The Concept Assembly Zone

Tracking How Concepts Form Across Transformer Depth

James Henry *Independent Researcher* jamesrahenry@henrynet.ca

April 5, 2026

Abstract

Mechanistic interpretability methods commonly extract concept representations by identifying the single optimal layer of a Transformer's residual stream where class separation peaks. This "best layer" heuristic is computationally efficient and empirically grounded, but it captures a snapshot of a process rather than the process itself. We introduce the **Concept Assembly Zone** (CAZ): a layer-level event where the model allocates geometric directions in activation space to serve one or more concepts. We formalize the CAZ through three layer-wise metrics — Separation, Concept Coherence, and Concept Velocity — and derive principled methods for identifying CAZ boundaries without manual layer sweeps. A CAZ is not a concept: it is the computational event where the model organizes its geometry to make a concept measurably separable. Multiple concepts may share a CAZ, and a single concept typically participates in multiple CAZes across depth. Empirical validation across 30 models from 7 architectural families and 7 concepts reveals that the separation curve S(l) is frequently **multimodal**, with a scored detection method uncovering an additional category of subtle assembly events ("gentle CAZes") that are invisible to standard peak detection but are causally active in 95% of cases. The framework generates seven testable predictions, of which five have been empirically evaluated, and provides a reference implementation in the open-source rosetta_tools library (v1.0.0).

---

1. Introduction

The dominant paradigm in mechanistic interpretability extracts concept representations by identifying the single "best layer"—the residual stream depth at which a linear probe or difference-of-means (DoM) vector achieves maximum class separation [Zou et al., 2023; Arditi et al., 2024]. This heuristic is computationally convenient and empirically grounded. It is also, by design, a snapshot: it identifies the peak of a process rather than characterizing the process itself.

Transformers are iterative dynamical systems. Each layer applies a sequence of attention and MLP operations that *write* new information into the residual stream, modifying and extending what prior layers contributed [Elhage et al., 2021]. A concept observed at Layer 15 was shaped by Layers 10 through 14 before it; the best-layer heuristic tells us where the concept is most legible, not how it arrived there.

This paper introduces the **Concept Assembly Zone** (CAZ) framework, which extends the interpretability toolkit from anatomy—*where is the concept most visible?*—to dynamical flow—*how does the concept form?* The CAZ is defined as a layer-level event

where the model allocates geometric directions in activation space such that a semantic concept becomes measurably separable. A CAZ is not the concept itself — each CAZ is a location in the layers where the model's geometry expresses influence to serve a concept.

The framework has immediate practical implications:

1. **Richer extraction.** CAZ-windowed extraction methods may capture information present in the assembly dynamics that single-layer methods do not.
2. **Principled intervention.** Ablation at different points in the CAZ chain produces qualitatively different effects. The framework provides a geometric basis for selecting intervention depth.
3. **Dark matter connection.** The structured residual left unexplained by sparse autoencoders [Engels et al., 2024] may partially correspond to in-progress concept construction within CAZes — transitional representations that resist linear decomposition at any single layer.
4. **Cross-model transfer.** Concept directions extracted at a CAZ in one model can be aligned to equivalent directions in another model via a single rotation, enabling transferable interpretability tooling across architectures.
5. **Understanding alignment training.** CAZ profiles provide a lens for studying what preference optimization changes in a model — not whether concepts exist, but where and how they are assembled.
6. **Concept inventory.** By tracking which geometric directions are allocated at each CAZ and which remain unaligned with any human concept probe, the framework provides a systematic approach to cataloguing what a model computes — both the named and the unnamed.

The CAZ framework generates specific, falsifiable predictions. Section 4 states seven such predictions; Section 6 reports empirical results for five of them across 30 models and 7 architectural families. The reference implementation is provided as rosetta_tools [Henry, 2026], an open-source Python library. We are explicit about the assumptions the framework inherits from the broader interpretability literature.

---

2. Background

2.1 The Residual Stream and Concept Representation

The residual stream formulation [Elhage et al., 2021] treats each layer's output as an additive contribution to a shared communication channel. Attention heads and MLPs read from and write to this stream; the final residual vector is projected onto the unembedding matrix to produce logits. This architecture makes layer-by-layer tracking of concept geometry natural: we can ask, at each layer l, how well the current residual stream separates two contrastive classes.

2.2 Difference-of-Means and Linear Artificial Tomography

DoM extracts a concept direction V_concept ∈ ℝ^d as the normalized difference between class-conditional mean activations at the chosen layer [Zou et al., 2023]. Linear

Artificial Tomography (LAT) uses a similar contrastive approach. Both methods produce a single vector at a single depth—a precise and useful representation of where the concept is most geometrically legible. The CAZ framework asks what additional information about concept formation might be recoverable from the layers surrounding that peak.

2.3 Abliteration and Intervention Depth

Arditi et al. [2024] demonstrated that refusal behavior across 13 open-source models is mediated by a single direction removable via weight orthogonalization ("abliteration"). Independent replications have observed KL divergences between abliterated and unmodified models ranging from 3.16 to 5.71 — suggesting that while behavioral suppression is effective, the intervention also affects general model capabilities. The CAZ framework motivates a systematic study of intervention depth: rather than selecting a single layer by hyperparameter search, the CAZ profile identifies where concepts are being actively constructed versus where they are established, offering a principled basis for choosing intervention points (Section 4.1).

2.4 The Emerging Geometric Program

Gurnee et al. [2025] demonstrated that character counts are represented on low-dimensional curved helical manifolds in the residual stream, with attention heads performing geometric transformations on these structures. Engels et al. [2025] found circular multi-dimensional representations for temporal concepts (days, months, years) that are not decomposable into independent one-dimensional SAE features. Wollschläger et al. [2025] showed that refusal occupies multi-dimensional polyhedral concept cones with multiple independent directions. These findings establish a growing body of evidence for rich geometric structure in activation space, and have explicitly called for unsupervised methods to detect and characterize it. The CAZ framework is designed to complement this geometric program by providing a layer-indexed account of when such structures crystallize.

---

3. The Concept Assembly Zone

A CAZ is not a concept. It is a layer-level computational event where the model allocates geometric directions in activation space to serve one or more concepts. The concept is the human label we project onto the geometry; the CAZ is the machinery the model uses to organize that geometry. A single CAZ may host multiple concepts simultaneously (48% do), and a single concept typically participates in multiple CAZes across depth (mean 3.4 per concept per model). The distinction matters: when we say "credibility has a CAZ at layer 10," we mean that layer 10 is where the model allocates a geometric direction that, when measured with a credibility probe, shows strong class separation — not that layer 10 "contains" credibility.

Terminology

Term	Definition
CAZ	A location in the layers where the model's geometry expresses influence to serve a concept
CAZ Profile	The full sequence of CAZes for one concept in one model
Black hole	Dominant CAZ with score > 0.5 — strong, concentrated assembly event
Gentle CAZ	Subtle CAZ with score < 0.05 — causally active but invisible to standard detection
Embedding CAZ	CAZ at the embedding boundary, driven by token-level features rather than transformer computation
Active CAZ	CAZ within the transformer layers, driven by attention and MLP computation. Active CAZes are the primary subject of this framework
CAZ score	Composite metric: prominence × coherence boost × √width
Separation S(l)	Fisher-normalized centroid distance between contrastive classes at layer l
Coherence C(l)	Explained variance ratio of the primary separating direction at layer l
Velocity v(l)	Smoothed rate of change of S(l) across layers

3.1 Concept Lifecycle

By tracking the residual stream across model depth, concept formation is empirically distinguishable as a sequence of allocation events rather than a single assembly followed by decay.

Early layers (Context and Syntax)

In early layers, the residual stream primarily resolves local context, grammar, and surface token relationships. Projecting contrastive datasets into this space generally produces heavily entangled activations; the separation metric is near zero. The model has not yet committed to a semantic trajectory.

However, the scored detector does find CAZes at layers 0–1 in some models (36 cases across the 30-model validation set). Most are gentle (score < 0.1) and reflect **embedding leakage** — concept-associated tokens having distinctive embeddings that create passive separation before any transformer processing occurs. We term these **embedding CAZes** to distinguish them from **active CAZes** where the model's attention and MLP computations allocate geometry to serve a concept. The embedding CAZ signal is a property of the tokenizer and training corpus, not a computational decision by the model. Practitioners using CAZ profiles should be aware that embedding CAZes may not respond to the same interventions as active CAZes.

CAZ Chain (Concept Allocation)

As the residual stream deepens, geometric directions are allocated to concepts at discrete layer-level events — the CAZes. A persistent direction in activation space may serve one concept at shallow layers, be reallocated to a different concept at a mid-depth CAZ, and reallocated again at a deeper one. Empirically, 48% of CAZ layers host 2+ concepts peaking simultaneously, and the most universal features across architectures are those that rotate through multiple concept alignments across depth (Section 6). For a given concept, the separation metric S(l) reflects when that concept holds a strong claim on one or more geometric directions. A single concept typically participates in multiple CAZes (mean 3.4 per concept per model under scored detection), ranging from dominant events ("black holes," score > 0.5) to subtle allocation events ("gentle CAZes," score < 0.05) that are nonetheless causally active in 95% of cases.

Late layers (Logit Projection)

In the final layers, the model transitions from abstract representation to concrete next-token prediction. The residual stream is projected toward the unembedding matrix. Concept geometry may degrade or re-entangle as abstract directions give way to vocabulary-specific structure. This is the only phase where separation genuinely decays — earlier apparent "decay" between CAZ peaks is better understood as reallocation of the direction to a different concept.

3.2 Layer-Wise Metrics

Let h_l^(i) ∈ ℝ^d be the residual stream activation at layer l for sample i, and let A, B be contrastive classes with conditional means $\bar{h}$_A^(l), $\bar{h}$_B^(l) and within-class covariance matrices Σ_A^(l), Σ_B^(l).

Separation Metric

We define the separation at layer l using a Fisher-normalized criterion [Bishop, 2006, §4.1.4]:

$$S(l) = ||\bar{h}\_A^{(l)} - \bar{h}\_B^{(l)}||_2 \;/\; \sqrt{[(1/2)(\mathrm{tr}(\Sigma\_A^{(l)}) + \mathrm{tr}(\Sigma\_B^{(l)}))]}$$

In plain terms: separation asks "if I gave you a sentence and asked whether it expresses credibility or not, how easily could you tell from the model's internal state at this layer?" A high S(l) means the model's activations for credible and non-credible text have moved far apart relative to how spread out each group is. A low S(l) means the two groups are still jumbled together. Tracking S(l) across layers reveals where the model begins to "make up its mind" about a concept.

Raw centroid distance is misleading when cluster dispersion varies across layers. Early layers tend toward diffuse, high-variance representations; normalization by within-class spread corrects for this. Mahalanobis distance [Mahalanobis, 1936] would account for full covariance structure but is numerically unstable without regularization in high-dimensional activation spaces. Fisher normalization provides the appropriate tradeoff between geometric fidelity and computational feasibility for initial experiments.

Concept Coherence

Separation alone is insufficient: two classes could exhibit identical centroid separation while one forms a tight cluster and the other a diffuse cloud. We track Concept Coherence as the explained variance ratio of the first principal component of the between-class direction at each layer:

$$C(l) = \lambda_1^{(l)} / \Sigma_i \lambda_i^{(l)}$$

where $\lambda_i^{(l)}$ are the eigenvalues of the pooled activation covariance at layer l, projected onto the contrastive subspace. A concept is *well-formed* when both S(l) and C(l) are high: the classes are far apart *and* the separating direction is geometrically clean.

In plain terms: coherence asks "is the concept encoded as a single clean direction, or is it smeared across many dimensions?" A high C(l) means the model has committed to one dominant direction for separating the classes — the concept has crystallized into a sharp geometric feature. A low C(l) means the separation exists but is spread across multiple directions, making the concept harder to extract with a single vector. Coherence distinguishes a concept that is clearly formed from one that is still diffuse.

Concept Velocity

To identify CAZ boundaries, we compute the rate of geometric divergence between layers. Because raw layer-to-layer differences are noisy, we apply a smoothed estimate:

$$v_{concept}(l) = (1/(2k+1)) \Sigma_{j=l-k}^{(l+k)} [S(j) - S(j-1)]$$

where k is the smoothing half-window. A practical heuristic is $k = \lfloor L/24 \rfloor$, where L is total model depth, yielding k=1 for 12–24 layer models, k=2 for 48-layer models, and k=3 for 72-layer models. This scales the smoothing window proportionally to model depth and prevents false CAZ boundary detection from single anomalous layers. The appropriate value of k should ultimately be determined empirically — for models where ground-truth concept boundaries can be established via ablation, the k value that maximizes boundary prediction accuracy is preferred.

In plain terms: velocity asks "is the concept forming right now, or has it already formed?" Positive velocity means separation is increasing — the model is actively constructing the concept at this layer. Negative velocity means separation is decreasing — the concept is being degraded or reallocated. Zero velocity means nothing is changing. The velocity curve marks the boundaries of the CAZ: it goes positive when assembly begins and negative when it ends.

3.3 CAZ Boundary Detection

Single-Region Detection (Velocity-Based) When the S(l) curve is unimodal, CAZ boundaries are derived from the velocity profile:

- **CAZ Entry** (l_start): The first layer where v_concept(l) exceeds a sustained positive threshold θ_+.
- **CAZ Peak** (l_max): The layer where S(l) reaches its absolute maximum. This corresponds to the "best layer" of conventional interpretability.

- **CAZ Exit** (l_end): The layer where v_concept(l) becomes consistently negative, marking the onset of post-CAZ degradation.

The conventional best-layer heuristic extracts V_concept at l_max. CAZ-aware extraction uses the full interval [l_start, l_end].

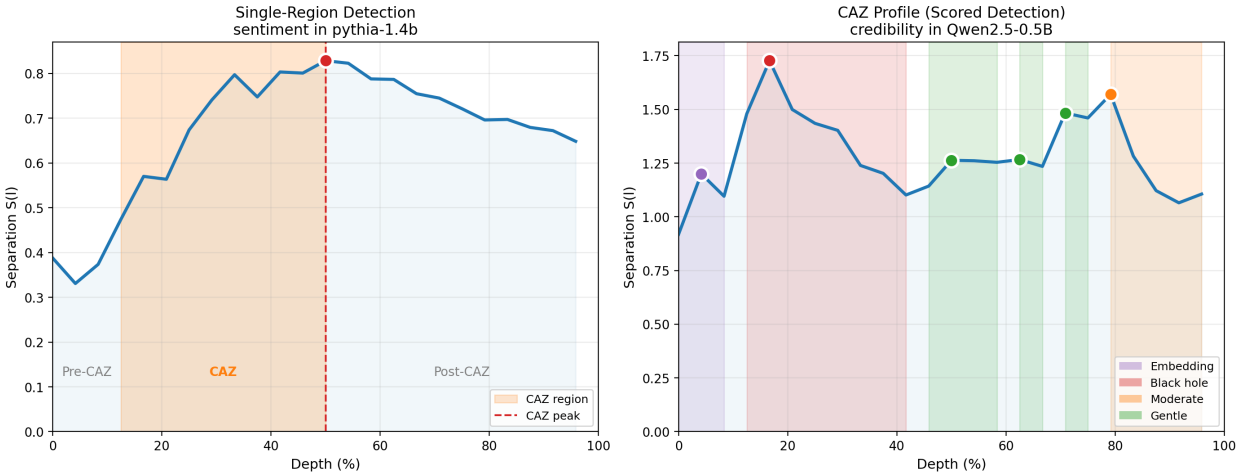

Figure 1: Figure 1: Single-region detection (left) versus scored CAZ profile (right). The single-region detector identifies one assembly zone for sentiment in Pythia-1.4b. The scored detector reveals six CAZes for credibility in Qwen2.5-0.5B, including embedding, black hole, moderate, and gentle events.

Multi-Region Detection (CAZ Profiles) Empirical analysis across 30 models reveals that the S(l) curve is frequently **multimodal** — a single concept can produce multiple significant local maxima at different depths. In these cases, the velocity-based boundary detector wraps a single contiguous zone around the global maximum and is blind to secondary peaks.

The **CAZ Profile** generalizes the single-region CAZ to a sequence of assembly regions:

1. Detect all significant local maxima in S(l) using prominence-based peak detection.
2. Identify **saddle points** — the local minima between consecutive peaks — as natural region boundaries.
3. Each region spans from one saddle to the next, with the first region starting at layer 0 and the last ending at the final layer.

A CAZ Profile is characterized by: - **n_regions**: Number of distinct assembly regions (1 = unimodal, 2+ = multimodal) - **dominant region**: The region with the highest peak separation - Per-region: start, peak, end, width, peak separation, coherence, rise/fall asymmetry

Scored Detection The prominence-based detector requires a threshold to determine which peaks are significant. A fixed threshold (e.g., 10% of global max separation) is arbitrary and risks discarding subtle but causally active assembly events. We replace it with a composite scoring system:

```
CAZ score = prominence × coherence boost × √width
```

with a 0.5% prominence floor. This scores each detected region on a continuous scale rather than applying a binary keep/discard decision. The score naturally separates dominant assembly events (black holes, score > 0.5) from subtle ones (gentle CAZes, score < 0.05) while retaining both for analysis. The scored detector is the primary detection method; the fixed-threshold detector is retained for backward compatibility.

3.4 Multi-Layer Concept Extraction

Three extraction methods should be compared empirically:

1. **Delta PCA**: PCA on layer-to-layer residual deltas Δh_l = h_l - h_(l-1) within [l_start, l_end]. This captures what each layer *adds*—the construction process itself.
2. **Windowed PCA**: PCA on raw activations h_l across [l_start, l_end]. This captures the cumulative concept direction as it evolves through the assembly zone.
3. **Single-layer (baseline)**: The standard DoM vector at l_max.

If methods (1) or (2) consistently outperform (3) on downstream steering and classification tasks, this validates the dynamical framing as more than descriptive vocabulary. If they do not, the framework may still provide useful theoretical structure while failing to improve practical extraction—which is itself worth establishing.

3.5 Sub-Representations

When a concept's S(l) curve is multimodal, the dom_vector (first principal component of contrastive activations) at each peak defines a distinct linear direction. Empirical measurement shows these directions are **geometrically distinct**: across all multimodal concept × model pairs, the cosine similarity between the shallow and deep peak dom_vectors averages 0.2–0.4. The two peaks are not the same feature at different amplitudes — they are different linear features that both happen to separate the same contrastive classes.

This implies that a single human concept label ("credibility", "negation") maps to **multiple sub-representations** at different processing depths. Interpretive evidence suggests:

- **Shallow sub-representations** form near the embedding layers, likely driven by lexical cues (concept-associated words).
- **Deep sub-representations** form in mid-to-late layers, likely driven by compositional processing (contextual inference, scope, pragmatic reasoning).

The transition between sub-representations at the saddle point is abrupt, not gradual: layer×layer cosine similarity matrices show block-diagonal structure, and adjacent-layer cosine similarity dips sharply at the saddle point (to as low as 0.35 in some models), indicating a phase transition between distinct encoding regimes.

3.6 Depth-Stratified Convergence

Cross-architecture alignment confirms that sub-representations are **independently universal**: depth-matched alignment (shallow↔shallow, deep↔deep) significantly exceeds cross-depth alignment (shallow↔deep, deep↔shallow) across all 6 tested concepts ($p < 0.01$ for each). Deep↔deep alignment is consistently the strongest, suggesting the compositional sub-representation is the most universal feature that models converge on.

This refines the Platonic Representation Hypothesis [Huh et al., 2024]: representational convergence is not monolithic but **stratified by processing depth**, with each stage of concept assembly converging independently across architectures. Full alignment results are reported in the companion validation paper [Henry, 2026b].

---

4. Testable Predictions

The CAZ framework generates seven predictions that are in principle falsifiable with existing open-weight models and standard interpretability tooling.

4.1 Optimal Ablation Depth

Prediction 1: The suppression-to-damage ratio varies systematically with intervention depth relative to the CAZ.

For a given concept, extract the concept direction at each layer via DoM. Apply orthogonal projection at each layer l independently. Measure: (a) behavioral suppression rate on targeted prompts, (b) KL divergence from the unmodified model on unrelated prompts. Plot the ratio (a)/(b) as a function of layer.

The original formulation predicted that mid-CAZ ablation would produce the best ratio — intervening during assembly to allow downstream layers to route around the missing information. Empirical results [Henry, 2026b] show a more nuanced picture: in models with redundant encoding (Pythia, GPT-2, OPT), the concept direction persists in the residual stream at all layers post-assembly, so suppression is layer-invariant while capability damage is lowest at late layers. The optimal intervention point is therefore *post-CAZ*, not mid-CAZ. In models with sparse encoding (Qwen, Gemma), layer specificity is greater and the CAZ location matters more. The prediction is **revised**: optimal ablation depth depends on the model's encoding strategy, and the CAZ profile identifies which strategy applies.

4.2 Architecture-Stable CAZ Positioning

Prediction 2: CAZ boundaries are concept-specific but architecture-stable.

Different concepts should have different CAZ windows within the same model. However, the *relative* ordering of those windows — as a fraction of total model depth — should be consistent across architectures. Absolute depth percentages may be family-specific, but relative concept ordering should be universal.

Status: Confirmed for relative ordering across 7 architectural families [Henry, 2026b].

4.3 CAZ Width and Concept Abstraction

Prediction 3: CAZ width correlates with concept abstraction level.

More abstract concepts (e.g., "trustworthiness," "moral valence") should have wider CAZ windows than concrete ones (e.g., "negation," "plurality"), because abstract concepts require more iterative construction across attention layers. This is testable by comparing l_end - l_start for concepts at different levels of semantic abstraction as operationalized by, for example, depth in WordNet or scores on standard concreteness rating datasets.

Status: Initial support. Affective and epistemic concepts show wider CAZs than relational and syntactic concepts [Henry, 2026b].

4.4 Post-CAZ Degradation as Logit Interference

Prediction 4: Post-CAZ re-entanglement correlates with unembedding matrix structure.

The degradation of clean concept geometry in late layers is not noise but a structural consequence of preparing the residual stream for logit projection. Concepts whose associated vocabulary tokens are distributionally similar in the unembedding space—close in embedding distance—should show more post-CAZ degradation than concepts with distributionally distinct vocabulary. This would explain why some concepts retain clean geometry into late layers (their vocabulary is well-separated) while others degrade early (their vocabulary clusters).

Status: Not yet tested. Structural analysis shows post-CAZ decay is gentle, but the correlation with unembedding structure has not been directly measured.

4.5 Depth-Stratified Representational Convergence

Prediction 5: Cross-architecture alignment is depth-matched.

When a concept has multiple assembly regions, the sub-representation at a given processing depth should align more strongly with the corresponding-depth sub-representation in other architectures than with a different-depth sub-representation. Specifically, after Procrustes rotation, cosine(shallow_A, shallow_B) > cosine(shallow_A, deep_B) and cosine(deep_A, deep_B) > cosine(deep_A, shallow_B).

Status: **Confirmed.** Tested across 6 concepts and 56 model pairs. Depth-matched alignment significantly exceeds mismatched for all 6 concepts ($p < 0.01$ for each). Full results in the companion validation paper [Henry, 2026b].

4.6 Lexical vs. Compositional Sub-Representations

Prediction 6: Shallow peaks encode lexical features; deep peaks encode compositional features.

The dom_vector at the shallow assembly peak should correlate with token embedding vectors for concept-associated words (e.g., "reliable", "dubious" for credibility). The dom_vector at the deep peak should show lower correlation with token embeddings and higher dependence on multi-token contextual patterns.

Status: **Not supported by initial test.** Token embedding probing (cosine similarity between peak dom_vectors and concept-relevant token embeddings) yields near-zero values (~0.02) at both peaks, with no significant difference (Wilcoxon p = 0.82). Neither peak resembles raw token embeddings. The lexical/compositional distinction may operate at a higher level of abstraction than direct embedding alignment — the shallow feature could depend on token identity through multi-layer composition rather than literally pointing toward any single token's embedding vector. Requires alternative experimental designs: per-token position attribution, attention knockout, or probing classifiers trained at each peak.

4.7 Multi-Modality as Architectural Property

Prediction 7: Multi-modality prevalence is determined by architecture, not scale.

The fraction of concepts showing multimodal S(l) curves should vary more between architectural families (attention mechanism, activation function, training data) than between scales within a family.

Status: **Supported with nuance.** Multi-modality does not correlate with model parameter count ($\rho = 0.11$, $p = 0.63$) but varies dramatically by family: Qwen 2.5 shows deep, prominent bimodality (valley depths of 26–36% between peaks); Gemma 2 shows subtle structure (valley depths of 8–15%) that falls below the 10% prominence threshold used for peak detection. The architectural difference is in degree of sub-representation separation, not binary presence/absence.

---

5. Relationship to Existing Work

The CAZ framework operates in the same space as several established interpretability methods. Each captures different aspects of model internals; the CAZ contribution is the layer-indexed dynamical view — tracking *when* representations form, not just *what* they are.

5.1 Methodological Context

Sparse Autoencoders (SAEs) decompose activations at a fixed layer into interpretable monosemantic features [Cunningham et al., 2023; Bricken et al., 2023]. SAEs answer "what features exist at layer L?" The CAZ framework answers "how does the feature at layer L relate to the feature at layers L−1 and L+1?" The approaches are complementary: SAE features at a given layer are a snapshot; CAZ

tracking reveals which snapshots are part of the same evolving computation. Engels et al. [2024] found that SAE "dark matter" — structured residual that resists linear decomposition — accounts for roughly 50% of the error vector. The CAZ framework offers a candidate mechanism: in-progress concept construction within a CAZ produces transitional representations that are neither the input feature nor the output feature. These transitional states may be precisely what resists decomposition at any single layer.

Centered Kernel Alignment (CKA) [Kornblith et al., 2019] measures representational similarity between layers or models by comparing activation kernel matrices. CKA provides a global similarity score between two representation spaces but does not identify *which* features are shared or how they evolve across depth. The CAZ framework tracks individual concept directions layer by layer, producing a trajectory rather than a scalar comparison. CKA could serve as a complementary validation tool — high CKA between two layers would be expected within a CAZ (the representation is being refined, not replaced) and low CKA at a saddle point between CAZs (the representation is being reallocated).

Linear probing [Belinkov, 2022; Alain & Bengio, 2017] trains classifiers on activations at each layer to measure concept presence. Probing accuracy curves are closely related to the separation metric S(l) — both measure how distinguishable two classes are at a given depth. The CAZ framework adds the velocity metric (rate of change of separation) and the coherence metric (geometric quality of the separating direction), which together identify not just where a concept is present but where it is actively being constructed. Probing also requires training a classifier per layer; the CAZ metrics are computed directly from activation statistics.

Representation Engineering (RepE) [Zou et al., 2023] extracts concept directions for honesty, morality, power-seeking, and related concepts via contrastive stimuli, then uses those directions for monitoring and steering. The CAZ framework can directly inform RepE's operational decisions. First, the CAZ profile identifies where a concept is being actively constructed versus where it is established, providing a principled basis for choosing intervention depth. Second, the model's encoding strategy determines whether layer selection matters at all: in models with redundant encoding, the concept direction persists across all post-CAZ layers, so RepE can intervene anywhere with equivalent effect. In models with sparse encoding, the CAZ peak is the critical intervention point and intervening elsewhere may miss the target. Third, the scored CAZ detector reveals that a concept may have multiple intervention-worthy layers — RepE steering applied at a gentle CAZ may produce different behavioral effects than steering at the dominant peak.

5.2 Related Empirical Findings

Manifold interpretability. Gurnee et al. [2025] found curved manifolds in middle layers for character counting and explicitly called for unsupervised geometric discovery methods. The CAZ framework provides a formalism for identifying *where* in the layer stack such manifolds crystallize — the assembly zone is precisely where curved manifold structure should be most geometrically coherent.

The geometry of refusal. Arditi et al. [2024] and Wollschläger et al. [2025] establish

the geometric structure of refusal — a single removable direction (abliteration) and multi-dimensional concept cones, respectively. The CAZ framework extends this by asking not just *what* the geometry is but *when* it forms, and using that temporal structure to identify optimal intervention points.

Multi-dimensional concept structure. Engels et al. [2025] found circular multi-dimensional representations for temporal concepts that are not decomposable into independent one-dimensional SAE features. Wollschläger et al. [2025] showed refusal occupies polyhedral concept cones. These findings establish that rich geometric structure exists; the CAZ framework provides a layer-indexed account of when such structures crystallize.

The Platonic Representation Hypothesis. Huh et al. [2024] proposed that models trained on different data and architectures converge on shared representations. The CAZ framework enables a depth-stratified test of this hypothesis: rather than measuring global alignment, we can ask whether convergence differs at shallow versus deep processing stages.

---

6. Proof of Concept

To demonstrate that the CAZ metrics and detection methods produce meaningful results, we present a minimal example on GPT-2-XL (48 layers, 1.5B parameters) using 7 concepts with 100 contrastive pairs each.

6.1 CAZ Detection

The separation curve S(l) for credibility in GPT-2-XL peaks at layer 46 (96% depth) with S = 0.736 — the strongest signal among the 7 concepts tested. The scored detector identifies 3 CAZes for this concept: a dominant peak at L46, and two gentle CAZes at earlier layers that are invisible under a 10% prominence threshold but produce >20% separation suppression when ablated.

Across all 7 concepts in this single model, the framework detects a consistent ordering by assembly depth: relational and syntactic concepts assemble in the 75–81% range; affective and epistemic concepts cluster in the 92–96% range.

6.2 Scored Detection Reveals Hidden Structure

Lowering the detection threshold from 10% to 0.5% (scored detection) increases the number of detected CAZes from 7 to 23 in this single model. The additional 16 gentle CAZes are not noise — ablation confirms causal impact for the majority.

6.3 Scope of Validation

The framework has been validated across 30 models from 7 architectural families (Pythia, GPT-2, OPT, Qwen 2.5, Gemma 2, Llama 3.2, Mistral) spanning 70M to 7B parameters. Full empirical results — including multi-family scale ladders, structural analysis, cross-architecture alignment, and dark matter quantification — are reported

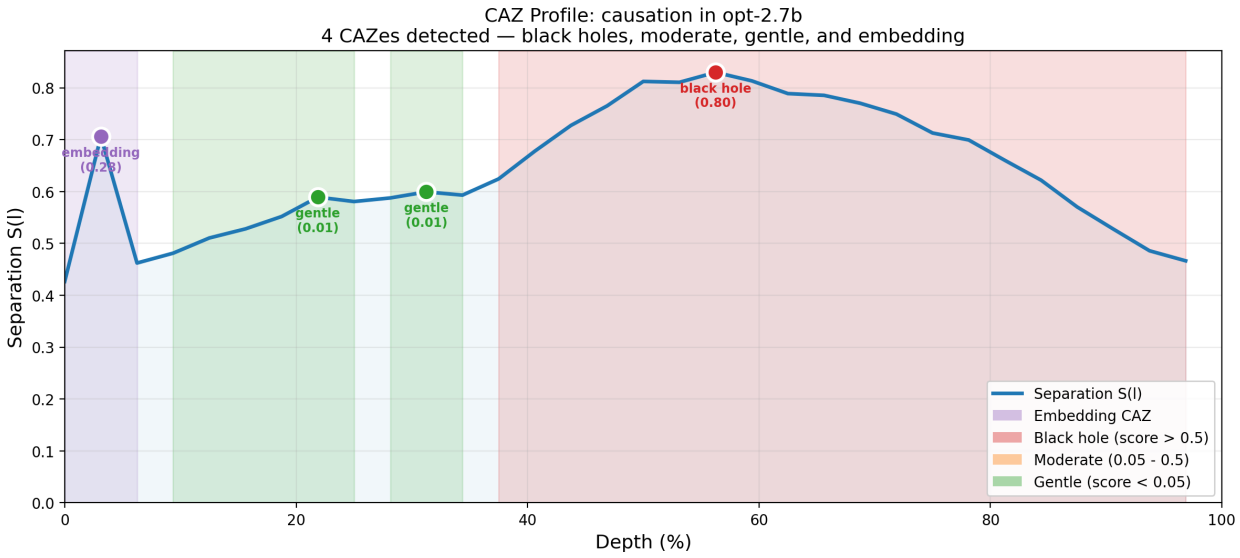

Figure 2: Figure 2: Scored CAZ profile for credibility in Qwen2.5-0.5B. Six CAZes detected: one embedding CAZ (layer 1), one black hole (layer 4), three gentle CAZes (layers 12, 15, 17), and one moderate (layer 19). The full CAZ chain is visible across model depth.

in the companion validation paper [Henry, 2026b]. The reference implementation is provided as rosetta_tools v1.0.0 [Henry, 2026].

---

7. Limitations

Known limitations include:

Synthetic contrastive data bias

This is the most significant limitation of the current work. All 7 concept datasets used for validation were generated by a single model (Claude Sonnet 4.5). Every CAZ location, every separation score, and every cross-architecture alignment number in this paper is measured against Claude's definition of each concept. If Claude's concept boundaries differ systematically from a human consensus definition, then what we are mapping is Claude's ontology projected onto other models, not a universal property of the models themselves. Multi-model consensus pair generation is planned to address this. The framework methodology is independent of the data source; only the specific empirical results are at risk.

Computational cost

The full CAZ extraction pipeline runs in under 2 minutes per model on a single NVIDIA L4 (22 GiB) for models up to 7B parameters. Frontier-scale models (70B+) will require larger GPUs or 8-bit quantization, which may affect metric precision.

Smoothing sensitivity

CAZ boundary detection depends on the smoothing parameter k and threshold θ_+. The current heuristic ($k = \lfloor L/24 \rfloor$) produces consistent results across 30 models but has not been formally validated against ground-truth concept boundaries.

Linearity assumption

The separation metric assumes the concept manifold is approximately linearly separable. For concepts with curved or multi-dimensional structure [Gurnee et al., 2025; Engels et al., 2025], kernel-based or topological metrics may be required. This limitation is shared with most of the current interpretability literature.

Token position dependence

The framework does not account for which token positions carry concept information. Zhao et al. [2025] showed that harmfulness and refusal encode at different token positions. CAZ boundaries may vary by token position as well as by layer.

Causal validation coverage

N-CAZ ablation establishes causal impact for 95% of detected CAZes across all score levels. The remaining 5% of non-causal detections at the gentle level may be noise or measurement artifact. Ablation testing has been performed on 15 of 30 models; full coverage is pending. Details in the companion validation paper [Henry, 2026b].

PRH measurement sensitivity

Cross-architecture alignment numbers are highly sensitive to the rotation estimation method. PCA compression to low-dimensional subspaces inflates alignment scores. The honest range is 24–74% depending on training data quality. Full reconciliation of measurement methods is reported in the companion validation paper [Henry, 2026b].

---

8. Conclusion

The Concept Assembly Zone framework provides a methodology for tracking how concepts form across transformer depth — moving from the "best layer" snapshot to a dynamical view of concept allocation.

The key contributions are:

1. **Three layer-wise metrics** (Separation, Coherence, Velocity) that characterize concept formation as a process, not a point.
2. **Scored detection** that reveals a spectrum of assembly events from dominant to subtle, replacing binary thresholds with continuous scoring.
3. **The CAZ-is-not-a-concept distinction** — CAZes are layer-level allocation events where the model organizes geometry to serve concepts. Multiple concepts share CAZes; single concepts participate in multiple CAZes across depth.
4. **Sub-representation tracking** across depth, revealing that human concept labels map to geometrically distinct directions at different processing stages.

Prediction	Claim	Status
P1	Optimal ablation depth relative to CAZ	Revised — depends on encoding strategy
P2	Architecture-stable CAZ ordering	Confirmed for relative ordering
P3	CAZ width correlates with abstraction	Initial support
P4	Post-CAZ degradation correlates with unembedding	Not yet tested
P5	Cross-architecture alignment is depth-matched	Strongly confirmed
P6	Shallow peaks are lexical, deep are compositional	Not supported by initial test
P7	Multi-modality is architectural, not scale-dependent	Supported with nuance

The framework has survived empirical stress-testing across 30 models while requiring revision of its core assumption from single-peak to multi-peak assembly — a revision that strengthened its explanatory power. Full empirical results are reported in the companion validation paper [Henry, 2026b].

The reference implementation is available as rosetta_tools v1.0.0 [Henry, 2026], an open-source Python library providing the full CAZ extraction, alignment, ablation, and feature tracking pipeline described in this paper.

References

- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*. https://arxiv.org/abs/2406.11717
- Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*. https://arxiv.org/abs/1610.01644
- Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., & Olah, C.

(2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, Anthropic. https://transformer-circuits.pub/2023/monosemantic-features/index.html

- Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*. https://arxiv.org/abs/2309.08600
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, Anthropic. https://transformer-circuits.pub/2021/framework/index.html
- Engels, J., Riggs, L., & Tegmark, M. (2024). Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research (TMLR)*, April 2025. *arXiv preprint arXiv:2410.14670*. https://arxiv.org/abs/2410.14670
- Engels, J., Michaud, E. J., Liao, I., Gurnee, W., & Tegmark, M. (2025). Not all language model features are one-dimensionally linear. *Proceedings of the International Conference on Learning Representations (ICLR 2025)*. *arXiv preprint arXiv:2405.14860*. https://arxiv.org/abs/2405.14860
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1), 49–55.
- Henry, J. (2026). rosetta_tools: Shared tooling for the Rosetta interpretability research program (v1.0.0). https://github.com/jamesrahenry/Rosetta_Tools
- Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The Platonic Representation Hypothesis. *Proceedings of the International Conference on Machine Learning (ICML 2024)*. *arXiv preprint arXiv:2405.07987*. https://arxiv.org/abs/2405.07987
- Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. *Proceedings of the International Conference on Machine Learning (ICML 2019)*, 3519–3529.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
- Gurnee, W., Ameisen, E., Kauvar, I., Tarng, W., Pearce, A., Olah, C., & Batson, J. (2025). When models manipulate manifolds. *Transformer Circuits Thread*, Anthropic, October 2025. *arXiv preprint arXiv:2601.04480*. https://arxiv.org/abs/2601.04480
- Wollschläger, T., Elstner, J., Geisler, S., Cohen-Addad, V., Günnemann, S., & Gasteiger, J. (2025). The geometry of refusal in large language models: Concept cones and representational independence. *Proceedings of Machine Learning Research (ICML 2025)*, 267, 66945–66970. *arXiv preprint arXiv:2502.17420*. https://arxiv.org/abs/2502.17420

- Zhao, J., Huang, J., Wu, Z., Bau, D., & Shi, W. (2025). Harmfulness and refusal are distinct concepts in language models. *Advances in Neural Information Processing Systems (NeurIPS 2025)*. *arXiv preprint arXiv:2507.11878*. https://arxiv.org/abs/2507.11878
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., & Hendrycks, D. (2023). Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*. https://arxiv.org/abs/2310.01405