produces an activation-level assignment of neurons to categories at this layer—the category structure lives in *which directions* the neurons write (`fc2` weights), not in a firing-rate gating of those writes.** This mirrors the language-model finding that the write mechanism is distributed across a coalition rather than concentrated in a few selectively-firing neurons (§3.6).

Summary. Three training objectives, next-token prediction (language), self-distillation (DINOv2), and classification (ViT-Base, AST), all produce the same per-dim sign organization. The read side of the circuit (residual stream detection, K/V preservation) is equally strong across modalities. The

write side (FFN specificity) is present in all models but sharpened by classification. The shared factor is the architecture: a residual stream with FFN writes.

4 Related Work

Sparse Autoencoders. SAEs decompose hidden states into overcomplete dictionaries of interpretable features (Bricken et al., 2023; Templeton et al., 2024; Cunningham et al., 2023). These methods assume features correspond to directions recoverable only through learned rotations, motivated by the superposition hypothesis (Elhage et al., 2022). Our work shows that for 175 semantic categories, comparable detection is achievable from the standard basis without training, and that the combinatorial capacity of sign patterns ($3^D - 1$ features from D dims) eliminates the bottleneck that motivates geometric packing (§5). Beyond the 175 curated categories, unsupervised discovery scales to 1500 features at 100% yield and 99% sparsity across all four language models, matching SAE-scale coverage without optimization. The SAE approach has since been carried to other modalities, with sparse autoencoders trained on CLIP vision-transformer activations (Joseph et al., 2025) and on audio latent spaces (Paek et al., 2025); both require training a dictionary per model. Our cross-modal results (§3.7) recover comparable per-feature structure in vision and audio transformers directly from the standard basis, with no dictionary to train.

Recent work raises concerns about SAE reliability: Makelov et al. (2025) demonstrate that SAEs produce interpretable features even on untrained transformers, questioning whether discovered features reflect learned computation. We test this directly: on randomly initialized models with identical architecture, per-dim category AUC drops to 0.60–0.68 (0/175 categories detectable on Qwen/Gemma), confirming that the structure we find requires training and is not an artifact of architecture or the discovery method.

The superposition hypothesis. Elhage et al. (2022) propose that transformers encode more features than dimensions via non-orthogonal geometric packing. We address this in §5: in real transformers, the cross-dim coupling that geometric superposition predicts is not observed (MI < 0.006 bits, vs. 0.05–0.10 in Elhage's toy bottleneck), and per-dim sign matching is a sufficient decoder both in real models and in the toy itself (71–100% feature recovery).

Individual neuron interpretability. Early work showed neurons encode concepts (Bau et al., 2017); the field moved away when "neurons are polysemantic" became consensus (Elhage et al., 2022). Our findings suggest polysemanticity may be localized to MLP interiors, while residual stream dimensions exhibit high per-dim specificity (AUC 0.80).

The privileged-basis question. Whether the standard basis is special has been debated in vision since the introduction of network dissection (Bau et al., 2017), which scores individual convolutional units by thresholding their activation maps against segmentation masks, and feature visualization (Olah et al., 2017), which finds basis directions interpretable more often than random ones but reports the question contested. The more sophisticated view holds that concepts are *learned, distributed* directions rather than individual axes: Net2Vec (Fong & Vedaldi, 2018) finds that "multiple filters are required to code for a concept," and TCAV (Kim et al., 2018) represents a concept as the normal vector to a linear classifier trained to separate concept examples from random ones, measuring importance by directional derivatives along it. What unites all of these, whether they favor axes or learned directions, is that they read a unit's activation *magnitude* and, in the distributed case, train a decoder to recover the direction. We ask a different question, and in a different place: not which direction encodes a concept, but whether the *sign* of a standard-basis coordinate in the residual stream is a sufficient, training-free decoder. The decomposition into sign (content) and magnitude (confidence) is what is new here; prior work neither separates the two nor reads features by sign agreement.

Linear probes. Probes (Alain & Bengio, 2017; Belinkov, 2022) confirm that representations contain information but require training per property. Our probe comparison reveals that trained probes converge to axis-aligned weights (99.9% sign agreement).

Logit lens. The logit lens (nostalgebraist, 2020) and tuned lens (Belrose et al., 2023) demonstrate that raw intermediate states are directly readable. Our framework extends this by explaining *why*: individual dimensions carry independent features, making intermediate states interpretable without any transformation.

Representation engineering. RepE (Zou et al., 2023) identifies concept "directions" for monitoring and control. Our work suggests these may correspond to subsets of dimensions with consistent sign patterns rather than geometric directions.

Across all these lines of work, the common assumption is that features require some transformation—learned or geometric—to decode. Our results show the standard basis already suffices.

5 DISCUSSION

Why do dimensions behave independently? We observe this as an emergent property of training, not an architectural guarantee. Circuit tracing provides a mechanistic sketch: FFN neurons write to specific dim subsets via `down_proj`. If neuron activations are sparse (few fire per token), each neuron writes to largely non-overlapping dimensions, inducing per-dim specificity. This is consistent with our MI finding (0.0014 bits) and with the structure surviving through learned projections. We emphasize that this is *functional* independence, cross-dimension capacity adds no practical value for feature reading, rather than a claim of strict statistical independence.

Per-layer coordinate systems. The dim-to-concept mapping is layer-specific: cross-layer Jaccard overlap is 0.042 (chance level). Dim 847 may encode "animal" at h_{24} but something entirely different at h_8 or h_{32}. The register file metaphor applies *within* each layer's output, not across the full stack. Each layer's FFN rewrites the residual stream into its own coordinate convention, and features discovered at h_L link specifically to layer $L-1$'s FFN neurons (not to neurons at other layers). This is why discovery must be performed per-layer and why prototypes are not transferable across layers. Full per-layer sweep data is provided in Appendix D.

Combinatorial coding. We discover 1500+ features from 2560 dimensions, with features using $\sim$897 dims each and sharing $\sim$35% by overlap. Features remain independently detectable (99% fire on $<$1% of vocabulary) because scoring uses sign agreement on each feature's own registered dim subset, not a global operation: two features can share dims without interfering, and the MI measurements ($<$0.006 bits; Table 3) confirm that shared dims introduce no practical coupling. The capacity is combinatorial rather than geometric: features need not compete for orthogonal directions.

Relationship to the superposition hypothesis. The superposition hypothesis (Elhage et al., 2022) proposes that transformers encode more features than dimensions by placing features at non-orthogonal directions that geometrically interfere, motivating learned rotations (SAEs) for decoding. Our results suggest the encoder/decoder distinction matters here: features are encoded as sign-magnitude pairs (sign carrying content, magnitude carrying confidence), and per-dim sign matching is a sufficient decoder regardless of whether the encoder packs features at non-orthogonal angles.

With D binary dimensions, the number of distinct sign-pattern features (subsets of k dims with specified signs) is:

$$\text{Capacity} = \sum_{k=1}^{D} \binom{D}{k} \cdot 2^k = 3^D - 1 \tag{7}$$

For $D = 2560$, this yields $\sim 10^{1220}$ possible features, effectively infinite. The address space is large enough that geometric packing is not the only available mechanism for representing many features.

We replicate Elhage et al.'s toy autoencoder ($x \mapsto W^\top W x$ with bottleneck $d \ll n$; full results in Appendix F). Superposition emerges as reported: more features than dimensions are represented, with non-orthogonal weight columns. 71–100% of the represented features are recoverable from per-dim sign patterns alone (AUC > 0.7); in a 20-dim bottleneck encoding 24 features, sign-based detection recovers 23/24 (96%). Sign-only reconstruction (Appendix F, Table 28) achieves 60–77% of full reconstruction quality at d=5, mirroring the sign-vs-full pattern in real LMs (Tables 1–2): both encode content in signs and refine with magnitudes.

We do not refute superposition: the toy's W has non-orthogonal columns and cross-dim sign MI of 0.05–0.10 bits (Table 28), real properties of its bottleneck regime. We add an empirical observation: the same toy already exhibits sign-as-content / magnitude-as-confidence, the structure we identify in real transformers, and per-dim sign matching is a sufficient decoder even where geometric interference is present. Real transformers do not show the toy's coupling signatures: pairwise MI is < 0.006 bits (Table 3, vs. 0.05–0.10 in the toy bottleneck), the cross-dim MLP buys nothing over per-dim

reading, and probes converge to axis-aligned weights. As the toy's bottleneck relaxes, its MI drops toward real-transformer levels (0.005 bits at n=200, d=20). The geometric coupling that defines superposition appears as a function of bottleneck pressure, and the regime where that pressure binds does not appear to be the regime real transformers operate in.

Contextual updating. The flipping dimensions are low-magnitude dims below the prototype threshold. The cross-category polysemy test (§3.4) shows context modulates the prototype read: a word scores higher against its category prototype in the matching context, so these context-dependent dims carry sense information rather than being noise; whether they encode other contextual properties (syntactic role, reference) is open. The high-magnitude, feature-registered dims preserve type identity, which explains why type-level prototypes generalize to context: they read from the stable high-magnitude partition. Contextual pairwise MI is lower than type-level, consistent with context diversifying rather than entangling dimensions.

The causal intervention is depth-gated, and read/write optima differ. The sign-flip intervention (§3.4) only changes behavior when applied near the output. On the 64-layer Qwen3-32B the effect is absent at the detection-optimal layer h_{48} but present at h_{60}; detection AUC at the two layers is nearly identical (0.91 vs 0.90), yet only the late patch is causally effective. The interpretation is consistent with the per-layer coordinate systems above: a sign pattern corrupted deep in the stack is re-derived by the intervening layers from earlier context before it reaches the logits, so the further the patch is from the output, the more it is overwritten. A feature is therefore *readable* across a wide band of layers but *causally flippable* only near the output—the read-optimal and write-optimal layers are distinct (12 layers apart on the 32B). This re-derivation also explains why early single-layer perturbation studies found sign flips inert: the corruption is corrected downstream. A full characterization of this depth-dependence is left to future work.

Sign suppresses but does not induce. The intervention is asymmetric. Flipping a feature's signs *away* from expected reliably suppresses the concept, but forcing them *toward* expected does not induce it: the high-confidence dimensions already carry the expected sign, so setting them is close to a no-op. Inducing a concept requires injecting *magnitude* (confidence), not setting sign (content)—consistent with the sign/magnitude decomposition throughout this work, and with the observation that activation-addition steering (e.g. on sparse-autoencoder features) operates by scaling a direction's magnitude rather than flipping signs. A magnitude-based induction primitive built from the type cache is a natural direction we do not pursue here.

Modality universality. The appearance of per-dim sign structure in DINOv2 and AST, trained on images and audio respectively with no shared data or vocabulary, suggests the structure is driven by the optimization process (gradient descent on transformer architectures) rather than properties of natural language specifically. The DINOv2 vs. ViT-Base comparison is particularly informative: same architecture, same evaluation data, but self-supervised vs. supervised training. Both exhibit per-dim sign structure, with classification merely sharpening it (11/12 vs. 9/12 superclasses, more dims per category above threshold). This pattern, structure present without supervision and amplified by supervision, implies that per-dim specialization is a convergent property of residual transformers trained by gradient descent, regardless of objective.

Why do different objectives converge? Why next-token prediction, self-distillation, and classification all produce the same per-dim structure is an open question. We note only that the shared factor is the architecture: a residual stream with additive FFN writes and sparse activation. A formal information-theoretic treatment is beyond the scope of this work.

Limitations. (1) Scale: tested on 4–32B language models and base-sized (86M parameter) vision/audio models; whether the structure persists at 70B+ is open, though consistent results from 4B to 32B suggest no degradation. (2) The causal sign-flip (§3.4) suppresses a concept reliably and concept-specifically, but is bounded in two ways. It is *directional*: flipping signs away from a feature suppresses it, but forcing signs toward a feature does not induce it—induction is a magnitude operation we do not provide a primitive for here (§5). It is also *depth-gated*: the effect requires patching near the output, as deeper corruptions are re-derived downstream (§5); we characterize the full depth-dependence only partially. The high detection AUC (0.97–0.99) also suggests a non-interventional use: real-time monitoring of per-dim sign patterns for safety-relevant features, enabling early stopping or fallback. (3) Type-level prototypes read type identity from the stable high-magnitude dims; systematically cataloging the contextual-role features encoded by the flipping 42% is future

work. (4) Cross-architecture variation in per-dim AUC tracks the dimension-to-vocabulary ratio rather than parameter count: Mistral ($D/V \approx 0.13$) achieves 0.844, Qwen3-32B (0.034) and Qwen 3.5-4B (0.010) reach 0.792–0.801, and Gemma (0.010) reaches 0.772. Models with more dimensions per vocabulary token achieve sharper per-dim discrimination. The same axis governs the contextual polysemy read (§3.4): three models reach 77–80% accuracy, while Qwen3-32B—whose absolute per-dim sharpness is lower despite its size—stays near chance at every layer tested, indicating the contextual read inherits the per-dim sharpness that D/V controls. (5) The 175 curated language model categories are a subset of the model's full feature inventory; unsupervised discovery at 1500 features (100% yield) suggests the total is much larger, but we have not characterized the full count. (6) The DINOv2/AST experiments validate detection, K/V discovery, and FFN circuit tracing from the final layers; we have not yet tested per-layer sweeps on non-autoregressive architectures to confirm the same layer-progression pattern observed in language models.

6 Conclusion

We have presented evidence that individual dimensions in transformer hidden states function as independent binary registers, encoding semantic features via their signs and confidence via their magnitudes. Converging experiments validate this across three language model families and three non-language transformers: sign patterns alone carry predictive content (80–90% top-4096 without any learned decoder); 175 semantic categories are discoverable from a type cache with zero training, scaling to 1500 features unsupervised; per-dim independence holds empirically (MI < 0.006 bits, MLP adds zero AUC over per-dim reading); the same features survive K/V attention projections and trace back to axis-aligned FFN writes; and the identical method works on vision and audio transformers trained with self-supervised, supervised, and next-token objectives alike.

We do not claim the standard basis is the unique or optimal feature basis, only that it suffices for practical feature reading at quality comparable to trained methods. And it is cheap to use: the cache is the only artifact, the rest is bookkeeping over signs. The combinatorial capacity of sign patterns ($3^D - 1$ features) means this approach faces no inherent scaling limit as models grow larger.

These findings suggest that practical feature reading may not require the optimization overhead currently assumed. If the standard basis suffices, the primary challenge shifts from *finding the right rotation* to *characterizing what each dimension encodes at each layer*—a cataloging problem rather than an optimization problem.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

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

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, Anthropic*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread, Anthropic*, 2022.

Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Interspeech*, 2021.

Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering CLIP's vision transformer with sparse autoencoders. In *CVPR 2025 Workshop on Mechanistic Interpretability for Vision (MIV)*, 2025. arXiv:2504.08729.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, 2018.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Sparse autoencoders can interpret randomly initialized transformers. *arXiv preprint arXiv:2501.17727*, 2025.

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

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Nathan Paek, Yongyi Zang, Qihui Yang, and Randal Leistikow. Learning interpretable features in audio latent spaces via sparse autoencoders. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025. arXiv:2510.23802.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic*, 2024.

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 Koyber, Dawn Song, Matt Fredrikson, J Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Unsupervised Feature Discovery and SAE Comparison

A.1 SAE Head-to-Head

We directly compare against Google's Gemma Scope 2 SAE (16K features, trained on millions of contextual tokens) on Gemma 3-4B, layer 25—the layer at which Google released the SAE, so the comparison (and the K/V analysis of Table 9, which aligns to it) is pinned to h_{25} rather than the h_{34} discovery layer. For each of the 175 categories, we compare our sign prototype (50 anchors, $\sim$200 dims, zero training) against the SAE's best single feature (selected from 16K candidates by highest mean activation on anchor tokens). AUC is computed at the prototype level: tokens are scored by sign agreement fraction and then ranked.

Table 20: Sign prototype vs trained SAE (Gemma Scope 2, 16K features) on Gemma 3-4B.

Method	Mean AUC	Sign wins
Sign prototype (zero training)	0.952	—
SAE best single feature	0.824	173/175
SAE top-5 + LogReg	0.873	161/175
SAE top-10 + LogReg	0.932	124/175
SAE top-20 + LogReg	0.958	84/175

At the individual feature level (how SAEs are typically reported), sign prototypes win on 173/175 categories. Combining top-20 SAE features via trained logistic regression slightly surpasses sign prototypes (0.958 vs 0.952), but requires searching 100 candidate features per category, training a classifier, and encoding tokens through the SAE ($\sim$150 seconds for full vocabulary vs sub-second for sign scoring). The SAE itself requires millions of contextual tokens and significant GPU-hours to train; our method requires only a single forward pass per vocabulary token ($\sim$20 minutes, no gradient computation).

A.2 Unsupervised Feature Quality

The 1500 unsupervised features achieve high quality without human labels: mean max-score 0.906 (median 0.908), with 100% scoring ≥ 0.80 and 61.4% scoring ≥ 0.90 on their top-activating tokens.

B Probe and MLP Ablation

To test whether cross-dimension structure provides practical benefit, we train a 2-layer MLP (128 hidden neurons, ReLU) that can learn arbitrary nonlinear dimension combinations. Training data scaled from 50 to 2000 examples per category. 80/20 train/test split, test AUC reported.

Table 21: Probe ablation: does cross-dim structure exist at any training scale? (Qwen 3.5-4B, 175 categories, test AUC).

Train size (pos+neg)	Mean LogReg AUC	Mean MLP AUC	MLP wins ($\Delta > 0.01$)	Mean gap
100 (50+50)	0.9964	0.9931	4/175 (2%)	-0.003
400 (200+200)	0.9977	0.9957	0/175 (0%)	-0.002
1000 (500+500)	0.9991	0.9980	0/175 (0%)	-0.001
2000 (1000+1000)	0.9992	0.9984	0/175 (0%)	-0.001
4000 (2000+2000)	0.9992	0.9987	0/175 (0%)	-0.001

The MLP never outperforms LogReg on any category at any training scale. The gap is consistently negative: additional capacity is a liability, not an asset. Since LogReg converges to axis-aligned weights (no learned rotation) and the MLP with full cross-dim capacity adds nothing, cross-dimension structure provides no practical benefit beyond what per-dim reading already captures.

Hard-negative evaluation. Full-vocabulary AUC could be inflated by class imbalance. We test against semantically adjacent categories as hard negatives (animals vs food/body parts, weapons vs vehicles). Hard-negative AUC remains 0.92–1.0 (mean 0.977).

C Full Category Data

The 175 categories span:

Table 22: Feature strength tiers (Qwen 3.5-4B, h_{24}).

Tier	Representative categories	Count
Strong (≥ 0.85)	physics, country, body, language, electronics	14
Solid (0.80–0.85)	biology, math, metal, emotion, code, conjunction	72
Moderate (0.75–0.80)	weather, religion, fruit, tool, weapon, currency	75
Weak (0.70–0.75)	music_genre, greeting, narrative, sadness, bird	14

Strong features are ontologically discrete (physics terms are fundamentally unlike non-physics terms). Weak features are fine-grained subtypes ("bird" $\subset$ animal), gradient/degree categories ("sadness" as a degree of emotion), or context-dependent ("greeting" depends on pragmatic use). All 175 exceed null calibration.

D Layer Sweep and Per-Head Specialization

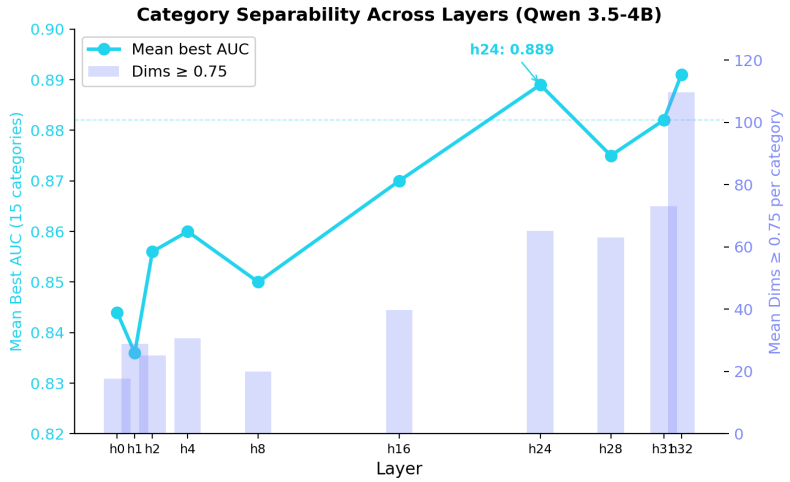

Figure 5: Category separability across layers (Qwen 3.5-4B, 15 categories). Categories are detectable at all layers (AUC $\geq$ 0.83 even at h_0), with separability peaking at h_{24}. The number of dims exceeding the 0.75 threshold increases progressively, reflecting growing feature density through the stack.

Table 23: Category separability across layers (Qwen 3.5-4B, 15 categories).

Layer	Mean Best AUC	Mean Dims ≥ 0.75
h_0	0.844	17.7
h_4	0.860	30.7
h_8	0.850	20.0
h_{16}	0.870	39.7
h_{24}	0.889	65.1
h_{28}	0.875	63.1
h_{32}	0.891	109.7

Categories are discoverable at all layers (AUC ≥ 0.84 everywhere), but feature density increases progressively. Layer 24 provides optimal balance of high AUC with moderate dim count. Layer 32's higher count reflects the prediction transition.

Gemma 3-4B layer sweep (175 categories, full vocabulary).

Table 24: Category discovery across layers (Gemma 3-4B, 175 categories). Gemma shows a U-shaped profile: strong embedding-level features (h_0), a middle-layer dip, and partial recovery at late layers.

Layer	Discoverable ($\geq$0.75)	Mean AUC	Mean Dims
h_0	153/175 (87%)	0.819	3.6
h_4	55/175 (31%)	0.742	0.5
h_8	63/175 (36%)	0.743	0.6
h_{12}	61/175 (35%)	0.739	0.5
h_{16}	50/175 (29%)	0.737	0.5
h_{20}	68/175 (39%)	0.743	0.8
h_{25}	114/175 (65%)	0.763	1.4
h_{28}	107/175 (61%)	0.763	1.4
h_{34}	137/175 (78%)	0.772	1.7

Gemma's embedding layer (h_0) achieves 87% discoverability from surface distributional features (code tokens, function words, grammatical categories). Middle layers (h4–h20) reorganize these representations, temporarily reducing per-dim readability to 29–39%. Late layers restore it, climbing from h_{25} (65%) to peak semantic separability at the final layer h_{34} (78%, mean AUC 0.772)—the layer we use for Gemma discovery. Unlike Qwen, whose peak sits at h_{24} before the prediction transition, Gemma's U-shaped profile places its peak at the final layer; the selection rule (peak category separability) is the same, only the depth differs. All 175 categories exceed null calibration ($p_{95} = 0.68$) at h_{25} onward.

Mistral 7B layer sweep (171 categories, full vocabulary).

Table 25: Category discovery across layers (Mistral 7B, 171 categories). Mistral shows monotonic improvement peaking at h20–h30 with 100% discoverability.

Layer	Discoverable ($\geq$0.75)	Mean AUC	Mean Dims
h_0	96/171 (56%)	0.761	1.8
h_4	167/171 (98%)	0.818	12.6
h_8	165/171 (96%)	0.807	11.4
h_{12}	168/171 (98%)	0.814	12.8
h_{16}	170/171 (99%)	0.824	25.3
h_{20}	170/171 (99%)	0.843	42.3
h_{24}	171/171 (100%)	0.842	43.6
h_{28}	169/171 (99%)	0.843	46.7
h_{30}	171/171 (100%)	0.840	47.2
h_{32}	171/171 (100%)	0.820	34.8

Mistral shows near-perfect discoverability from h_4 onward (96–100%), with mean AUC peaking at 0.843 in the h20–h28 range. The larger hidden dimension ($D = 4096$) provides more per-dim resolution throughout the stack, explaining Mistral's consistently higher AUC compared to the 2560-dim models.

Cross-layer Jaccard overlap. Mean Jaccard similarity of feature dim sets between non-adjacent layers is 0.042 (chance level for sparse subsets of 2560 dims). Dim assignments are layer-specific; the same physical dimension encodes different categories at different layers.

Per-head specialization (Gemma 3-4B, layer 25). The 4 attention heads show clear specialization:

- Head 1 (K): 80 feature dims, concentrated in negation (20), number (14), color (10)
- Head 1 (V): 91 feature dims, concentrated in animal (9), negation (16), emotion (14)
- Feature load varies $2\times$ across heads

E Random-Init Control

To verify that discovered features require training, we run identical discovery procedures on randomly initialized models (same architecture, random weights, seed=42).

Table 26: Random-init control: feature discovery on untrained models.

Model	Random AUC	Trained AUC	Discoverable ($\geq$0.75)	Unsupervised yield
Qwen 3.5-4B	0.608	0.801	0/175	0/200
Gemma 3-4B	0.603	0.763	0/175	0/200
Mistral 7B	0.679	0.844	11/175	1/200 (incoherent)

Random-init models produce no monosemantic, semantically coherent features. The residual AUC above 0.50 reflects input embedding similarity that passes through random projections (stronger for Mistral's smaller 32K vocabulary). The one Mistral "feature" (5 dims, tokens: "widely", "itched", "installed", "EqualTo", "adding") is semantically incoherent and fires on 3.4% of vocabulary (not monosemantic).

The trained-vs-random visualization (Figure 2) provides a complementary view: both produce an expanding envelope, but random weights show smooth uniform expansion while trained weights produce the characteristic pinch and asymmetric internal organization that enables per-dim feature encoding.

F Toy Superposition Model

To compare the toy regime against real transformers (§5), we replicate Elhage et al.'s toy autoencoder ($x \mapsto W^\top W x$ with ReLU, bottleneck $d \ll n$, sparsity s) and test whether per-dim sign patterns recover the "superposed" features.

Table 27: Toy superposition replication. The model encodes more features than dimensions (superposition emerges). Sign detection = fraction of represented features recoverable from per-dim sign patterns in the hidden state (AUC > 0.7), without any learned decoder or geometric projection.

Config	Features (n)	Bottleneck (d)	Represented	Sign detection	Comb. capacity (3^d-1)
s=0.05	20	5	10	80% (8/10)	242
s=0.02	40	5	6	100% (6/6)	242
s=0.01	100	5	7	71% (5/7)	242
s=0.05	100	20	24	96% (23/24)	3.5×10^9
s=0.02	200	20	22	100% (22/22)	3.5×10^9

Despite non-orthogonal encoding columns (confirming superposition), 80–100% of represented features are recoverable from sign patterns alone. The combinatorial capacity column shows that even with $d = 5$ dimensions, 242 sign-pattern addresses are available, far exceeding the 6–10 features the model actually represents. At $d = 20$, the address space (3.5×10^9) makes geometric packing entirely unnecessary.

This shows sign matching is a sufficient decoder of the toy's features. To characterize what the hidden state itself stores, we measure two further properties below.

Sign-vs-magnitude and cross-dim coupling in the toy. The detection result above shows sign matching identifies which feature is active, but does not characterize how the toy's hidden state stores content. We measure two further properties: sign-only reconstruction quality (forcing $|h_d|$ to the row-mean magnitude and passing through $W^\top$), and pairwise mutual information between dim signs in h (Eq. 6, same procedure as §3.2).

Table 28: Toy hidden-state analysis: sign-only reconstruction quality and pairwise MI between dim signs. Sign-only WMSE rescales each row of sign(h) to its mean $|h|$ before decoding. Real-transformer cross-dim MI (Table 3): 0.0004–0.005 bits.

Config	WMSE full	WMSE sign-only	Sign/full ratio	Mean MI (bits)	Max MI
n=20, d=5, s=0.05	0.0112	0.0148	1.3×	0.062	0.167
n=40, d=5, s=0.02	0.0039	0.0061	1.6×	0.095	0.198
n=100, d=5, s=0.01	0.0016	0.0028	1.7×	0.045	0.119
n=100, d=20, s=0.05	0.00086	0.0098	11.4×	0.015	0.150
n=200, d=20, s=0.02	0.00013	0.0062	47.4×	**0.005**	0.045

Two patterns emerge. First, sign-only reconstruction at the bottleneck regime (d=5) reaches 60–77% of full reconstruction quality (sign/full WMSE ratio 1.3–1.7×), the same sign-carries-content / magnitude-refines pattern observed in real LMs (Tables 1–2: sign alone preserves 49–63% top-1 and 72–93% top-5). The toy and real transformers agree on the encoding: signs carry content, magnitudes carry confidence. The large sign/full ratios at d=20 reflect tiny absolute errors (full WMSE near zero); sign-only reconstruction remains accurate in absolute terms.

Second, cross-dim MI distinguishes the regimes. The toy bottleneck at d=5 shows mean pairwise MI of 0.045–0.095 bits, 10–50× higher than real transformers (Table 3: 0.0004–0.005 bits). As the bottleneck relaxes (n=200, d=20, s=0.02), toy MI drops to 0.005 bits, entering real-transformer territory. The geometric coupling that defines superposition is a function of bottleneck pressure, not a fixed property of the architecture.

G Cross-Category Polysemy: Case Detail

The 77 cross-category polysemy cases (§3.4) each pair a polysemous word with a category-sense sentence and an other-sense sentence; we score the target word's contextual representation against the word's category prototype in both. "Correct" means the category-sense context scores higher. The target token is verified to be located in both sentences; cases failing this check are excluded (2 on Qwen, varies by tokenizer). We report all cases without selection.

Table 29: Representative cross-category cases (Qwen 3.5-4B, τ=0.70). Score = fraction of prototype dims whose contextual sign matches the category polarity.

Word	Category	Cat-sense	Other-sense	Δ
train	vehicle	0.931	0.448	+0.483
jeep	vehicle_land	0.783	0.217	+0.565
owl	bird	0.938	0.562	+0.375
pine	tree	0.935	0.613	+0.323
date	fruit	0.875	0.583	+0.292
seal	animal	0.750	0.525	+0.225
Representative failures (reported, not excluded):				
copper	metal	0.780	0.923	−0.143
calf	body	0.628	0.752	−0.124
whip	weapon	0.667	0.889	−0.222

Strong, ontologically distinct pairs (vehicle, tree, bird) separate cleanly; failures concentrate in categories whose senses are both concrete and adjacent (body parts, metals). Full per-case data is in the released results file.

Threshold robustness. The effect is not a tuned τ: accuracy and AUC are stable across $\tau \in [0.55, 0.70]$ and increase as the prototype broadens (lower τ). Peak pooled AUC: Qwen 0.768 (τ=0.55), Mistral 0.746 (τ=0.55), Gemma 0.738 (τ=0.58).

Qwen3-32B layer sweep. Because Qwen3-32B is near chance at its detection-optimal layer (h_{48}), we tested whether any other layer reads cross-category sense, building a separate type-cache and reading context at each. It does not: pooled AUC stays at 0.49–0.71 across all layers and thresholds, never approaching the 0.72–0.77 of the other models. Best pooled AUC per layer (at the AUC-maximizing τ):

Table 30: Qwen3-32B cross-category polysemy by layer (best pooled AUC over $\tau \in [0.55, 0.72]$ at full case coverage, n=75; higher-τ settings that drop cases are excluded). No layer reaches the 0.72–0.77 of the 4B–7B models.

Layer	best pooled AUC	at τ	accuracy
h_{40}	0.644	0.68	57%
h_{48}	0.610	0.65	55%
h_{54}	0.671	0.58	60%
h_{60}	0.616	0.62	61%

This is consistent with the low dimension-to-vocabulary ratio of Qwen3-32B (D/V=0.034) limiting per-dim category sharpness (Limitation 4); the weakness is layer-independent, not a matter of choosing the wrong readout depth.

H Causal Sign-Flip: Per-Category Detail

This appendix provides the per-category numbers behind the causal intervention summarized in §3.4. For each model and category we build the feature's sign prototype from the type cache, then during a live forward pass flip the prototype's signs at all positions and report the change in the mean logit of the category's target tokens (greedy decoding). Patch layers: Qwen 3.5-4B h_{24}, Gemma 3-4B h_{34}, Mistral 7B h_{24}, Qwen3-32B h_{60} (the near-output layer required by the depth-dependence of §5); the Gemma flip is applied at block 33, whose FFN writes the h_{34} residual used for detection (block $L-1$ writes h_L).

Table 31: Per-category target-logit change under sign-flip (full coalition, τ=0.6). **Away**: signs flipped away from expected. **Toward**: same dimensions and magnitudes forced to the expected sign (isolates sign from magnitude). **Random**: equal number of random dimensions. Away suppresses on every category that responds; toward and random do not.

Model	Category	Away	Toward	Random
Qwen 3.5-4B	animals	−11.55	+0.81	−2.24
	numbers	−13.90	−0.42	−2.93
	colors	−11.81	−2.68	−3.24
	food	−10.10	+0.73	+0.03
	countries	−10.22	−1.17	−0.98
Gemma 3-4B	animals	−20.32	+2.07	−3.59
	numbers	−19.11	+0.73	−3.31
	colors	−24.27	+2.12	−4.12
	food	−20.71	+1.73	−1.97
	countries	−24.46	+1.92	−4.32
Mistral 7B	animals‡	−1.86	−0.70	−1.19
	numbers	−7.16	−1.64	−2.27
	colors	−6.99	−0.74	−1.13
	food	−5.10	−0.21	−1.25
	countries	−6.23	−0.68	−3.94
Qwen3-32B	animals	−5.10	+0.29	−0.94
	numbers	−8.98	−2.31	−2.14
	colors	−7.38	−1.84	−0.17
	food	−6.12	−0.30	−0.03
	countries‡	+1.20	−0.97	−2.15

‡Each model has one weak category that does not respond (animals/Mistral, countries/Qwen3-32B); 4/5 categories respond strongly on every model.

Concept-specificity via the disjoint control. A natural concern is that flipping any large trained sign-coalition might degrade the model's general next-token competence rather than suppressing the specific concept. It does not. Because features share dimensions ($\sim$35% overlap), flipping a *different* concept's coalition partially flips the target's; the apparent cross-concept damage is entirely this overlap. Removing the shared dimensions—flipping only the part of another concept's coalition that is disjoint from the target—leaves the target essentially untouched, indistinguishable from a random flip of equal size.

Table 32: Disjoint specificity control (target-logit change on concept A). **A (full)**: flip A's own coalition. **Disjoint**: flip another concept's coalition with $A \cap B$ removed. **Random**: equal-size random flip. Disjoint $\approx$ random: the causal effect is concept-specific, not general damage.

Model	Concept A	A (full)	Disjoint	Random
Qwen 3.5-4B	animals	-11.55	-0.34	-0.65
	numbers	-13.90	-0.60	-2.39
	colors	-11.81	-0.17	-3.18
Gemma 3-4B	animals	-20.32	-3.47	-3.16
	colors	-24.27	-1.45	-2.27
	countries	-24.46	-1.77	-1.37
Mistral 7B	numbers	-7.16	-0.25	$+0.04$
	colors	-6.99	-0.41	-0.21
	food	-5.10	-0.13	-0.27
Qwen3-32B	numbers	-8.98	-0.11	-0.45
	colors	-7.38	-0.77	$+0.06$
	food	-6.12	-0.02	-2.08

Across all four models, flipping a disjoint coalition produces an effect statistically indistinguishable from a random flip, while flipping the concept's own coalition suppresses it by 5–24 logits. The causal effect is therefore specific to the feature, and the coalition—not any single dimension—is the unit at which it acts.