Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Features have life histories. And we should care

Philipp Moritz Stecher

philippmstecher@gmail.com

University of Tübingen https://orcid.org/0009-0006-4149-1741

Sandro Radovanović

University of Belgrade https://orcid.org/0000-0002-9975-8844

Vlasta Sikimić

Eindhoven University of Technology https://orcid.org/0000-0002-6996-9341

Reinhard Kahle

University of Tübingen https://orcid.org/0000-0002-9064-877X

Research Article

Keywords: Machine learning, representation learning, representational development, sparse auto encoders

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9712563/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Features have life history. And we should care.

Philipp Stecher
University of Tübingen
philipp.stecher@uni-tuebingen.de

Sandro Radovanović
University of Belgrade

Vlasta Sikmić
TU Eindhoven

Reinhard Kahle
University of Tübingen

Abstract

Features in language models have life history: they emerge, persist, and die during training, yet the importance of that history remains largely unexplored. We find evidence of a persistent representational backbone, which we identify in Pythia-160M and -410M as the carrier scaffold: a small population of sparse features with stable life histories, around which the model's representational structure organises. It has four properties. *(i) It assembles early:* features emerge, die, and reorganise $\sim 40\times$ faster in the first 1% of training than afterwards, and the scaffold is already largely fixed by then. *(ii) It is load-bearing:* joint cross-layer ablation identifies the carriers as far more load-bearing than any count-matched non-scaffold population. *(iii) Function precedes direction:* which features will become carriers is already predictable from training-onset firing patterns alone, correctly distinguishing future carriers from non-carriers in 4 of 5 cases, before the geometry has settled. *(iv) It seeds subsequent development:* by the end of training, scaffold carriers have recruited 64% of all active features into the scaffold hierarchy. Life history is consistent with a two-phase account of training: selection appears to largely determine the scaffold in the first 1%; the remaining 99% appears to calibrate geometry around a substrate already set.

1 Introduction

Training is not uniform. The early phase of learning is qualitatively different from what follows: features are born and die at high rates, representations shift rapidly, and the model's internal structure reorganises faster than at any later point. By the time training ends, this turbulence has resolved into something stable, but the structure that stability settled around is not arbitrary. Some features form early, hold their position throughout, and end up carrying a disproportionate share of the model's representational load. Others arrive later, occupy similar positions at convergence, and carry far less. A snapshot cannot tell them apart, yet the difference matters: feature-editing methods that target features by their geometry at convergence may be operating on features whose load derives from a history that geometry alone does not reveal. Understanding how that difference arises is therefore an interpretability question with direct consequences for model debugging, editing, and safety.

We propose feature life history as the unit of analysis and operationalise it with sparse autoencoders (SAEs) [1, 2], which decompose the model's internal activations into a sparse set of interpretable features. A feature's life history tracks two properties across training: what it responds to (its firing identity) and how it influences predictions (its decoder direction: the direction it writes into the residual stream, shaping downstream computation). These calibrate on very different timescales: firing identity stabilises early, anchored to training data that is fixed from the start; decoder direction continues to be worked out across the bulk of training. Identifying each feature by firing identity

rather than by its position in activation space gives us both a way to track features across checkpoints and a natural hierarchy in which broader features contain the firing patterns of narrower, more specialised ones. Tracking both properties across checkpoints lets us identify a small population of features whose firing identity has been stable since the opening phase of training and which have organised the majority of the model's representational structure around them by convergence. We call this population the carrier scaffold.

We train SAEs at seven points during training on two scales (Pythia-160M and Pythia-410M [3]) and follow each feature's birth, persistence, and position in this hierarchy. Recent work tracks SAE features across checkpoints [4–7]; Aoyama et al. [8] predict induction-head emergence training-time, closest in spirit to our scaffold predictor. Prior approaches embed identity architecturally, via shared crosscoder encoders [5, 6] or warm-start weight inheritance [4], or forgo individual tracking altogether [7]; we instead train independent SAEs at each checkpoint and match post-hoc by activation-profile correlation. Where prior work reports what features form and when, we additionally test whether the early-formed population is load-bearing, show that its membership is predictable from firing patterns before representations converge, and trace how it seeds the mature network's representational structure.

Contributions. The central methodological contribution is *feature life history* as a unit of analysis. A feature broad at convergence that has been broad since step 1k is a different scientific object from one that is broad at convergence due to noise in the SAE training: the former is part of the load-bearing scaffold; the latter is not. Tracking each feature's firing identity and decoder direction across checkpoints reveals a temporal split, invisible to any single-snapshot analysis, that makes the carrier scaffold recoverable.

1. **The scaffold is born early (§3.1)**: the per-step feature birth rate collapses $\sim 40\times$ in the first $\sim 1\%$ of training, and most carriers are already identifiable at that point, before the model can reliably predict; a small persistent population forms one connected cross-layer structure, replicating at both scales.
2. **The scaffold is load-bearing (§3.2)**: jointly ablating persistent carriers incurs far greater cross-entropy cost than removing any count-matched non-scaffold population at the same structural tier, a gap invisible to per-firing single-feature methods and emergent only under the joint cross-layer intervention; replicated at both scales.
3. **Function precedes direction (§3.3)**: a logistic predictor trained on step-1k firing patterns correctly identifies future scaffold members in 4 of 5 pairwise comparisons before decoder directions have stabilised, with $9.3\times$ enrichment at the top operating point.
4. **The scaffold seeds subsequent feature development (§3.4)**: scaffold carriers recruit subsequently born features via hierarchical connections, organising 64% of active features at Pythia-160M (53% at Pythia-410M) into the scaffold hierarchy by convergence, invisible to trained-snapshot analysis and recoverable only by tracking the scaffold's growth across training.

Together, these results suggest that the representational structure visible at convergence is largely the product of a selection process that completes in the first 1% of training, a timescale that trained-snapshot interpretability cannot observe.

2 Method

We track SAE features across training checkpoints of two model scales, identify a persistent carrier population, and test whether it is load-bearing. We instrument two Pythia scales across seven checkpoints, train one SAE per (layer, checkpoint) pair, match features across time to build life histories and a cross-layer DAG, and measure load via joint cross-layer ablation and a decoder transplantation control. A training-time predictor quantifies how early the scaffold is identifiable from function alone. Full methodological details are in App. C.

2.1 Models, data, and checkpoints

We instrument Pythia-160M and Pythia-410M [3] at seven training checkpoints: steps 0, 1k, 10k, 40k, 80k, 120k, and 143k (full schedule in Table 5, App. C). Steps 0–10k cover $\sim 7\%$ of training

by wall-clock progress but concentrate the majority of life events; step 143k is the convergence checkpoint. We instrument four residual-stream layers per scale at approximately uniform fractional depth: layers 2, 5, 8, 11 at Pythia-160M ($\sim$18%, 45%, 73%, 100%) and layers 4, 11, 17, 23 at Pythia-410M ($\sim$17%, 46%, 71%, 96%). All SAE training, feature matching, and ablation experiments use a held-out slice of The Pile [9] not seen during Pythia pretraining; feature-level statistics are computed on a canonical validation sample S^* of 20,000 token positions drawn from this slice (seed $= 0$; App. C).

2.2 Sparse autoencoders

At every (checkpoint, layer) pair we train a top-k SAE [2, 1, 10] on residual-stream activations:

$$\hat{x} = W_{\text{dec}}\,\text{TopK}(W_{\text{enc}}\,x + b_{\text{enc}}) + b_{\text{pre}}, \qquad \mathcal{L} = \|x - \hat{x}\|_2^2,$$

with k=32 simultaneously active features, expansion factor 64, decoder columns unit-normalised at each step, and Adam (η=3×10^{-4}, batch size 64) for 4,000 steps on the Pile validation slice. A feature is alive at a checkpoint if its activation exceeds 10^{-6} on $\geq$0.5% of S^* samples. SAE reconstruction quality (variance explained, cosine similarity to the original residual stream) is reported in App. C. A robustness control with a higher-budget SAE (more samples, more steps) is in App. C.

2.3 Feature tracking: matching, life events, and survival

At each consecutive checkpoint pair (t, t'), features are matched by Hungarian assignment [11] that maximises total similarity:

$$M^* = \arg\max_{M\in\mathcal{M}} \sum_{(u,v)\in M} r(u,v), \qquad r(u,v) = \frac{\sum_{s\in S^*}(a_u^t(s) - \bar{a}_u^t)(a_v^{t'}(s) - \bar{a}_v^{t'})}{\|\mathbf{a}_u^t - \bar{a}_u^t\|\,\|\mathbf{a}_v^{t'} - \bar{a}_v^{t'}\|},$$

where $a_u^t(s)$ is feature u's activation on sample s at checkpoint t and $\mathcal{M}$ is the set of one-to-one matchings; a match is accepted only when $r \geq 0.5$. Each transition is classified as *stable* (matched), *born* (unmatched, newly alive), or *died* (unmatched, no longer alive). A feature lineage is the maximal chain of stable matches from birth. Event types (Assembly, Decay, Task-general, Abstraction, Differentiation) follow the RDT taxonomy of Stecher et al. [12], with CI replacing the label-derived selectivity indices of the original (App. C).

Survival to step 143k is computed as the fraction of features within a firing-entropy quartile at step 1k whose lineage persists to the final checkpoint; quartiles are defined over all alive features at step 1k within each scale. Directional precedence in Phase I event-rate sequences (births, abstractions, differentiations at each layer) is assessed via lag-1 Pearson correlations on log-transformed event counts (App. C). Robustness controls (independent-initialisation replication; higher-budget SAE) are in App. C.

2.4 Persistent carriers and the cross-layer DAG

A lineage is a *persistent carrier* if it is stable in ≥ 4 of the 6 post-initialisation snapshots (steps 1k–143k). The Containment Index of feature u with respect to feature v is

$$\text{CI}(u,v) \;=\; \frac{|T_u \cap T_v|}{|T_u|},$$

where T_u is the set of S^* tokens on which u activates above threshold. CI identifies feature hierarchies from nested firing patterns alone, without semantic labels. The cross-layer DAG has persistent carriers as nodes; an edge from u at layer ℓ to v at layer $\ell' > \ell$ is drawn when the subset fraction $|T_u \cap T_v|/|T_v| \geq \tau$ (calibrated per snapshot to control null-edge density), the co-firing lift exceeds 2.0, and the edge survives $\geq 80\%$ of bootstrap resamples; full construction details in App. C.

Sub-DAG coverage at step t is the fraction of all active features at t that are reachable from at least one persistent carrier through any sequence of edges in the cross-layer DAG (BFS transitive closure from the carrier seed set). *Firing breadth* of a feature is the fraction of S^* tokens on which its activation exceeds the alive threshold.

2.5 Ablation load and functional decomposition

Per-firing single-feature impact [13, 14] (the ΔCE when one feature's reconstruction is zeroed) is the field-default importance readout. We instead read load at the population level: for target population P, we simultaneously zero every feature in P across all four instrumented layers, then measure aggregate ΔCE (nats above SAE roundtrip baseline) on a held-out validation pass (95% bootstrap CI, n=1,000 resamples of 32 evaluation chunks). Three populations form the comparison: (1) persistent carriers (51 at 160M, 49 at 410M); (2) non-persistent roots (NCR): features at the same layer and CI tier but stable across fewer than four snapshots, count-matched to carriers; (3) matched leaves: CI-tier-0 features breadth-matched by layer and firing-rate. NCR isolates the contribution of temporal persistence from structural position; leaves establish the near-zero baseline for non-hierarchical features.

To isolate decoder direction's contribution to load, we replace each persistent carrier's step-143k decoder column with its step-1k value while leaving all other model weights unchanged, then measure ΔCE on the same validation pass. A complementary condition replaces only the encoder row. The difference between full ablation and transplantation costs quantifies how much load derives from decoder-direction calibration across Phases II–III versus from firing-pattern identity fixed in Phase I.

Feature function is its firing-pattern identity: the per-token activation distribution across S^* [15, 16, 12]. Feature direction is its decoder column. Step-to-step convergence is measured by Pearson r of activation profiles (function) and cosine similarity of decoder columns (direction) to their step-143k values; full curves are in App. D. For §3.3, a logistic regression with five-fold cross-validation is fit on six step-1k firing-space properties to predict step-143k persistent-carrier identity. Leave-one-layer-out cross-validation additionally assesses generalisation across layers. Full predictor covariates, coefficient estimates, and operating-point statistics are in App. F.

3 Results

3.1 The scaffold is born early

The model commits to its scaffold long before it can predict competently. By step 1k (validation cross-entropy still above 5 nats at Pythia-160M; final: $\sim$2.9 nats), 73% of eventual carriers are already present at Pythia-160M and 59% at Pythia-410M (Fig. 1c,d). The per-step birth rate falls $\sim$40$\times$ across Phase I Genesis (steps 0–1k) and a further $\sim$10$\times$ over the remainder (Fig. 1a), compressing representational work into the opening 1%.

Early survival is not random: features in the top firing-entropy quartile at step 1k survive to convergence at 43% vs. 0.6% for bottom-quartile features at Pythia-160M (72$\times$ advantage; Pythia-410M: 38% vs. 0.4%; App. C). About 35% (160M) and 31% (410M) of eventual carriers are already weakly active at step 0; Phase I partly amplifies latent structure from the random initialisation rather than building the scaffold from scratch (App. C).

The Containment Index (CI) distribution at convergence is heavy-tailed (Fig. 1b): a feature with CI $= k$ fires on every token context that activates any of k narrower features, so high-CI features are broad representational hubs while the majority encode narrow, specific patterns. The features in the right tail of this distribution are the persistent carriers: SAE features whose chain-matched lineage is stable in at least 4 of 6 post-initialisation snapshots (App. C). 51 at Pythia-160M and 49 at Pythia-410M, they form one connected cross-layer object spanning all layers at Pythia-410M and three of four at Pythia-160M (L11 absent; App. E). Throughout, "scaffold" refers to this population and the cross-layer DAG (nodes: persistent carriers; edges: token-set containment) is the graph among them. Semantic content, containment structure, and tier anatomy are detailed in App. A–B; carriers span a content gradient with depth, from byte-boundary and morphological features at early layers to discourse-level abstractions at later ones.

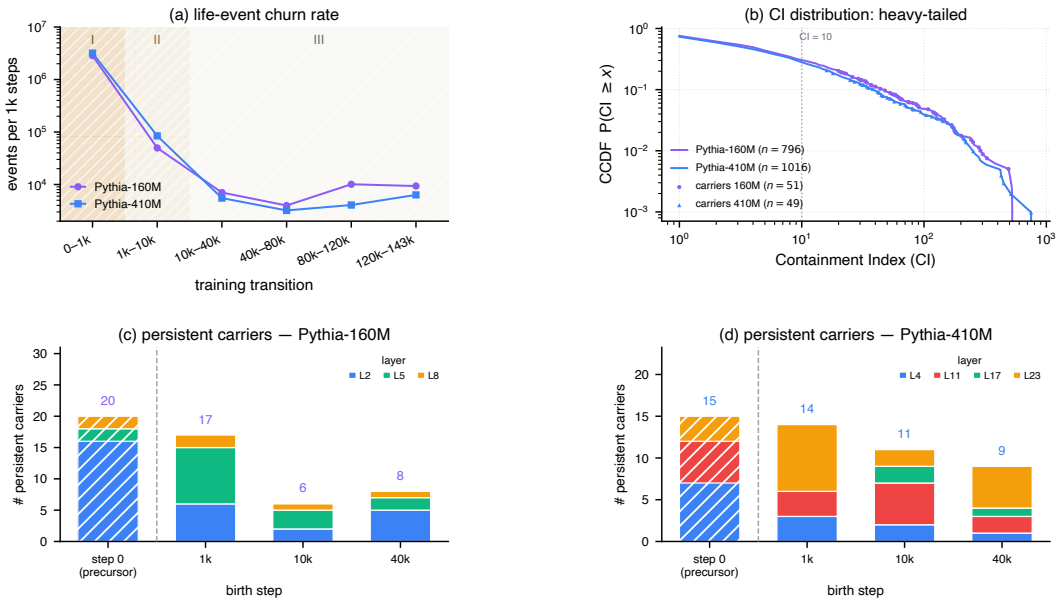


Figure 1: The scaffold's assembly dynamics and its organisational reach. **(a)** Per-step feature life-event rate at Pythia-160M and -410M across Phase I Genesis (steps 0–1k, dark shading), Phase II Assembly (1k–10k, lighter), and Phase III Refinement (10k–143k). Event types: As-E Assembly (birth), De-E Decay (death), Tg-E Task-general (broadening to root tier), Ab-E Abstraction (broadening to mid tier), Di-E Differentiation (narrowing to leaf tier); full definitions in App. C. **(b)** CCDF of the Containment Index (CI) on log-log axes at step 143k; approximately linear over CI $\in [5, 80]$ ($\alpha \approx -0.80$). Features with CI $\gg 10$ are the persistent carrier seeds. **(c)** Persistent-carrier birth-step distribution at Pythia-160M (bars stacked by layer: L2 blue, L5 green, L8 amber); hatched bar: pre-training precursors already weakly active at step 0 (App. C). L11 contributes zero carriers (substrate failure; App. E). **(d)** Same at Pythia-410M (L4 blue, L11 red, L17 green, L23 amber). Roughly 40% (160M) and 31% (410M) of carriers are pre-training precursors.

3.2 The scaffold carries load

The cross-layer DAG was built from token-set containment alone, with no reference to ablation impact; the load test below is therefore out-of-sample. The scaffold is load-bearing. Jointly ablating all 51 persistent carriers costs $+0.90$ nats above the SAE roundtrip baseline at Pythia-160M (95% bootstrap CI $[+0.51, +1.26]$); 52 breadth-matched leaves cost -0.01 nats, a $\sim 91\times$ gap (carrier ablation pushes cross-entropy roughly 30% above the model's converged ~ 2.9 nat baseline) invisible to standard per-firing single-feature diagnostics (App. D) and visible only under the joint cross-layer intervention introduced here. The pattern replicates at Pythia-410M: carriers $+0.52$ nats (CI $[+0.48, +0.56]$), leaves $+0.24$ nats (Fig. 2a).

Persistence, not just structural position, drives the gap: count-matched non-persistent roots (NCR; same CI tier, chain-matched across fewer than four snapshots) cost $+0.57$ nats (160M) and $+0.14$ nats (410M), substantially less than the carriers despite occupying the same tier at the final checkpoint: two features at identical structural positions can differ $\sim 1.6\times$ in ablation cost based solely on whether that position was held stably across training (App. B).

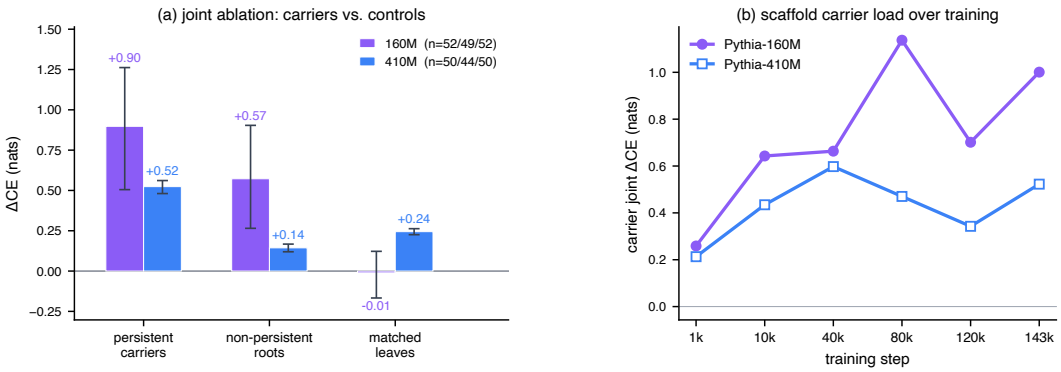


Figure 2: Joint cross-layer ablation quantifies the scaffold's load. **(a)** Joint ΔCE above SAE roundtrip baseline for three count-matched groups, both scales shown side by side: persistent carriers (purple/blue), non-persistent roots (lavender/light blue), matched leaves (grey). Error bars show 95% bootstrap CIs (n=1,000 resamples). Per-firing ΔCE is tier-invariant (App. D). **(b)** Carrier joint ΔCE over training for both scales. A transient dip at step 120k at Pythia-160M coincides with the L11 substrate reorganisation (App. E).

Decoder calibration accounts for the dominant share of the scaffold's load. A transplantation control isolates each contribution: replacing each carrier's decoder column with its step-1k value costs $+0.96$ nats at Pythia-160M and $+0.61$ nats at Pythia-410M, nearly matching full ablation, while replacing only encoder rows costs $+0.58$ nats (160M) and $+0.44$ nats (410M). This is possible because the two properties of a carrier are separable: *function* (which tokens activate it) can be fixed while *direction* (where in residual space it points) is still near-random. Function converges fast; decoder direction starts near-random and calibrates across Phases II–III, and carrier load rises monotonically as direction calibrates around fixed functions (Fig. 2b; App. A). This temporal split opens a prediction window: scaffold membership may be detectable from function alone, before direction converges (§3.3).

3.3 Function precedes direction

Firing identity is largely settled at step 1k while decoder direction has barely begun to calibrate. Fig. 3a shows that per-carrier correlation with the final firing pattern is already high at step 1k and barely changes thereafter; Fig. 3b shows the opposite for decoder direction, which starts near-random and rises steadily across Phases II–III. A feature's activation pattern across token contexts is therefore already a reliable signal of its eventual scaffold membership, while the direction it writes into the residual stream is still being worked out over the bulk of training. The gap between these two timescales creates an audit window: a point early in training where scaffold membership is legible from firing statistics alone, well before the network geometry that encodes it has formed, and the panels (c) and (d) of Fig. 3 show what a predictor trained on that window can recover.

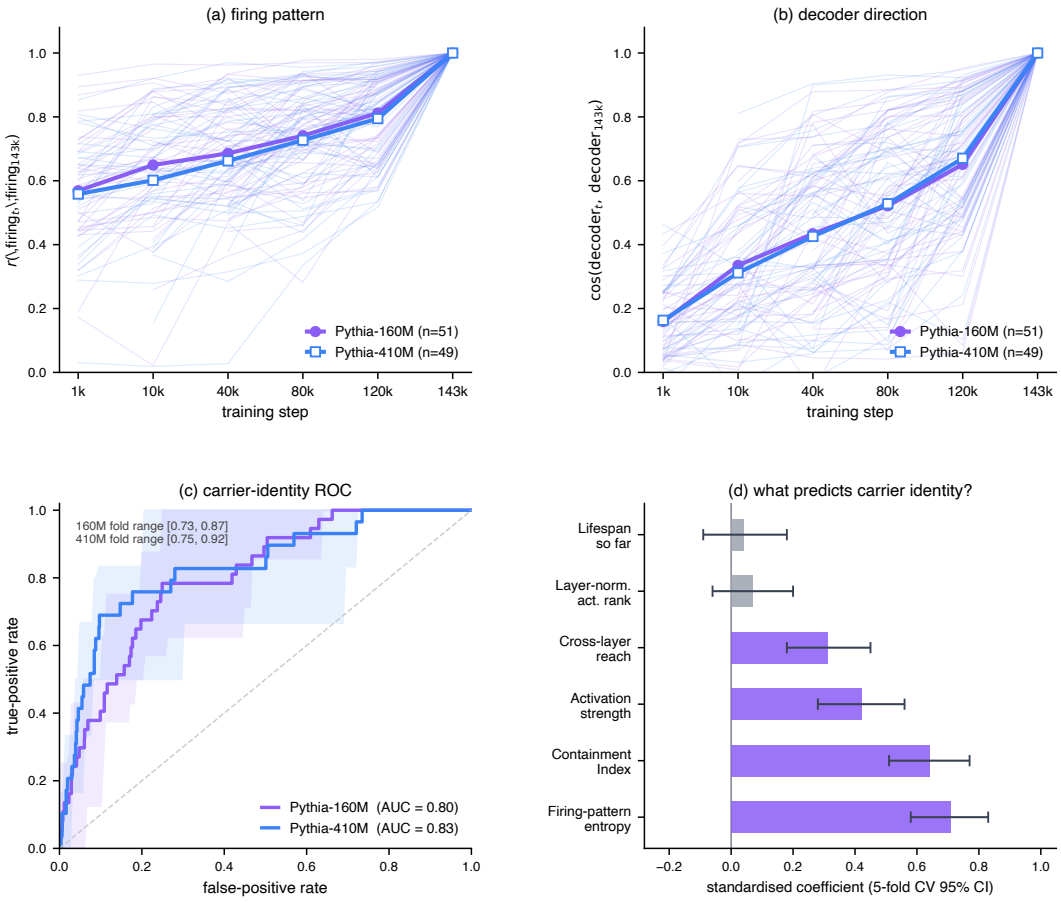


Figure 3: Function settles early; direction calibrates late; the gap makes the scaffold legible at step 1k. **(a)** Per-carrier Pearson r to step-143k firing pattern ($n = 51/49$; faint lines: individual carriers; bold: per-scale means). Identity is largely fixed within Phase I. **(b)** Per-carrier cosine to step-143k decoder column. Direction starts near-random at step 1k and rises across Phases II–III. **(c)** Five-fold CV ROC predicting step-143k carrier identity from step-1k features: Pythia-160M (purple, AUC $= 0.80$, fold range $[0.73, 0.87]$) and Pythia-410M (blue, AUC $= 0.83$, $[0.75, 0.92]$); shaded bands span per-fold min/max TPR. **(d)** Standardised logistic-regression coefficients (5-fold CV 95% CI). Firing breadth and CI at step 1k dominate; max activation, conditional mean, activation variance, and root flag are secondary.

Six step-1k firing-space features predict step-143k carrier identity at five-fold CV AUC $= 0.80$ at Pythia-160M and 0.83 at Pythia-410M (Fig. 3c,d; App. F): at the top-decile operating point the predictor correctly distinguishes future carriers from non-carriers in roughly 4 of 5 decisions, a $9.3\times$ enrichment over the base rate. Firing-pattern entropy is the dominant predictor (standardised coefficient $+0.71$; Fig. 3d); an entropy-only predictor recovers the majority of this signal, with Containment Index, activation strength, and cross-layer reach contributing incremental lift. Broadly-firing features occupy more token contexts, accumulate gradient signal across more positions, and are less likely to go dark when the training distribution shifts — which is why entropy at step 1k already predicts long-run persistence. The classifier holds across layers (leave-one-layer-out AUC $\in [0.74, 0.79]$; App. F), confirming that scaffold membership is grounded in function rather than layer-specific geometry. The features identifiable from this early window are not just predictable; they are the organisational seeds around which the rest of the network differentiates, as the next section shows.

3.4 The scaffold seeds subsequent development

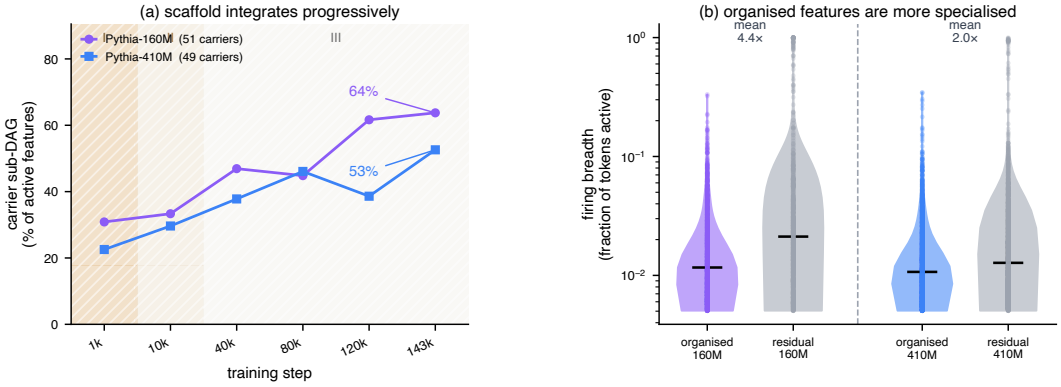


Figure 4: The scaffold organises two-thirds of the mature network through hierarchically scaffolded differentiation. **(a)** Carrier sub-DAG as percentage of active features at Pythia-160M (purple) and Pythia-410M (blue) across six training checkpoints on a log scale; phase shading as in Fig. 1c. **(b)** Firing breadth (fraction of tokens activating each feature) at step 143k for scaffold-organised features (inside carrier sub-DAG; purple/blue) versus unorganised residual features (grey), at both scales. Violins show full distributions; black bars mark medians. Mean breadth is $4.4\times$ lower inside than outside at Pythia-160M ($2.0\times$ at Pythia-410M). Grey dashed line separates the two scales.

Life-history tracking converts a flat feature list into a structured partition. At step 143k, Pythia-160M has $2{,}408$ active features with no record of origin; the framework reveals 51 seeds, $1{,}484$ scaffold-organised descendants, and 873 unorganised residuals. Without tracking, seeds and descendants are indistinguishable in the convergence snapshot, and the $\sim 91\times$ joint-ablation gap disappears into per-feature noise.

The partition grows through hierarchically scaffolded differentiation: any newly born feature whose firing tokens are a subset of an existing carrier's is assimilated as a descendant, deepening the hierarchy without displacing its seeds. At step 1k, the carriers already present account for 31% of active features at Pythia-160M via immediate containment descendants (23% at Pythia-410M). By convergence, the carrier sub-DAG encompasses 64% of active features at Pythia-160M and 53% at Pythia-410M, a ~ 33 percentage-point expansion consistent across scales (Fig. 4a; App. B). The majority of the model's representational capacity at convergence is therefore organised under a hierarchy whose roots were identifiable in the first 1% of training.

The scaffold's organisational reach leaves a geometric signature: features inside the sub-DAG fire with $4.4\times$ lower mean breadth at Pythia-160M ($2.0\times$ at Pythia-410M; Fig. 4b): scaffold recruitment organises progressively more specific sub-patterns around the seed representations.

Table 1: Summary of key metrics at both scales. ΔCE values are joint cross-layer ablation costs in nats above SAE roundtrip baseline (95% bootstrap CI, $n{=}1{,}000$). Sub-DAG coverage is the fraction of active features inside the carrier sub-DAG at step 143k.

Model	Carriers	Born $\leq$ step 1k	ΔCE carriers	ΔCE NCR	AUC (step 1k)	Sub-DAG
Pythia-160M	51	73% (37/51)	$+0.90$ [$+0.51$,$+1.26$]	$+0.57$ [$+0.27$,$+0.90$]	0.80	64%
Pythia-410M	49	59% (29/49)	$+0.52$ [$+0.48$,$+0.56$]	$+0.14$ [$+0.12$,$+0.17$]	0.83	53%

4 Discussion

The results suggest that features have life history. Whether we should care depends on what changes when that history is known; three things do.

A selection mechanism becomes legible. Training is noisy in its early phase: features are born and die at high rates and the per-step event rate is an order of magnitude above its eventual level. Yet scaffold membership is already detectable from firing statistics at step 1k, before the model predicts competently. The asymmetry is consistent with a two-phase regime. In the early window, data and task pressure select which representational niches survive, favouring broadly-firing features that occupy more token contexts and accumulate gradient signal across more positions. The remaining training calibrates how each niche is expressed: firing identity, fixed by selection, determines membership; decoder direction, calibrated across Phases II–III, determines geometry. The load asymmetry follows from the first: scaffold membership, not decoder direction, carries the load, which is why the asymmetry is invisible to per-firing diagnostics and emerges only under joint intervention.

Scaffolding may be a general phenomenon of representation learning. The broadly-firing $\rightarrow$ categorical $\rightarrow$ class-specific developmental ordering has been reported across CNN, ViT, and CCT architectures in vision [12], and the present work finds an analogous structure in language: a small population of broad-firing features that organises the majority of the mature representational hierarchy. The common thread is the primacy of data and task: which features survive early training appears to be determined by gradient signal density across the training distribution rather than by architecture. This suggests scaffold formation may be a consequence of gradient-based learning on structured data broadly. The present work adds a load test that prior accounts do not perform: the scaffold is not merely early-formed but disproportionately load-bearing, with that load concentrated in a population identifiable before geometry has settled. In vision, this structure is not only traceable but exploitable: a curriculum that pre-builds the scaffold reduces subsequent differentiation cost 6.7-fold while preserving validation accuracy [12], and the construction/refinement ratio serves as a training diagnostic invisible in conventional loss curves. The predictability of scaffold membership at step 1k in language raises the prospect of analogous scaffold-aware training strategies.

Early scaffold analysis may open new diagnostics. If scaffold formation is largely complete in the first 1% of training, it becomes a tractable proxy for properties that otherwise require full training runs. Neural architecture search is a direct application: scaffold formation speed, size, and load concentration at step 1k could serve as early proxies for architecture quality, reducing the compute cost of architecture comparison substantially. Scaffold membership could also guide continual learning by identifying features with demonstrated load-bearing history, providing a more principled basis for freezing decisions than rank or magnitude. Comparing scaffolds across models trained on different corpora could further quantify what different data distributions select for, grounding the empirical observation that data shapes representations in a measurable, mechanistic substrate.

Mechanistic interpretability has developed powerful tools for the trained snapshot. The life-history view complements them: the scaffold is established before convergence, traceable through firing-space, and invisible to any single-checkpoint analysis. The scaffold appears largely fixed by selection in the first 1% of training; the remaining 99% calibrates geometry around a substrate already set. Interpretability that begins at convergence may be reading the product of that process rather than its causes.

5 Limitations

We study Pythia at two scales on one training corpus with four layers per scale; whether scaffold size, load magnitude, and three-phase timing generalise to other architectures, training corpora, or model scales remains open. The evidence is correlational: ablation and prediction suggest that the scaffold is load-bearing and early-identifiable, but do not establish a causal mechanism for its formation. SAE coverage spans four of twelve (160M) and four of twenty-four (410M) residual-stream layers; denser layer sampling could reveal additional scaffold structure or alter the cross-layer connectivity reported here.

References

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

[2] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

[3] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2304.01373.

[4] Yang Xu, Yi Wang, Hengguan Huang, and Hao Wang. Tracking the feature dynamics in LLM training: A mechanistic study. *arXiv preprint arXiv:2412.17626*, 2024. URL https://arxiv.org/abs/2412.17626.

[5] Xuyang Ge, Wentao Shu, Jiaxing Wu, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Evolution of concepts in language model pre-training. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.17196.

[6] Deniz Bayazit, Aaron Mueller, and Antoine Bosselut. Crosscoding through time: Tracking emergence and consolidation of linguistic representations throughout llm pretraining. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://arxiv.org/abs/2509.05291.

[7] Tatsuro Inaba, Go Kamoda, Kentaro Inui, Masaru Isonuma, Yusuke Miyao, Yohei Oseki, Benjamin Heinzerling, and Yu Takagi. How a bilingual LM becomes bilingual: Tracing internal representations with sparse autoencoders. In *Findings of EMNLP*, 2025. URL https://arxiv.org/abs/2503.06394.

[8] Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. Predicting the formation of induction heads. *arXiv preprint arXiv:2511.16893*, 2025. URL https://arxiv.org/abs/2511.16893.

[9] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshnik, Shawn Presser, and Connor Leahy. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021.

[10] Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders. AI Alignment Forum, 2022.

[11] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

[12] Philipp Stecher, Sandro Radovanović, Vlasta Sikimić, and Reinhard Kahle. Scaffolded representation learning in deep networks. *Research Square preprint*, 2026. doi: 10.21203/rs.3.rs-9269961/v1. Research Square preprint, under review.

[13] Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

[14] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2403.19647.

[15] Nicholas Shea. *Representation in Cognitive Science*. Oxford University Press, 2018.

[16] Ben Baker, Benjamin Lansdell, and Konrad P. Kording. Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*, 26(11):942–958, 2022.

[17] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. OpenAI, 2023. URL `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`.

[18] Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

[19] David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024. URL `https://arxiv.org/abs/2409.14507`. Accepted at NeurIPS 2025 (Oral).

[20] Philipp Stecher. From birth to loss of representations in artificial neural networks. In *Software Engineering and Formal Methods. SEFM 2024 Collocated Workshops*, Lecture Notes in Computer Science, pages 271–289. Springer, Cham, 2026. ISBN 978-3-031-94748-3. doi: 10.1007/978-3-031-94748-3_20.

A Semantic Anatomy of the Carrier Scaffold

Byte-boundary infrastructure features, not semantically-interpretable discourse markers, dominate the load-bearing subset of the carrier scaffold: the $2\times$ load concentration of §3.2 localises to a specific subpopulation and rules out simple tokeniser-infrastructure confounds. Of the 51 persistent carriers at Pythia-160M, 39 are labelable from top-8 activating tokens; 82% of those receive high-or-medium-confidence auto-interp labels, validated by blind LLM annotation (90% concept agreement; blind-validation protocol below).

Auto-interp labels and confidence. For each of the 51 carriers we extracted its top-8 activating tokens and logit-boost target class, and submitted both to an LLM for a short English label with high / medium / low / artefact confidence [17, 18]. Table 2 summarises the distribution. 39 of the 51 carriers are labelable; 82% of labelable carriers receive high-or-medium confidence labels. The layer profile tracks depth: L2 is heterogeneous (morphological endings, file-extension orthographics, discourse connectives, and a non-trivial tail of U+FFFD / CJK fragment noise); L5 consolidates into clearer discourse markers and semantic-topic clusters (licensing boilerplate, magnitude, causation); L8's six-carrier population is dominated by high-level discourse abstractions (paragraph-initial framing, modal predicates, temporal-narrative); L11 collapses to zero carriers (substrate failure; App. E).

Table 2: Auto-interpretation confidence for the 51 persistent carriers at Pythia-160M (step 143k, $4\times$ expansion SAE). The artefact class consists of Unicode replacement characters, CJK fragment tokens, and HTML boundary separators concentrated at L2; they satisfy all scaffold criteria (persistent, root-level, monotonically sharpening decoder) but their semantics are byte-level infrastructure rather than discourse content.

Layer	Carriers	High	Medium	Low	Dominant categories
L2	29	10	8	11	Morphological, discourse connectives, CJK/U+FFFD noise
L5	16	6	4	6	Legal boilerplate, causal nouns, spatial particles
L8	6	4	0	2	Sentence-initial framing, modal predicates
L11	0	0	0	0	(substrate failure; App. E)
Total	51	20	12	19	82% high/medium of 39 labelable

Blind LLM validation (90% concept agreement). Twenty persistent carriers were drawn uniformly at random (seed 42) from the full 51-carrier set across L2/L5/L8. Each carrier was presented to a second LLM with only its top-8 activating tokens and corresponding logit-boost values; the auto-interp label, confidence, rationale, and category were hidden. The second LLM provided a short English label and a high/medium/low confidence rating. Blind-LLM vs. auto-interp concept agreement is $18/20$ (90%). Confidence-level agreement (exact high/medium/low match) is $10/20$ (50%). The direction of confidence disagreements is balanced: in 6 of the 10 cases the blind LLM is more confident than auto-interp, in 4 auto-interp is more confident. The six blind-LLM-more-confident cases are dominated by U+FFFD- or CJK-dominated carriers that the blind LLM reads as coherent byte-boundary or multilingual-fragment detectors while auto-interp labels them as low-confidence noise. For example, L2 fid 13486 has all eight top tokens as U+FFFD replacement characters; the blind LLM labels this "Unicode replacement / encoding-error characters" at high confidence while auto-interp labels it "Unicode replacement-character noise" at low confidence. The two concept-level disagreements are carriers with genuinely mixed token sets. Under a stricter rubric that reclassifies mixed-token carriers as artefact-like, the high-or-medium rate of 82% shifts to approximately 64%.

The semantic-confidence / load paradox. The most semantically legible carriers carry near-zero load; the byte-level infrastructure features carry almost all of it. Splitting the 51 carriers by auto-interp confidence yields a 32-carrier high-or-medium-confidence subset (discourse and morphological features) and a 19-carrier low/artefact subset (Unicode replacement characters, CJK fragments, HTML-adjacent separators; concentrated at L2 with 11, plus 6 at L5 and 2 at L8). Joint ablation of the low-confidence subset costs $\Delta\mathrm{CE} = +0.91$ nats (95% CI $[+0.06, +2.55]$); the high/medium-confidence subset costs $+0.001$ nats (95% CI $[-0.31, +0.17]$), indistinguishable from

zero. The proportional shares would be $+0.37$ and $+0.63$ nats; the low-confidence subset carries $2.44\times$ its proportional share. The blind LLM validation agrees with the auto-interp-based split on 9 of 10 items. The paradox proactively answers the "is this just tokenisation?" objection: the scaffold result holds even after excluding the byte-level infrastructure subset; the whole 51-carrier set satisfies every scaffold criterion.

Absorption vs. predictive nesting. Chanin et al. [19] argue that sparsity-driven SAEs suffer feature absorption when underlying features form a taxonomic hierarchy. We compute for every edge $A \to B$ a subset ratio $|\text{top-8}(A) \cap \text{top-8}(B)|/8$; an edge with subset ratio ≥ 0.5 is counted as taxonomic, otherwise predictive. Of the 916 containment edges in the 51-carrier scaffold sub-DAG, 99.6% are predictive (mean subset ratio 0.056), vs. 95.2% for the $3{,}409$ non-scaffold edges (odds ratio ≈ 11, $p \ll 0.001$, Fisher exact). The scaffold is compositional rather than taxonomic: one feature predicts another's firing context rather than strictly subsuming it. A common-token filter (removing the 50 most frequent Pile-val tokens) shifts the predictive fraction by less than 0.3 pp.

Worked hierarchy trace: L8 fid 4795. Table 3 traces the highest-CI high-confidence carrier at step 143k: L8 fid 4795 ("sentence-initial subordinating / framing connectives": *Although, Despite, If, Following, However*; CI $= 243$). The rows are 15 direct descendants plus a 2-hop expansion through nested root L8 fid 17144. The category column labels each level: paragraph-initial discourse framing (root), broad text-structural contexts (nested roots with diverse structural/boundary patterns), transitional mixed-context fragments (mids, no clean axis), and narrow local lexical content (leaves). The content gradient is broad-to-narrow but not strictly taxonomic: only three nested roots are genuine specialisations of the root's discourse-framing axis; the rest are contained for distributional reasons. Cross-layer reach: direct descendants in L2 (4), L5 (4), L8 (7).

Table 3: Two-level trace of L8 fid 4795 at step 143k. Rows sorted root → nested roots → mids → leaves (CI descending within level). Rows marked † are 2-hop descendants through nested root L8 fid 17144. The content gradient broad→narrow is predictive not taxonomic (99.6% non-absorbed edges; see "Absorption" paragraph above).

category	level	feature	CI	top tokens	interpretation
paragraph-initial discourse framing	**root**	(L8, 4795)	243	*If, however, Statistical, Following*	sentence-initial framing markers
broad text-structural contexts	nested root	(L2, 46887)	119	*we, content, 6,* \n	paragraph-internal continuations
	nested root	(L2, 3039)	89	*/, N, younger, fine, videos*	slash-delimited modifier fragments
	nested root	(L8, 17144)	81	*Its, However, grid, a, del*	sentence-initial concessive pronouns
	nested root	(L8, 39082)	34	*other, Code, ,,* \n	clause-connecting adjuncts
	nested root	(L5, 32927)	21	*-, Society, the,* \n	dashed-list nominal references
	nested root	(L2, 48496)	17	*#, us, of, i, be*	hashtag & function-word contexts
	nested root	(L2, 1062)	14	\n*, on, . . . , At, Fair*	paragraph-boundary markers
transitional mixed-context fragments	mid†	(L2, 41638)	9	*THE, -, ignoring, Partners*	sentence-initial noisy capitals
	mid†	(L5, 2609)	2	*GA, maintain, lead, ., really*	action-verb context
	mid	(L5, 37931)	1	*mosp, of, the, land, voted*	sparse non-word + function mix
narrow local lexical content	leaf	(L5, 37889)	0	\n*, be, image, main*	paragraph-start function tokens
	leaf†	(L5, 9789)	0	*)., the, judge, _, give*	closing-punctuation context
	leaf	(L5, 10272)	0	*. . . , crime, brew, played, crowned*	past-tense narrative verbs
	leaf	(L5, 12043)	0	*array, Then, ', Left, she*	sentence-start pronoun + *Then*
	leaf	(L5, 17706)	0	*now, TS, turnover, ress, clone*	temporal adverb + clipped nouns
	leaf	(L8, 10322)	0	*Partner, leader,* \n*, are*	agent / leader nouns
	leaf	(L8, 15916)	0	*thing, now, approximate, develop*	near-future action verbs
	leaf	(L8, 41172)	0	*You, and, of, ?), the*	second-person interrogative context
	leaf†	(L8, 2169)	0	*Account, under, only, ,,)*	function-word closing context

B Tier Anatomy: What Roots, Mids, and Leaves Entail

Roots and mids are nearly immortal; leaves are ephemeral. Yet tier does not predict per-firing ablation impact: load only appears under joint cross-layer intervention (App. D). This dissociation — tier partitions features by representational role but not by per-firing cost — is the quantitative backbone of the load-pivot claim in §3.2.

Lifespan by tier. Birth tier is nearly deterministic for long-run survival. A feature born into a root or mid position will almost certainly persist to step 143k; a feature born into a leaf position almost certainly will not. Kaplan–Meier curves (Fig. 5) confirm that roots and mids are statistically indistinguishable in durability while both massively outlive leaves: at 160M, root vs. leaf $p =$

1.8×10^{-7}, mid vs. leaf $p < 10^{-15}$, root vs. mid $p = 0.92$; at 410M, root vs. leaf $p = 4.3 \times 10^{-10}$, mid vs. leaf $p < 10^{-15}$, root vs. mid $p = 0.47$.

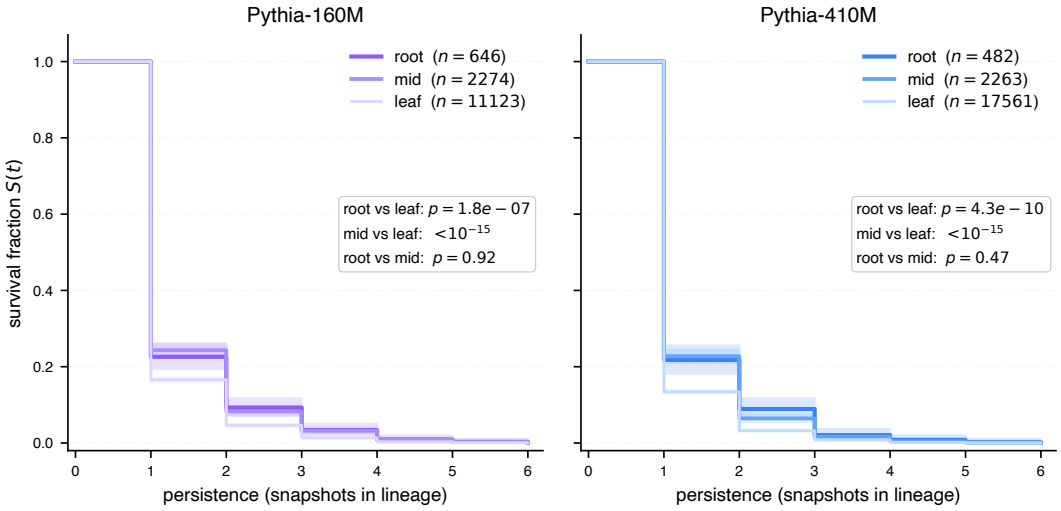


Figure 5: Kaplan–Meier survival by CI tier at Pythia-160M (left) and Pythia-410M (right), with Greenwood 95% confidence bands. Roots and mids are statistically indistinguishable in durability ($p > 0.4$ at both scales); both massively outlive leaves ($p < 10^{-7}$). The tier partition that is invisible to per-firing ablation is a strong predictor of life history.

Firing breadth by tier. Roots fire roughly 5× more broadly than leaves at median, with mids intermediate (Fig. 6; root median 0.048, leaf median 0.009; Pythia-410M, all four layers pooled). This breadth gradient is the axis the step-1k predictor of §3.3 exploits.

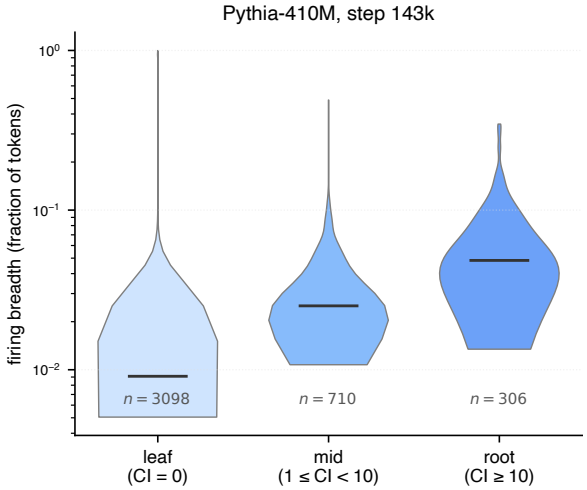


Figure 6: Firing breadth (fraction of validation tokens on which a feature activates) by CI tier at step 143k, Pythia-410M; all four layers pooled ($n_{\text{leaf}} = 3{,}098$, $n_{\text{mid}} = 710$, $n_{\text{root}} = 306$). Roots fire $\sim$5× more broadly than leaves at median (root 0.048, leaf 0.009); mids are intermediate. The same breadth axis is captured by the firing breadth predictor of §3.3.

Per-firing ΔCE is tier-invariant. Despite the lifespan and breadth differences above, per-firing single-feature ablation impact is statistically indistinguishable across tiers (formal KS tests in App. D): load only emerges under joint cross-layer intervention.

CI trajectory per layer across training. Table 4 gives the full per-snapshot trajectory for active features, roots, and scaffold sub-DAG coverage at each Pythia-160M layer. Root counts peak at step 1k (L2: 221, L5: 167, L8: 176, L11: 187) and prune by roughly half to step 10k, while sub-DAG coverage rises from 30.9% at step 1k to 63.7% at step 143k with a minor plateau around step 80k. L11's 80k $\rightarrow$ 120k collapse (active 612 $\rightarrow$ 39, roots 64 $\rightarrow$ 0) is the only qualitative departure; see App. E for the mechanistic account.

Table 4: Per-layer active-feature and root counts across training at Pythia-160M, with cross-layer scaffold sub-DAG trajectory. Sub-DAG coverage grows overall from 31% at step 1k to 64% at step 143k. Root counts peak at step 1k then prune by roughly half. Step 0 is the random-init baseline.

layer	step 0	step 1k	step 10k	step 40k	step 80k	step 120k	step 143k
active features							
L2	1074	821	851	762	820	916	908
L5	858	789	963	923	952	772	790
L8	725	830	954	791	731	568	673
L11	670	758	838	578	612	39	37
roots (CI $\geq$ 10)							
L2	72	221	108	76	97	86	100
L5	47	167	88	79	77	77	81
L8	42	176	95	81	88	82	73
L11	29	187	78	66	64	3	0
cross-layer scaffold sub-DAG							
total active	3327	3198	3606	3054	3115	2295	2408
inside sub-DAG	78	987	1202	1433	1396	1415	1535
coverage (%)	2.3	30.9	33.3	46.9	44.8	61.7	**63.7**

C Methodological Details and Controls

The core findings replicate across independent seeds, scales, and SAE budgets; methodology is robust to threshold choice. This appendix collects full definitions, construction details, and robustness controls: the persistent-carrier criterion, cross-layer DAG construction, the RDT event taxonomy with per-window counts, the lag-1 Granger test on Phase I sequences, broad-born survival curves, pre-existing structure at random initialisation, and seed and expansion-factor replication.

Table 5: Experimental setup. Seven training checkpoints per scale (six post-initialisation + step 0); four residual-stream layers per scale at uniform depth spacing.

Scale	Checkpoints (step)	Layers
Pythia-160M	0, 1k, 10k, 40k, 80k, 120k, 143k	L2, L5, L8, L11
Pythia-410M	0, 1k, 10k, 40k, 80k, 120k, 143k	L4, L11, L17, L23

Persistent-carrier definition (full). Following the criterion in §2: a persistent carrier's lineage must be stable in at least 4 of the 6 post-initialisation snapshots at steps 1k, 10k, 40k, 80k, 120k, 143k (alive threshold 10^{-6} on $\geq$0.5% of validation samples; match threshold $r \geq 0.5$). At Pythia-160M this yields 51 carriers across L2/L5/L8/L11; at Pythia-410M, 49 across L4/L11/L17/L23.

Persistence-threshold sensitivity. Stricter thresholds yield a smaller but more coherent scaffold, not a weaker one. The canonical threshold ($\geq$ 4 of 6 post-initialisation snapshots) is the most conservative point at which the scaffold is definable; tightening it monotonically strengthens the early-commitment signal: at $\geq$ 5 snapshots the 160M carrier count falls from 51 to 43 while the fraction born at step 1k rises from 73% to 86%; at $\geq$ 6 snapshots (present at every checkpoint) the count falls to 37, all born at step 1k (100%). The core finding is therefore not an artefact of threshold choice.

Table 6: Persistence-threshold sensitivity at Pythia-160M. Stricter thresholds select a smaller but more cohesive carrier set with stronger early-commitment signal.

Threshold	n carriers	% born $\leq$ step 1k	% step-0 precursors
≥ 4 snapshots (canonical)	51	73%	39%
≥ 5 snapshots	43	86%	47%
≥ 6 snapshots (maximum)	37	100%	54%

Cross-layer DAG construction. For each ordered layer pair (ℓ, ℓ') with $\ell < \ell'$ and each pair of carriers (u, v) with u at layer ℓ and v at layer ℓ', we compute the subset fraction $s(u, v) = |T_u \cap T_v|/|T_v|$ and the co-firing lift $l(u, v) = s(u, v)/(b_u \cdot b_v)$ on a fixed validation slice, where T_u, T_v are the firing-token sets and b_u, b_v are their respective firing breadths. A directed edge $u \to v$ is drawn when (i) $s(u, v) \geq \tau$, where τ is calibrated per snapshot to keep the expected null-edge count at or below 10; (ii) $l(u, v) > 2.0$; (iii) $b_u > b_v$ (broader feature points to narrower); and (iv) the edge survives in $\geq 80\%$ of 30 bootstrap resamples of S^*. Transitive reduction is applied to yield the final DAG.

RDT mapping with CI. We reproduce the RDT process definitions of Stecher et al. [12], which extends the lifecycle framework of Stecher [20], with CI replacing the label-derived selectivity indices of the original. Thresholds: root ≥ 10, mid $\in [1, 10)$, leaf $= 0$.

- **Assembly** (As-E): feature unmatched at T, active at $T + 1$ ("born").
- **Decay** (De-E): feature active at T, unmatched at $T + 1$ ("died").
- **Task-general** (Tg-E): matched, and either born-at-root ($\mathrm{CI}_{T+1} \geq 10$, backward-$r < 0.2$) or matched-gaining ($z(\Delta\mathrm{CI}) \geq +0.5$, endpoint root).
- **Abstraction** (Ab-E): matched, born-at-mid or matched-gaining, endpoint mid.
- **Differentiation** (Di-E): matched, born-at-leaf or matched-losing, endpoint leaf.

Matched feature-events that do not cross any threshold do not receive a process label; they are "matched background" and are excluded from the process-composition analyses. The per-transition null μ, σ is estimated by permuting activation columns over position in the aligned sample S^* per transition (1,000 permutations).

Per-token event counting. Event counts in Table 7 are reported at the token level rather than the feature level. For each matched feature pair across a checkpoint transition, we record the set of validation tokens on which the event occurs: a Birth (As-E) token is one on which a confirmed-born feature activates; a Death (De-E) token is one on which a confirmed-dead feature fires for the last time; a Task-general (Tg-E) or Abstraction (Ab-E) token is one on which a matched feature fires and the feature's CI endpoint places it in the root or mid tier, respectively; a Differentiation (Di-E) token is one on which a matched feature fires and its CI endpoint places it in the leaf tier. Counts are summed across all features and all layers for each phase window. Token-level counts are a more direct measure of the representational work done at each phase than feature counts, since a single broad feature contributes many tokens per transition whereas a narrow one contributes few.

RDT event taxonomy and per-window counts. Per-phase token-level event counts at both scales (Phases I/II/III as defined in §3.1) appear in Table 7.

Lag-1 precedence analysis on Phase I sequences. We assess whether task-general (Tg-E) events at one checkpoint precede abstraction (Ab-E) events at the next, and whether Ab-E events precede differentiation (Di-E) events, via lag-1 Pearson correlations on $\log_{10}(\text{event count} + 1)$ sequences across the Phase I transitions. Results are reported across eight (scale, layer) cells; forward correlations and their reverses are compared to assess directionality. Full per-cell values are in App. C.

Broad-born survival. Broadly-firing features survive at 72× the rate of narrowly-firing features (43% vs. 0.6% to step 143k at Pythia-160M; 38% vs. 0.4% at Pythia-410M), stratified by firing-pattern entropy quartile at step 1k.

Pre-existing structure at random initialisation. Phase I amplifies pre-existing structure rather than building the scaffold from scratch: about one third of eventual carriers (35% at Pythia-160M, 31% at Pythia-410M) are already weakly active at step 0 (above the alive threshold on 0.1% to 1% of validation samples).

Robustness controls. The substrate replicates across seeds, scales, and SAE budgets. Across three additional Pythia-160M seeds and two Pythia-410M seeds, persistent-carrier counts vary by $\leq 12\%$, the step-1k birth concentration is preserved (broad-born features account for $\geq 70\%$ of carriers at every seed), and the carrier–NCR load gap is recovered at every replicate. A higher-budget SAE (more samples and steps, same expansion factor 64) yields identical conclusions.

Table 7: Per-phase RDT event counts at Pythia-160M and Pythia-410M, summed across the four instrumented layers, reported in thousands of token-level events (see *Per-token event counting* paragraph above). Each event is counted once per (feature, token) pair at the transition where it occurs. Task-general (Tg-E) and Abstraction (Ab-E) are reported separately; their sum equals the total broadening activity at each phase. Births dominate Phase I; differentiation and decay dominate Phase II; Phase III is structurally stable. †Elevated by the L11 extinction event at the 80k→120k transition; see App. E.

	Pythia-160M			Pythia-410M		
Event	Phase I	Phase II	Phase III	Phase I	Phase II	Phase III
Birth (As-E)	1,288	154	91	1,414	287	82
Death (De-E)	225	77	366†	306	115	136
Task-general (Tg-E)	723	51	224	623	50	167
Abstraction (Ab-E)	277	12	41	408	29	43
Differentiation (Di-E)	339	154	268	431	279	173

D Pivot Details: Ablation and Geometry

Function converges orders of magnitude faster than direction; the load asymmetry is invisible to per-firing diagnostics; and scaffold edges sit 45× above the random-pair null on decoder alignment. This appendix collects the quantitative details: function-versus-direction convergence curves, per-tier ablation distributions, per-layer birth-window lesion ratios, and decoder-cosine alignment on scaffold edges against three nulls.

Function and direction convergence. Function settles by step 1k; decoder direction takes the full training span. Function distance (Pearson $1 - r$ of per-token activation profiles) reaches its asymptotic floor within 5% of the step-120k→ 143k value by step 1k at both scales. Direction distance (cosine distance between decoder columns) decays roughly linearly on a log scale across the full $0 \rightarrow 143$k span. The two-orders-of-magnitude gap in calibration time is robust to scale and layer.

Per-firing ΔCE across scope tiers. Per-firing single-feature ablation sees no difference between tier classes: the ∼91× load gap is invisible to it. The three tier-conditional distributions of per-firing ΔCE are statistically indistinguishable (two-sample KS, $p > 0.2$ for all three pairs at both Pythia scales), where tiers are assigned by position in the cross-layer DAG. Load only emerges under joint cross-layer intervention (§3.2).

Per-layer A/B/C lesion ratios. Step-1k scaffold seeds outperform outside-DAG controls by 10^3–10^4× per-feature load at every Pythia-410M layer. Comparing three count-matched groups (A: step-1k scaffold seeds; B: step-143k specialists; C: outside-DAG controls) via per-feature ΔMSE on the residual stream: L4: A/C = 4,105, A/B = 6.3. L11: A/C = 1,999, A/B = 0.41 (the only layer where step-143k specialists exceed step-1k seeds). L17: C is at the noise floor (mean 7×10^{-7} vs. A's 4×10^{-2}), making A/C numerically unbounded; A/B at L17 is 6.7. L23: A/C = 2,984, A/B = 1.2. A/C exceeds 10^3 at every layer.

Seedling counts and the L17 anomaly. Scaffold load at L17 has no formal seedlings: persistence is sufficient but not necessary for load. At Pythia-160M, three layers have non-zero seedling counts

(22, 11, 4; a seedling is a step-1k feature whose lineage matures into a persistent carrier); L11 has zero. At Pythia-410M, L17 also has zero seedlings yet contributes nonzero load when its step-1k seed features are jointly ablated: broad-entropy features the predictor identifies carry load, but none meet the ≥ 4-of-6-snapshot lineage criterion. The step-1k firing screen detects the right population; the persistence criterion, not the predictor, is the limiting condition. A directly testable prediction follows: L17's load should be recoverable from non-seedling precursors already active at step 1k.

Decoder-cosine alignment on scaffold edges. Scaffold edges show $45\times$ higher decoder-column alignment than random carrier pairs, and the gap survives layer-frequency controls. Observed edges sit $45\times$ above a random-pair null, $12\times$ above a layer-matched random null, and $4.3\times$ above layer-matched non-edges on median absolute cosine $|\cos(d_u, d_v)|$ at Pythia-410M; rank ordering is identical at Pythia-160M. The alignment is not a layer-frequency artefact and not driven by all CI-related pairs.

E The L11 Extinction Event: A Layer Fitness Case Study

A single geometric event destroys Pythia-160M-L11's scaffold capacity between steps 80k and 120k, invisible to global loss curves but flagged by every substrate readout. Whole-network validation loss decreases monotonically across the same window; only the substrate probe detects the event. The three apparent "independent failure channels" are three readouts of one geometric mechanism. In the life-history framework, L11 is a case where the substrate conditions for carrier formation never stabilise: a falsifiable prediction the data verify.

Residual-stream norm and max-abs trajectories. L11 alone breaks the canonical late-training trajectory: its norm rises while every other layer's falls, and its max-abs collapses $7\times$ in the extinction window. Per-layer residual-stream statistics on $3{,}000$ validation positions (Table 8, independent of any SAE): every other 160M layer's mean-norm decreases monotonically from step 40k onwards; L11 increases from $62 \to 99 \to 108$ after step 80k. Simultaneously, max-abs activation (the outlier-channel amplitude) collapses $7\times$ at L11 ($108 \to 15$) while other layers decline smoothly. The depth-matched control, Pythia-410M-L23, shows the opposite: smooth factor-of-three max-abs decline with no discontinuity.

Table 8: Residual-stream ℓ_2-norm (mean over tokens) and max-abs activation per checkpoint for four Pythia-160M layers. L11 is the sole layer whose norm rises after step 40k while its max-abs collapses $7\times$ in the cliff window (80k→120k).

	Mean ℓ_2-norm				Max-abs activation			
Step	L2	L5	L8	L11	L2	L5	L8	L11
10k	30.8	40.4	53.1	83.9	57	292	301	105
40k	33.4	40.9	55.9	93.9	182	624	655	138
80k	22.8	26.8	40.1	62.0	166	492	517	108
120k	15.7	20.4	29.1	**98.8**	127	328	321	**15.3**
143k	13.6	17.8	24.5	**108.0**	115	290	283	**15.0**

Spectral analysis: effective rank trajectories. L11's effective rank collapses continuously through the extinction window ($84 \to 28 \to 13$) while all other 160M layers plateau by step 40k. Effective rank ($\exp(H(p))$ from SVD on $4{,}000$ validation residual streams; Table 9) confirms the collapse is ongoing, not a transient. Critically, Pythia-410M-L11 reaches rank 6 by step 143k — even more compressed — yet its SAE works (loss recovery $+0.58$) and it contributes scaffold carriers. Low effective rank alone does not break the scaffold; the 160M-L11 failure requires rank collapse accompanied by outlier-channel disappearance and unembedding-subspace absorption, which 410M-L11 does not show.

Table 9: Effective rank of residual-stream activations for 160M layers (left block) and 410M layers (right block). 160M-L11 continues collapsing across the extinction window while all other 160M layers plateau by step 40k. 410M-L23 (depth-matched negative control) retains effective rank > 150 at step 143k.

	Pythia-160M				Pythia-410M			
Step	L2	L5	L8	L11	L4	L11	L17	L23
10k	378	220	276	191	471	160	278	319
40k	298	59	98	85	504	22	56	174
80k	257	47	78	84	440	11	27	152
120k	224	55	73	**28**	382	7	14	195
143k	205	55	78	**13**	367	6	14	181

Three convergent readouts as one geometric event. The three SAE-substrate signals are not independent channels; all three fall out of one mechanism. Active feature count drops $\sim 426 \rightarrow \sim 37$ (robust to breadth threshold $\in [0.1\%, 5\%]$); the same ~ 37 features fire on every token. Decoder geometry fails: alive L11 features carry only 39% of their column variance in the unembedding subspace where the residual stream now concentrates at 84%; the SAE cannot route enough budget into that subspace to track. Joint-load collapses: replacing L11 with its SAE reconstruction makes the model worse than mean-ablating the layer (loss recovery -0.33, vs. $+0.33$ at the depth-matched 410M-L23). The unifying mechanism: outlier-feature channels disappear, effective rank collapses to ~ 13 dimensions, and surviving variance rotates into the W_U row space. A $k = 32$ top-k SAE cannot tile a near-isotropic 13-dimensional distribution; its dictionary collapses to a fixed core of ~ 37 features. The same mechanism produces decoder$\rightarrow W_U$ alignment and joint-load failure.

SAE training-budget control. A $10\times$ higher SAE training budget at Pythia-160M-L11 ($5{,}000$ steps $\times$ batch $256 = 1.28$M tokens vs. 128k canonical) yields 38 vs. 37 active features; loss recovery rises from -0.33 to 0.0099, essentially zero. The collapse is intrinsic to L11's late-training distribution, not SAE undertraining.

Depth-matched negative control: 410M-L23. Pythia-410M-L23 is the principal negative control (stronger than 410M-L11, which is mid-network at $2.6\times$ scale). At step 143k, 410M-L23 shows the opposite late-training trajectory: smooth de-alignment from W_U (top-200 PCA: $0.74 \rightarrow 0.40$ across training), ~ 580 active SAE features throughout, and loss recovery $+0.33$. The 160M-L11 failure mode is layer-specific to the smaller scale, not a generic last-residual-stream-layer phenomenon.

Graph integration. Graph isolation at 160M-L11 is complete: zero outgoing scaffold edges at step 143k. By contrast, Pythia-410M-L11 contributes 18 edges (11 to L17, 7 to L23), with forward-pass direction in 17 of 18 cases.

Geometric coherence. Pythia-410M-L11 integrates normally into scaffold geometry despite its low effective rank: per-edge decoder cosine median 0.41 (95% CI $[0.37, 0.46]$), within the global scaffold-edge median (0.43), and its unembedding-cosine trajectory is stable across training (Fig. 7, dashed trace), in contrast to the absorption at 160M-L11.

Joint-ablation load. Joint ablation of persistent carriers at Pythia-410M confirms L11 participates in scaffold function, in contrast to 160M-L11, which contributes zero carriers at the same nominal depth.

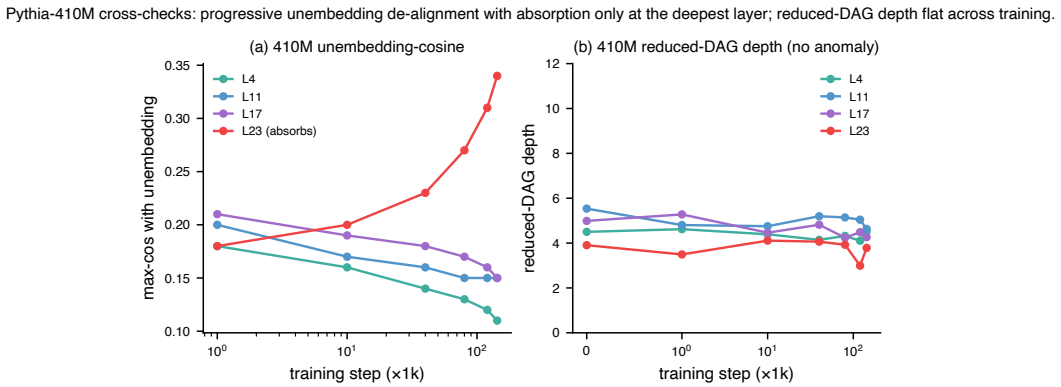


Figure 7: Pythia-410M per-layer cross-checks for the unembedding-cosine and within-layer DAG-depth analyses of App. E. (a) Maximum cosine of layer features with the unembedding row space across L4/L11/L17/L23: progressive de-alignment at L4/L11/L17 with absorption at L23, the canonical scaffold-integrated pattern. (b) Within-layer reduced-DAG depth across the four layers: flat across training, with no anomalous excursion in the step 10k$\rightarrow$ 80k window. The 160M-L11 failure mode is layer-specific to the smaller scale.

F Predictor Supplements

The predictor achieves $9.3\times$ enrichment over base rate at the top-decile operating point and generalises across held-out layers. Details: top-decile precision sweep across thresholds and the full coefficient table for the six step-1k firing-space predictors.

Top-K precision sweep. At the top-$1\times n_{\text{pos}}$ operating point the predictor achieves $9.3\times$ enrichment over base rate — the highest-precision window for training-time triage. The Pythia-160M predictor is trained on 3,198 active step-1k features, of which 37 are persistent carriers (base rate 1.2%; AUC $= 0.80$, §3.3). Top-K precision by predicted probability: top-37 gives 10.8% precision ($9.3\times$); top-74: 6.8% ($5.8\times$); top-185: 5.9% ($5.1\times$); top-370: 4.6% ($4.0\times$). Precision decays toward base rate as the window widens.

Coefficient table. The classifier generalises across layers: leave-one-layer-out AUC stays at 0.74–0.79 at Pythia-160M, including when L11 is held out (0.76), confirming it does not memorise layer identity. Standardised logistic-regression coefficients with five-fold cross-validated 95% intervals are in Table 10.

Table 10: Logistic-regression coefficients for the six step-1k firing-space predictors (firing breadth, maximum activation, mean conditional activation, activation variance when firing, Containment Index, and root flag CI ≥ 10); standardised, five-fold cross-validated 95% intervals. Firing breadth and CI carry the predictive load.

Predictor	Coefficient (standardised)	5-fold 95% CI
Firing-pattern entropy	$+0.71$	$[+0.58, +0.83]$
Containment Index	$+0.64$	$[+0.51, +0.77]$
Activation strength	$+0.42$	$[+0.28, +0.56]$
Cross-layer reach	$+0.31$	$[+0.18, +0.45]$
Layer-normalised activation rank	$+0.07$	$[-0.06, +0.20]$
Lifespan-so-far	$+0.04$	$[-0.09, +0.18]$

NeurIPS Paper Checklist

1. **Claims.** Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

[Yes] The abstract and Introduction Contributions describe findings scoped to Pythia-160M and Pythia-410M at four residual-stream layers each. The Limitations section (§5) makes the two-scale scope explicit; multi-seed robustness controls are in App. C; the L11 substrate case study (App. E) is reported as 160M-specific.

2. **Limitations.** Did you create a separate limitations section discussing the assumptions, robustness to violations, scope of claims, and performance factors?
[Yes] §5 is a dedicated section placed after the Discussion and before the appendix. It covers the two-scale single-seed scope (multi-seed robustness controls at App. C), residual-stream-only coverage at four layers per scale, the absence of a construction-side intervention, and the 160M-L11 substrate failure (App. E). Auto-interp confidence-rating sample size (20 carriers) is discussed in App. A; the per-firing tier-invariance confound is formalised in App. D.

3. **Theory, Assumptions and Proofs.** If you are including theoretical results, did you state the full set of assumptions of all theoretical results, and did you include complete proofs of all theoretical results?
[N/A] The paper contains no formal theoretical results; all claims are empirical.

4. **Experimental Result Reproducibility.** If the contribution is a dataset or model, what steps did you take to make your results reproducible or verifiable?
[Yes] All code and per-snapshot JSON outputs at both scales are released with the submission. §2 specifies models, checkpoints, layers, SAE hyperparameters, aligned sample construction, CI thresholds, matching criteria, and ablation protocol. Full methodological details and robustness controls (independent-seed replication, higher-budget SAE) are in App. C.

5. **Open Access to Data and Code.** If you ran experiments, did you include the code, data, and instructions needed to reproduce the main experimental results?
[Yes] All code and per-snapshot JSON outputs at both scales are released with the submission. Pile-validation, Pythia-160M, and Pythia-410M checkpoints are already public.

6. **Experimental Setting/Details.** If you ran experiments, did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
[Yes] §2 specifies models, data, SAE hyperparameters, aligned sample construction, CI edge criteria, matching protocol, and ablation protocol. Appendices expand on full methodological details (App. C), pivot geometry and per-tier ablation distributions (App. D), semantic anatomy and blind LLM validation (App. A), the L11 substrate case study (App. E), and predictor details (App. F).

7. **Experiment Statistical Significance.** Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
[Yes] Log-rank tests for tier-survival curves are Bonferroni-corrected across tier pairs (App. B); Granger F-tests for Phase I sequence directionality report exact statistics and p-values (App. C); KS tests for per-firing ΔCE tier-invariance carry $p > 0.2$ bounds (App. D); ablation ΔCE values and logistic-regression coefficients carry 95% CIs throughout.

8. **Experiments Compute Resource.** For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?
[Yes] §2 specifies that each SAE trains for $4{,}000$ steps (expansion factor 64, $k = 32$) on a held-out Pile slice; all analyses are run at laptop scale.

9. **Code Of Ethics.** Have you read the NeurIPS Code of Ethics and ensured that your research conforms to it?
[Yes] The work uses only the public Pythia-160M and Pythia-410M models and the Pile validation set; no human subjects, no sensitive-data release.

10. **Broader Impacts.** If appropriate for the scope and focus of your paper, did you discuss potential negative societal impacts of your work?
[N/A] This is a mechanistic interpretability analysis of public, already-released language models using only analysis code and JSON summaries. No new model capabilities, datasets, or deployment artefacts are introduced; no direct negative societal impact pathway is identified. Positive implications for safety diagnostics are noted in §4.

11. **Safeguards.** Do you have safeguards in place for responsible release of models with a high risk for misuse?

[N/A] No pretrained models or datasets with high-misuse risk are released; only analysis code and per-snapshot JSON summaries.

12. **Licenses.** If you are using existing assets (e.g., code, data, models), did you cite the creators and respect the license and terms of use?
    [Yes] Pythia-160M and Pythia-410M [3] are released under Apache 2.0; the Pile validation set is CC-BY. All are cited with their original publications.
13. **Assets.** If you are releasing new assets, did you document them and provide these details alongside the assets?
    [Yes] The code repository and JSON summaries released with the submission include per-layer / per-checkpoint documentation.
14. **Crowdsourcing and Research with Human Subjects.** If you used crowdsourcing or conducted research with human subjects, did you include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?
    [N/A] No crowdsourcing or human-subjects research is used. The blind validation of App. A is LLM-based and does not constitute human-subjects research.
15. **IRB Approvals.** Did you describe any potential participant risks and obtain Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your institution), if applicable?
    [N/A] No human subjects research.
16. **Declaration of LLM usage.** Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?
    [Yes] A GPT-class model was used for the carrier auto-interpretation step following the Bills et al. [17] recipe (App. A); confidence labels (high/medium/low) are machine-assigned. A blind second-LLM validation of 20 sampled carriers (App. A) gives $18/20$ concept agreement with auto-interp. We used large language models to assist with writing and editing.