THE TWO CLOCKS OF REPRESENTATION LEARNING IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation learning runs on two clocks. The token clock, which tokens a feature answers to, is largely done within the first percent of pretraining. The wiring clock, how it answers within those tokens, calibrates to the end of the run, in step with the direction the feature writes. We track sparse-autoencoder features across checkpoints of three Pythia scales and OLMo-2-1B, splitting each activation profile exactly into token-set and context parts. By step 1k the token set reads $r = 0.71$–0.78 against its final form, while context selectivity and direction stand at 0.22–0.42 and move together to the end. The window the two clocks open is where a feature's mature role becomes readable. Containment, the nesting of one feature's firing set inside another's, ranks at step 1k which persisting features will organise the mature network, at AUC 0.74–0.89 with nothing fitted, while which features persist stays near chance. Those persistent, high-containment features are the model's carriers. They cost 1.7–$1.9\times$ as much to remove as populations matched on layer, breadth and activation energy, and ablating a containment parent cuts its child's firing by 11% at the two smaller Pythia scales.

1 INTRODUCTION

What does it mean for a representation to have formed? Most analyses of language-model representations observe the trained model alone, and so only the endpoint. A trained model shows what a feature became. Which of its properties were present early, which were acquired later, and whether they settled together are readable only by following individual features through training.

Take a feature that answers to the word *firm*. Three things describe it. It fires on *firm* rather than on other tokens. Among occurrences of *firm* it prefers *law firm* to *firm handshake*. And whenever it fires it writes one vector into the residual stream. The first two are its *function*, properties of the input. The third is its *direction*, a property of the output. The natural expectation is that function settles on one schedule and direction on another.

Training draws the line elsewhere, and it runs through function. At the first checkpoint, which tokens a feature answers to, its *token set*, is already near its converged form, while how it chooses among contexts, its *context selectivity*, lags far behind and catches up on the schedule of the direction. The two rates are the *token clock* and the *wiring clock*. Across three Pythia scales and OLMo-2-1B the token set reads $r = 0.71$–0.78 against its converged form, context selectivity 0.22–0.41 and the basis-aligned direction 0.27–0.42. Context and direction then move together feature by feature and are still turning at 84% of training (§3.1).

We read the clocks by tracking sparse-autoencoder (SAE) features across pretraining (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2025). At each checkpoint we train an independent SAE, match features across time by their activation profiles rather than by feature index or decoder direction, and split each profile exactly into a per-token mean and an orthogonal context remainder (§2.3). That record, function and direction followed across checkpoints, is a feature's *life history*. The token set a feature settles on is largely input-constrained, since sibling runs trained from different initialisations form the same token sets by step 1k. Roughly a third of persistent features at the Pythia scales trace to a single token a random network already detects (Heap et al., 2025), with their context selectivity at chance (App. B).

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

Recent work tracks SAE features across checkpoints through crosscoders (Lindsey et al., 2024; Ge et al., 2026; Bayazit et al., 2026) or warm-started dictionaries (Xu et al., 2024), and Inaba et al. (2025) and Sawmya et al. (2025) read population statistics instead. These studies establish that features keep changing through pretraining, with most directions turning sharply near step 1k and rotating slowly afterwards (Ge et al., 2026), and that a feature's labelled meaning settles well before its direction, which keeps drifting into late training (Xu et al., 2024). Bayazit et al. (2026) date when a feature becomes causally important. We ask which part of it is fixed by then. At the layer level, Raghu et al. (2017) find representations converging bottom-up. What looks like bottom-up convergence is two clocks read as one. Shallow features carry more token set and more of them trace to initialisation, so their profiles settle early while on the wiring clock they calibrate last (§3.1). Tigges et al. (2024) find circuits consistent across training and scale.

A second literature reads training as a sequence of stages, from critical periods (Achille et al., 2019; Kleinman et al., 2024) through sudden transitions in loss or circuitry (Olsson et al., 2022; Chen et al., 2024) to developmental stages located by the local learning coefficient (Hoogland et al., 2025). Those accounts ask when the model changes. The two clocks ask which part of a feature changes first. Aoyama et al. (2026) forecast the emergence of induction heads (Olsson et al., 2022) from statistics fixed before training begins, whereas the gap between our clocks opens an early window on mature structural position.

Containment, the nesting of one feature's firing set inside another's, is the relational readout that makes structural position legible inside that early window (§2.2). Tracing life histories to convergence picks out a persistent, high-containment population, the *carriers*, collectively a model's *scaffold*. Carriers are a statistical stratum. Their counts, load and organisation replicate under independent dictionaries, while exact membership shifts from one to the next (Leask et al., 2025; Paulo & Belrose, 2026).

Contributions. Four things are new.

1. **The token set settles first, and context selectivity, an input property, settles with the output direction (§3.1).** The token set of persistent features is near its converged form at the first checkpoint, and at the Pythia scales for roughly a third of features at initialisation, while context selectivity and direction stand together far from theirs and move together to the end, at all four models and against nulls matched to each clock.
2. **Early containment predicts mature position among persisting features, and survival stays near chance (§3.2).** At the first checkpoint containment already ranks which persisting features will end as roots, beyond what breadth gives, and the signal is not carried by the settled token set. The ranking holds under a trackability control at three of the four models. Which features persist is near chance from containment at every model, and from breadth at all but OLMo-2-1B, where breadth reaches 0.61.
3. **Carriers are disproportionately load-bearing (§3.3).** Removing the carriers jointly raises next-token loss by 1.7–$1.9\times$ as much as removing non-carrier populations matched on both firing breadth and activation energy, with disjoint intervals at every model, and the ordering reproduces under independent SAE initialisation, though at Pythia-410M the estimate depends on which control is matched (App. D).
4. **Ablating a containment parent moves its child, in the SAE readout (§3.4).** Ablating a containment parent reduces its child's firing by 11% at the two smaller Pythia scales, against no median effect for matched controls, with most of that coupling already present in the opening phase.

2 METHOD

We instrument three Pythia scales across seven training checkpoints and train one SAE per (layer, checkpoint) pair. We match features across time to build life histories and a cross-layer containment hierarchy, then measure load by joint cross-layer ablation. A training-time predictor quantifies how early a surviving feature's mature position is identifiable from function alone. Full methodological detail is in App. A, and Table 1 collects the terms used throughout.

Table 1: Terms used throughout the paper, with the section that defines each.

Term	Meaning
Token set	the tokens a feature answers to, read as its mean activation per token (§2.3)
Context selectivity	its response within a token across contexts, what remains after the per-token mean (§2.3)
Direction	the decoder column through which it writes into the residual stream (§2.1)
Token clock	how far the token set has converged to its final form (§3.1)
Wiring clock	the same for context selectivity and direction, which move together (§3.1)
Containment, CI	the parent fires wherever the child fires, and CI counts the features beneath one (§2.2)
Carrier, scaffold	a feature persistent through training that ends as a root, and all of a model's carriers (§2.2)
Load	the added next-token loss from projecting out a population's decoder columns (§2.3)

2.1 Models, data, and sparse autoencoders

The three scales are Pythia-160M, -410M, and -1B (Biderman et al., 2023). Training runs from step 0 to step 143k and is concentrated early, with steps 0–10k covering $\sim$7% of training but most feature life events (schedule and layers in Table 4). Four residual-stream layers per scale are instrumented at approximately uniform fractional depth. All SAE training, matching, and ablation use a held-out Pile slice (Gao et al., 2020) not seen during pretraining. Feature-level statistics use a canonical validation sample S^* of $20{,}000$ token positions. OLMo-2-1B (Team OLMo, 2025; Allen Institute for AI, 2025), which differs from Pythia in architecture, tokeniser, and training corpus, runs through the identical pipeline as a cross-architecture replication, with its checkpoint schedule and remaining setup details in App. F.5.

At every (checkpoint, layer) pair we train a top-k SAE (Gao et al., 2025; Bricken et al., 2023; Sharkey et al., 2022) on residual-stream activations: $\hat{x} = W_{\text{dec}}\,\text{TopK}(W_{\text{enc}}\,x + b_{\text{enc}}) + b_{\text{pre}}$, with loss $\mathcal{L} = \|x - \hat{x}\|_2^2$. We use k=32 simultaneously active features, expansion factor 64 over the residual width ($d_{\text{model}} = 768/1024/2048$ at 160M/410M/1B and 2048 at OLMo-2-1B), and decoder columns unit-normalised at each step. Single-layer decoder insertion preserves 0.75–0.92 of mean-ablation-referenced loss everywhere except the failed 160M-L11 substrate (Table 5). A feature is *alive* at a checkpoint if it activates on at least 0.5% of S^* samples.

2.2 Feature tracking and the containment hierarchy

At each consecutive checkpoint pair, features are matched one-to-one by Hungarian assignment (Kuhn, 1955) on the Pearson correlation of their activation profiles over S^*, and a match is accepted when $r \geq 0.5$. Because matching is one-to-one, feature splits and merges are scored as births and deaths, a limitation the event taxonomy inherits. Each transition classifies a feature as *stable* (matched), *born* (unmatched, newly alive), or *died* (unmatched, no longer alive). A *lineage* is a maximal chain of stable matches from birth, a functional equivalence class under the matching rule, and "a feature" in this paper always means a lineage rather than a coordinate in any one SAE. Birth and death follow the developmental taxonomy of Stecher et al. (2026).

A feature is a *carrier* if it is present and chain-matched in at least 4 of the 6 post-initialisation snapshots and ends in the high-containment tier at convergence ($\text{CI} \geq 10$, defined below). A model's carriers together form its *scaffold*.

Containment is a relational readout of functional identity, built from firing token sets. An edge $u \to v$ records that the narrower feature v fires almost only where the broader feature u fires, far above chance. Informally, the parent fires broadly and the child within a subset of those contexts. We call u the *containment parent* and v the *child*. The *Containment Index* $\text{CI}(f)$ counts the features reachable from f in this DAG. High-CI features are *roots*, features with nothing beneath them *leaves*, and the remainder *mids*. *Sub-DAG coverage* at step t is the fraction of active features reachable from at least one carrier. *Firing breadth* is the fraction of S^* tokens on which a feature activates.

Membership is dictionary-relative. Independent SAE seeds recover the scaffold's statistical signature (counts, load, the same substrate failure), and 98–100% of carriers still have a counterpart at $r \geq 0.5$ in an independently initialised dictionary (median best r 0.74–0.76, App. F). What does not survive is exact one-to-one membership under our strict multi-checkpoint tracking criterion (Jac-

card 0.12–0.17, App. F), consistent with evidence that SAEs do not find canonical units (Leask et al., 2025; Paulo & Belrose, 2026). Carrier claims are therefore population-level claims about a statistical stratum rather than about a recoverable feature list. A second line of work questions the construct itself, since coverage admits pairs whose parent and child are unrelated and sparsity over hierarchical features absorbs a parent into its children (Cao et al., 2026; Chanin et al., 2025; Bussmann et al., 2025; Costa et al., 2025; Grandien et al., 2026). Both mechanisms act among the latents of one dictionary, so App. F.2 reads the DAG with that channel deleted.

2.3 Convergence clocks, ablation load, prediction, and causal mediation

We call a population *load-bearing* when removing it causes a larger increase in next-token prediction loss. We read load at the population level rather than per feature. For a target population, we simultaneously project out every member's decoder contribution across all four instrumented layers from the unmodified model, then measure the aggregate $\Delta\mathrm{CE}$ on a held-out validation pass, with 95% bootstrap intervals over $1{,}000$ resamples of 32 held-out evaluation chunks. This clean-model projection is the primary readout because its baseline is dictionary-independent. An SAE-roundtrip variant serves as the comparison, and the carrier-versus-control ordering agrees across both readouts. Projecting out many non-orthogonal components pushes activations off-manifold, so no cost is interpreted in isolation, and every control undergoes the identical intervention. Five count-matched control populations isolate confounds: non-persistent roots (structural tier), matched leaves (the non-hierarchical baseline), and non-carriers matched on firing breadth, on total activation energy, and on both, which hold layer and count fixed while varying persistence. A decoder-transplantation control (each carrier's converged decoder column replaced by its step-1k value) isolates the load owed to direction calibration (full detail on the readout, its variants, and all controls above in App. D).

For §3.1, function convergence is the Pearson correlation of a feature's activation profile over S^* to its step-143k profile and direction convergence the cosine of its decoder column to its step-143k column. The profile splits exactly into a *token-set part*, the feature's mean activation per token, and a *context part*, the remainder, orthogonal to it, so the raw correlation is the sum of a token-set and a context contribution and each part has its own clock (App. B). For §3.2, a five-fold cross-validated logistic regression predicts step-143k carrier identity from six step-1k firing-space properties (App. B).

Containment (§2.2) is an observational definition. To test whether the relation is causally coupled, rather than an artefact of co-occurrence or shared geometry, we ablate the parent feature u and measure the effect on the child v's firing, at the containment edges whose parent lies in a shallower layer than its child, against a control edge matched on breadth and Containment Index but without a containment relation to the same child. Edges inside one layer, and edges whose parent lies deeper than its child, are treated separately (App. E). Firing is measured within the SAE encoder, so this is SAE-roundtrip-conditional coupling and makes no claim about clean-model behaviour.

Four further controls (lift-matched sources, a cross-layer decoder-cosine graph, an independently initialised SAE, held-out token chunks) and a per-checkpoint trace of the coupling's onset are in App. E.

3 Results

Each result supplies the premise for the next: the clock gap opens a window in which the token set has settled while the wiring clock is still turning (§3.1), §3.2 reads that window, §3.3 shows that what it selects is load-bearing, and §3.4 that the selected structure transmits load.

3.1 The token set settles before context selectivity and direction

Across all persistent features, the token set settles first, and context selectivity settles with the direction rather than with the token set (Fig. 1a–b). Both halves of function read the input, so both should settle on one schedule. Context selectivity settles on the direction's instead, and the two together are the wiring clock. An early token set would follow from token statistics alone, so the finding is the schedule of the rest of function.

Table 2 gives the step-1k readings. The token set stands at $r = 0.71$–0.78 of its converged form, while context selectivity reads 0.22–0.41 and the basis-aligned direction 0.27–0.42. The token set

Table 2: Step-1k readings relative to step 143k, per-model means over the 437/489/528/586 features alive at step 1k: function whole and split into its token-set and context parts, and direction raw and after an orthogonal map of the step-1k basis onto the step-143k basis. Brackets give the 95th percentile of each reading's own matched null (App. B). The direction-matched rows rebuild the features on decoder cosine.

Reading at step 1k [own null]	Pythia-160M	Pythia-410M	Pythia-1B	OLMo-2-1B
All persistent features: function r, whole	0.59 [0.04]	0.58 [0.04]	0.57 [0.03]	0.54 [0.03]
token-set part	0.74 [0.13]	0.78 [0.12]	0.71 [0.12]	0.71 [0.14]
context part	0.22 [0.03]	0.27 [0.02]	0.32 [0.02]	0.41 [0.02]
All persistent features: direction cosine, raw	0.20 [0.07]	0.25 [0.07]	0.30 [0.05]	0.10 [0.04]
basis-aligned	0.27 [0.06]	0.33 [0.05]	0.42 [0.03]	0.40 [0.04]
Direction-matched chains: function r, whole	0.52 [0.07]	0.57 [0.05]	0.69 [0.04]	0.68 [0.03]
Direction-matched chains: direction cosine	0.39 [0.07]	0.43 [0.06]	0.47 [0.05]	0.20 [0.04]

leads in 86–95% of individual features. Every gain in function after the first checkpoint is contextual (App. B).

Context selectivity and direction are one clock rather than two. At every checkpoint from step 1k to step 120k the two readings sit within 0.10 of each other at every model, and from step 10k on they correlate feature by feature at $r = 0.43$–0.78. At step 120k, 84% of the way through training, both still read 0.6–0.75 while the token clock reads 0.86–0.91 (App. B).

The token set is largely input-constrained. Traced one hop back to step 0, 42/41/21% of persistent features at Pythia-160M/-410M/-1B are already seeded by a single token that a random network detects (Heap et al., 2025). Their token-set part reads $r = 0.77$–0.90 against its converged form at that step, while their context part is near chance. The seed is kept to convergence while the feature widens into a category around it. Features without a seed form their token set from context by step 1k and then read the same clocks (App. B).

The clocks also order depth, and the split reverses that ordering. Read whole, function converges bottom-up at step 1k, as Raghu et al. (2017) found for whole layers. Split, the shallowest layer's context selectivity is the furthest from converged, trailing the deepest by 0.17–0.22 at three models and by 0.07 at Pythia-1B, and it still trails at step 120k at three of the four, while the token clock is flat across depth (App. B, Table 8).

The lead survives a change of matching rule. Chains are built by firing-pattern matching, so a firing-based clock could look settled by construction. Rebuilt on decoder cosine instead, every model still puts function ahead of direction at step 1k (Table 2, lower block), as the token set predicts. Each clock also clears its own matched null at step 1k by a wide margin, from 4 null standard deviations for direction up to 14 for context selectivity. Reading each clock against its own identifiability ceiling narrows the step-1k gap and leaves it open. Carriers, 2–6% of the population, run the same schedule on every clock (App. B). What marks them out is containment position (§3.2) and load (§3.3).

3.2 Early containment predicts mature position, while survival stays near chance

Step 1k is where the churn ends, and so where an early reading can be taken. The per-step rate of feature life events falls 36–38× within the first ∼1% of training once the dictionary-noise floor is subtracted (22–29× uncorrected, Fig. 2a, and after correcting for warm-up, App. C). By that step most eventual carriers are already present, a fraction that falls as scale grows (Table 3).

The carrier label has two parts, persistence through training and a root position at convergence, and containment predicts them differently. Ordering step-1k features by containment, with nothing fitted, ranks the eventual roots among persisting features at AUC 0.74–0.89. It ranks persistence itself at chance, 0.53–0.58, against 0.46–0.61 for firing breadth. Position in the hierarchy is therefore set inside the first percent of training and kept, while survival is settled later. The same ranking scored against the joint label runs higher still (Fig. 1d), which says in part only that containment rank is

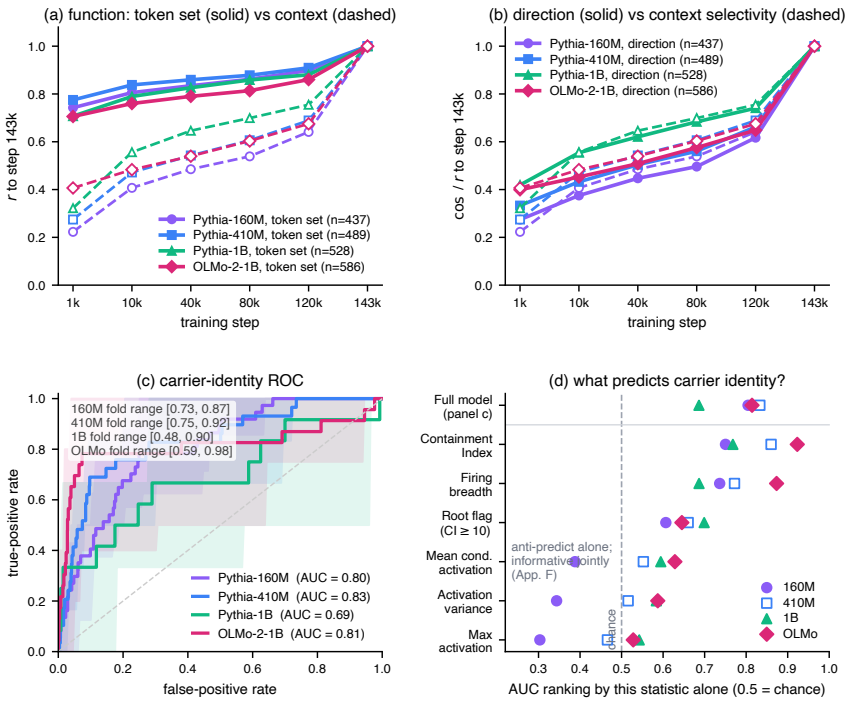


Figure 1: The lag between the token clock and the wiring clock opens an early predictive window. **(a)** Function split by token: mean correlation to step 143k of the token-set part (solid) and the context part (dashed) over all persistent features alive at step 1k ($n = 437/489/528/586$). **(b)** The context part (dashed, as in a) against the basis-aligned decoder direction of the same features (solid). Step-0 readings and the carrier subset are in App. B. **(c)** Five-fold cross-validated ROC for the joint step-143k carrier label (persistence and root position) from step-1k firing statistics. Bands show foldwise min/max TPR (Pythia-1B: 12 positives, fold AUC $[0.48, 0.90]$). **(d)** Univariate AUCs of raw step-1k statistics (0.5 is chance, no model fitted), with the multivariate model shown above. The conditional sign of the activation statistics is discussed under the univariate cross-check in App. B.

stable from step 1k on, since a root position is part of that label. A five-fold cross-validated logistic fit on six step-1k firing statistics reaches AUC 0.69–0.83 (Fig. 1c, Table 3), below the unfitted containment ranking at three of the four models.

Two controls locate the signal in containment itself. Scored only on pairs from the same breadth decile, containment still ranks carriers at 0.64–0.84, while breadth within containment deciles is at chance. Recomputed from per-token means alone, it ranks at only 0.54–0.64 (App. B), so relational position is legible before context selectivity has converged.

The prediction survives a trackability control. Carriers are easier to track across checkpoints than non-carriers, which could inflate the AUC on its own. Holding re-match reliability constant with an independent second dictionary keeps the prediction at three of the four models (App. B). Pythia-1B has too few positives for the control.

This is a ranking rather than a list. At a base rate near one percent the top of the ranking holds only a few carriers (App. B), and a fully prospective test shows what that costs. Selecting the top step-1k features by containment with the persistence label withheld keeps 10 of 51 to step 143k at Pythia-160M and 7 of 49 at Pythia-410M, too few to separate their removal cost from chance (App. D).

3.3 Carriers are disproportionately load-bearing

The scaffold is load-bearing. Removing the carriers jointly costs 1.7–1.9× what removing a non-carrier population matched on both firing breadth and activation energy costs, with disjoint bootstrap

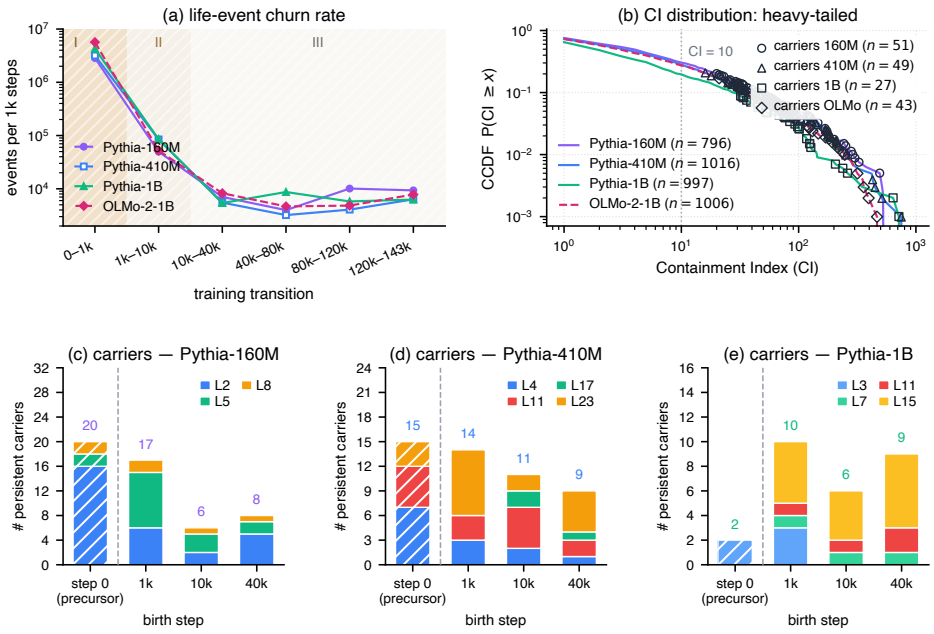


Figure 2: The scaffold assembles within the opening $\sim 1\%$ of training. **(a)** Token-level life-event rate (each event counted once per (feature, token) pair, plotted per 1k training steps) at Pythia-160M, -410M, -1B, and OLMo-2-1B (magenta, dashed, checkpoints fraction-matched to the Pythia schedule) across Phase I (steps 0–1k, dark shading), Phase II (1k–10k, lighter), and Phase III (10k–143k). **(b)** CCDF of the Containment Index (CI) at step 143k on log-log axes, approximately linear over CI $\in [5, 80]$ ($\alpha \approx -0.76$), with the carriers in the right tail. **(c–e)** Carrier birth-step distributions at the three Pythia scales, stacked by layer. Hatched bars mark pre-training precursors already weakly active at step 0 (39%/31%/7% of carriers at 160M/410M/1B). L11 contributes zero carriers at 160M.

Table 3: Summary metrics at the three Pythia scales and, below the rule, OLMo-2-1B. The ΔCE columns give joint decoder-projection costs in nats against the clean-model baseline for carriers and breadth-matched non-carriers, with 95% bootstrap CIs. Sub-DAG gives coverage at step 143k, and AUC gives five-fold CV prediction from step 1k. All values use the shared-initialisation configuration (reseeded and re-derived variants in App. D and App. F).

Model	Carriers	Born $\leq$ step 1k	ΔCE carriers	ΔCE control	AUC (step 1k)	Sub-DAG
Pythia-160M	51	73% (37/51)	+0.48 [+0.44,+0.52]	+0.15 [+0.14,+0.17]	0.80	64%
Pythia-410M	49	59% (29/49)	+0.52 [+0.49,+0.54]	+0.15 [+0.13,+0.16]	0.83	53%
Pythia-1B	27	44% (12/27)	+0.33 [+0.30,+0.35]	+0.14 [+0.13,+0.15]	0.69	29%
OLMo-2-1B	43	53% (23/43)	+0.36 [+0.33,+0.38]	+0.17 [+0.15,+0.19]	0.81	36%

intervals at every model (Fig. 3a). Fig. 3a gives the breadth-only and energy-only matches beside it. Carriers were selected from their life histories alone, with ablation cost held out, so the gap is out-of-sample.

The ordering survives a change of dictionary. Re-initialising all 24 checkpoint SAEs at Pythia-160M lowers the breadth-only ratio there from $3.2\times$ to $1.6\times$, the weakest figure any robustness check produced, and none brought a ratio below $1.5\times$ (App. D).

Persistence carries load of its own at fixed containment tier. Count-matched non-persistent roots cost roughly a quarter of what the carriers cost at Pythia-410M and a fifth at Pythia-1B under the roundtrip harness, with disjoint intervals at both (App. D).

The wiring clock enters here too. Transplanting each carrier's step-1k decoder column into the converged model costs $+0.96$ and $+0.61$ nats at Pythia-160M and -410M, yet a random direction

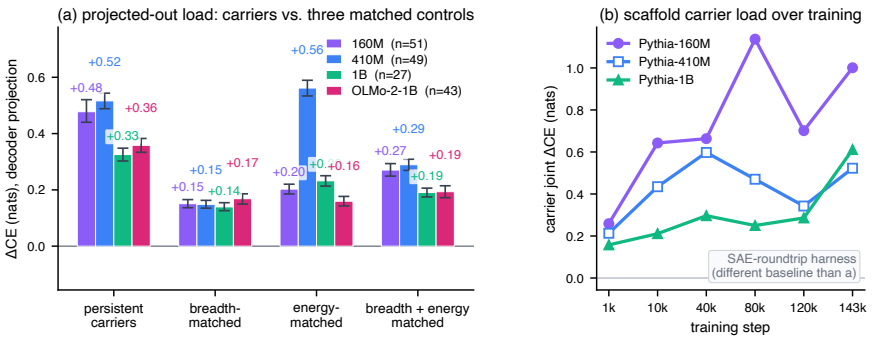


Figure 3: Carriers cost more to remove than every count-matched control. **(a)** Joint ΔCE from projecting out each group's decoder contribution (clean-model baseline) for carriers and three non-carrier controls matched on firing breadth, on activation energy, and on both, at the three Pythia scales and OLMo-2-1B. Carriers cost 2.1–3.5× the breadth-matched control, 0.9–2.4× the energy-matched control and 1.7–1.9× the jointly matched control, the last with disjoint intervals at every model (App. D). Error bars show 95% bootstrap CIs. **(b)** Carrier joint ΔCE over training (SAE-roundtrip harness) for the three Pythia scales. The step 120k dip at Pythia-160M coincides with the L11 reorganisation.

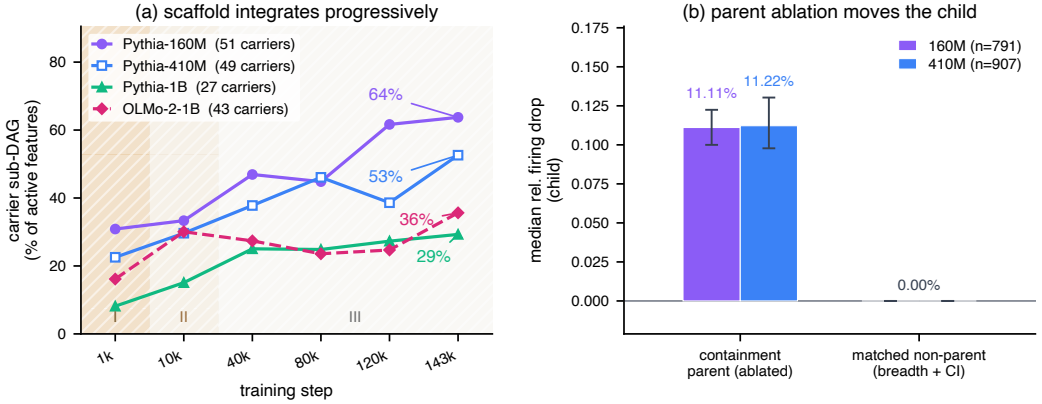


Figure 4: The scaffold organises a large share of the mature network, and its containment relations show causal coupling under feature intervention in the SAE readout. **(a)** Carrier sub-DAG as a percentage of active features across six checkpoints on a log scale, at Pythia-160M, -410M, -1B, and OLMo-2-1B (magenta, dashed), with phase shading as in Fig. 2a. **(b)** Median relative firing drop from ablating each containment parent and measuring its child directly, for the parent-ablated arm against its own matched-control arm, at Pythia-160M (791 edges from 250 parents) and Pythia-410M (907 edges). Error bars are 95% bootstrap CIs on the median. Firing breadth inside and outside the sub-DAG is in Fig. 7.

at the same angular distance costs the same, so the calibration carries real cost and no particular direction is special (App. D). Carrier load rises across training as directions calibrate around fixed functions (Fig. 3b, SAE-roundtrip harness). If the scaffold bears load, the relation that defines it should carry something between features, which is the last thing to test.

3.4 Ablating a containment parent moves its child, in the SAE readout

The containment edges of §2.2 span a large part of the network. At convergence the carrier sub-DAG covers 29–64% of active features, with 1,484 of Pythia-160M's 2,408 organised beneath its 51 carriers (Fig. 4a, Table 3). Most of that reach comes from containment rather than persistence. Equally broad, high-CI non-carriers recover much of the same coverage, and the persistence increment on top is modest and varies with scale (App. F.4).

Containment is an observational definition, so the relation could reflect co-occurrence alone. We test it by ablating the parent in the SAE-roundtrip readout of §2.3, at the two models with a balanced control pool. That readout measures coupling inside the encoder, a different quantity from the loss-based load of §3.3 (App. E). At Pythia-160M, removing the parent u cuts the child's median firing by 11% against 0% for a breadth- and CI-matched control (Fig. 4b). Pythia-410M reproduces this at 11% against 0%.

Four controls place the effect in the containment relation rather than in the pair. Sources matched on co-occurrence with the same child move it by 1%. A second graph built from cross-layer decoder-cosine alignment couples at 8.9% over an almost disjoint edge set. An independently initialised SAE reproduces the effect at 10%, and held-out token chunks come within half a point of the in-sample estimate. Splitting and absorption act inside one dictionary, and every edge here joins two (App. F.2). The coupling is a property of the model, and it reaches 61–71% of its converged strength by step 1k (App. E).

4 Discussion

A feature's development has two parts, and the model treats them differently. Its token set is input-constrained, since a random network already detects the seed and sibling runs form 85–88% of the same features by step 1k (App. B). Its context selectivity and direction are built, both starting near chance and still moving at 84% of training. Where the token set needs no agreement, context selectivity and direction are the feature's contract with its readers and settle as the readers do. The coarse-before-fine reading holds, as does the two-phase picture of Ge et al. (2026), with the token clock running in the statistical-learning phase and the wiring clock in the feature-learning one. What is new is where the line falls, through the middle of what function usually means, feature by feature and in the same place at four models.

Bottom-up convergence, a fact about layers since Raghu et al. (2017), belongs to the token clock alone (§3.1).

Every method that follows a feature inherits the same division. Activation-profile matching, ours included, mostly matches token sets, and warm-started dictionaries and crosscoders (Xu et al., 2024; Ge et al., 2026) have yet to be read with the split, so a developmental claim should say which part of a feature it is about.

Inside the window the two clocks open, the model commits to part of its representational organisation, and later training builds computation around what survives that commitment. Among persisting features, position in the containment hierarchy is legible at step 1k and kept while context selectivity and direction are still turning, and membership is settled later (§3.2). That population is also the expensive one to remove, at 1.7–$1.9\times$ a matched control (§3.3). Critical-period work makes the commitment claim for whole networks (Achille et al., 2019; Kleinman et al., 2024), and early-phase results date it without naming what is committed (Frankle et al., 2020). Carriers are a candidate unit, and declining plasticity (Dohare et al., 2024) is a reason to expect later computation to organise around them rather than replace them. Both readings go beyond what we measure, and a prospective selection keeps too few features to close the loop (§3.3). The direct test is to perturb predicted carriers early, resume pretraining, and ask whether the mature scaffold is restored.

One finding would overturn this reading, and it has not appeared. If the wiring clock were dictionary noise, two dictionaries on one run would disagree about where it stands. They agree at 0.56 and 0.47 (App. B).

5 Limitations

Evidence covers four layers per model at one SAE width and mainly one dictionary seed, and every claim is population-level (§2.2). Part of the step-1k gap is how identifiably an SAE recovers each clock (App. B). Pythia-1B is underpowered (12 step-1k carriers), OLMo-2-1B's early direction cosine is partly a basis effect (App. B), and the energy-only control closes the load gap at Pythia-410M ($0.92\times$). The roundtrip load trend (Fig. 3b) awaits replication across seeds (App. D). Containment is read through a top-k decomposition that distorts hierarchies, and App. F.2 bounds that channel without testing hierarchy-aware alternatives (Bussmann et al., 2025; Cao et al., 2026).

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and to support qualitative data analysis. GPT-5.6 Sol produced the candidate feature labels for the carrier auto-interpretation step, following the Bills et al. (2023) recipe, with machine-assigned confidence tiers, and Claude Opus 4.8 served as the blind second rater for those labels (App. F.6). Claude Opus 4.8 and Claude Opus 5 assisted in implementing the analysis pipeline and were additionally used to edit code, create figures, and draft and edit the manuscript, directed and reviewed by the authors at every step. Hypotheses, experimental design, theoretical development, and interpretation of results are the authors' own, and the remaining required-disclosure tasks, including synthetic data generation, dataset cleaning, proofs, and translation, do not apply to this work. Every quantitative claim, analysis, and experimental artefact was produced and checked against its source data by the authors, who are responsible for the contents of this submission.

ETHICS STATEMENT

This work is a mechanistic interpretability analysis of public, already-released language models (Pythia, OLMo-2; Apache 2.0) using held-out public validation text (the Pile; CC-BY). No human subjects, no new models or datasets, and no sensitive data are involved. The main societal implication concerns model auditing and editing, where methods that remove or modify features based on converged geometry underestimate the cost of edits to early-formed persistent structure, a gap this paper's causal and load-bearing results quantify directly. We see no direct path to harmful applications beyond those general to interpretability research.

REPRODUCIBILITY STATEMENT

All code and per-snapshot JSON outputs for all three Pythia scales and OLMo-2-1B are included with the submission. The Pythia (Biderman et al., 2023) and OLMo-2 (Team OLMo, 2025) checkpoints and the Pile validation set (Gao et al., 2020) are public. §2 specifies models, checkpoints, layers, SAE hyperparameters, aligned-sample construction, matching criteria, and the ablation protocol; the appendix, beginning at App. A, gives full methodological detail, seeds, and robustness controls. Each SAE trains for 2,000–8,000 steps. All analyses run at laptop scale. Every number printed in this paper is bound to a field of a shipped data file by a machine-checked ledger, and the checker runs on the supplement alone, with no GPU and no external drive.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=BkeStsCcKQ. arXiv:1711.08856.

Allen Institute for AI. OLMo-2-0425-1B model card. Hugging Face model repository, https://huggingface.co/allenai/OLMo-2-0425-1B, 2025. The 1B member of the OLMo 2 family (16 layers), released after the OLMo 2 report, with intermediate pretraining checkpoints published as revisions of the form stage1-step<N>-tokens<T>B.

Tatsuya Aoyama, Ethan Gotlieb Wilcox, and Nathan Schneider. Predicting the emergence of induction heads in language model pretraining. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2511.16893.

Deniz Bayazit, Aaron Mueller, and Antoine Bosselut. Crosscoding through time: Tracking emergence & consolidation of linguistic representations throughout LLM pretraining. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1353–1377, 2026. doi: 10.18653/v1/2026.acl-long.60. URL https://aclanthology.org/2026.acl-long.60/.

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)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

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.

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, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 6077–6101, 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

Tue M. Cao, Hoang X. Nhat, Raed Alharbi, Phi Le Nguyen, and My T. Thai. Tree SAE: Learning hierarchical feature structures in sparse autoencoders. *arXiv preprint arXiv:2605.07922*, 2026. URL https://arxiv.org/abs/2605.07922.

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. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://arxiv.org/abs/2409.14507.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=MO5PiKHELW.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.03093.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10365.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

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. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.04093.

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.

Nils Grandien, David Steinmann, Felix Friedrich, and Kristian Kersting. Do sparse autoencoders learn meaningful concept hierarchies? *arXiv preprint arXiv:2606.22994*, 2026. URL https://arxiv.org/abs/2606.22994.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Sparse autoencoders can interpret randomly initialized transformers. *arXiv preprint arXiv:2501.17727*, 2025. URL https://arxiv.org/abs/2501.17727.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2402.02364. Earlier arXiv title: The Developmental Landscape of In-Context Learning.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Tatsuro Inaba, Go Kamoda, Kentaro Inui, Masaru Isonuma, Yusuke Miyao, Yohei Oseki, Yu Takagi, and Benjamin Heinzerling. How a bilingual LM becomes bilingual: Tracing internal representations with sparse autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13458–13470, 2025. doi: 10.18653/v1/2025.findings-emnlp.725. URL https://aclanthology.org/2025.findings-emnlp.725/.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Aq35gl2c1k.

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.

Patrick Leask, Bart Bussmann, Michael T. Pearce, Joseph I. Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. Transformer Circuits Thread, 2022. URL https://arxiv.org/abs/2209.11895.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1fe56b043848b211030c202764c6a7-Abstract-Conference.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.05806.

Shashata Sawmya, Micah Adler, and Nir Shavit. The birth of knowledge: Emergent features across time, space, and scale in large language models. *arXiv preprint arXiv:2505.19440*, 2025. URL https://arxiv.org/abs/2505.19440.

Lee Sharkey, Dan Braun, and Beren Millidge. Interim research report: Taking features out of superposition with sparse autoencoders. AI Alignment Forum, 2022. URL https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition.

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. URL https://www.researchsquare.com/article/rs-9269961/v1. Posted 16 April 2026.

Team OLMo. 2 OLMo 2 Furious. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2501.00656. Full report on arXiv; a shorter version appeared at COLM 2025.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 40699–40731, 2024. doi: 10.52202/079017-1287. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/47c7edadfee365b394b2a3bd416048da-Abstract-Conference.html.

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.

A Experimental Setup and Definitions

This appendix specifies the models, SAE training, and the core definitions used throughout: the aliveness and matching thresholds, the persistent-carrier criterion, the containment DAG, and the reconstruction-fidelity check that licenses interpreting the dictionaries.

Table 4: Experimental setup. Seven training checkpoints per Pythia scale (six post-initialisation + step 0); four residual-stream layers per model at uniform depth spacing. OLMo-2-1B's six post-initialisation checkpoints are fraction-matched to the Pythia schedule and are labelled by their Pythia counterparts (1k, 10k, 40k, 80k, 120k, 143k) throughout the paper.

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
Pythia-1B	0, 1k, 10k, 40k, 80k, 120k, 143k	L3, L7, L11, L15
OLMo-2-1B	13k, 130k, 530k, 1,070k, 1,600k, 1,907k	L3, L7, L11, L15 (of 16)

SAE training details. Each SAE is trained with Adam ($\eta = 3\times10^{-4}$, batch size 64) for 2,000/4,000/8,000 steps at 160M/410M/1B, holding the token-per-feature budget approximately constant across scales; decoder columns are re-normalised to unit norm at each step. A feature is alive at a checkpoint if its activation exceeds 10^{-6} on at least 0.5% of S^* samples. All SAE training and analyses run at laptop scale on a single consumer machine; no accelerator cluster is required.

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 (Table 4), under the alive and match thresholds of §2.2, and its endpoint must sit in the high-containment tier (CI $\geq$ 10) at step 143k. The counts this yields are in Table 3 and the instrumented layers in Table 4.

DAG construction. The four instrumented layers' encodings are concatenated into one active set per snapshot, and every ordered pair of active features (u, v) in it is a candidate edge, whether the two sit in one layer or in different ones. For each pair we compute the subset fraction $s(u, v) = |T_u \cap T_v|/|T_v|$ and the co-firing lift $l(u, v) = s(u, v)/b_v = P(v \mid u)/P(v)$ on a fixed validation slice, where T_u, T_v are the firing-token sets and b_v is v's firing breadth (formula audited against the submitted implementation). 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. Nothing in the rule refers to depth, so the ordering is by firing breadth alone and the edges run inside a layer as well as between layers in both directions. App. F.2 gives the within-layer share and App. E the direction split, which decides what the ablation of §2.3 can test. Reported intervals throughout condition on the selected feature sets, the fixed evaluation chunks, and one SAE seed per (layer, checkpoint); they do not propagate uncertainty from dictionary initialisation or cross-checkpoint matching (the test-retest control below quantifies the former directly).

Life-event definitions. Births and deaths follow the developmental taxonomy of Stecher et al. (2026): a feature unmatched at T and active at $T+1$ is born, one active at T and unmatched at $T+1$ has died, with CI replacing the label-derived selectivity indices of the original (root CI $\geq$ 10, mid $\in [1, 10)$, leaf $= 0$). The per-transition event null is estimated by permuting activation columns over position in the aligned sample S^* (1,000 permutations).

Decoder-insertion fidelity. Single-layer decoder insertion preserves most of the model's performance at every instrumented layer of every model except the failed 160M-L11 substrate. For each (model, layer) at step 143k we replace the layer's residual-stream output with its SAE reconstruction and measure next-token CE on held-out tokens (127,500 tokens at the Pythia scales; 32,640 at OLMo-2-1B, CPU harness), against two references: the clean model and the layer replaced by

its mean activation vector over S^*. Loss recovered is $(\mathrm{CE}_{\mathrm{mean}} - \mathrm{CE}_{\mathrm{SAE}})/(\mathrm{CE}_{\mathrm{mean}} - \mathrm{CE}_{\mathrm{clean}})$: 1 is perfect reconstruction, 0 is no better than mean-ablating the layer. Table 5 reports all 16 cells. Values are 0.75–0.92 across models with one exception: 160M-L11 at 0.37, the substrate failure of App. F.6, which contributes no carriers and is excluded from all carrier claims. Insertion fidelity characterises the SAE as a measurement instrument; the load results themselves do not depend on it, because the primary readout projects decoder directions out of the unmodified model (§2).

Table 5: Single-layer decoder-insertion fidelity at step 143k. Cosine is the mean reconstruction cosine on S^*; ΔCE is the insertion cost against the clean model (nats/token); loss recovered is referenced to mean-ablation of the layer (see text). The 160M-L11 row is the substrate failure of App. F.6.

Model	Layer	Cosine	ΔCE	Loss recovered
Pythia-160M	L2	0.86	+0.97	0.92
	L5	0.86	+1.53	0.75
	L8	0.87	+1.49	0.75
	L11	0.97	+3.97	0.37
Pythia-410M	L4	0.91	+0.78	0.89
	L11	0.91	+1.08	0.85
	L17	0.90	+1.44	0.80
	L23	0.90	+1.73	0.82
Pythia-1B	L3	0.89	+1.40	0.80
	L7	0.89	+1.29	0.82
	L11	0.88	+1.29	0.82
	L15	0.87	+2.02	0.78
OLMo-2-1B	L3	0.89	+1.16	0.83
	L7	0.87	+1.19	0.81
	L11	0.89	+0.90	0.85
	L15	0.87	+1.77	0.87

B The Two Clocks: Convergence Timescales and What They Predict

This appendix reads three quantities for each feature, its token set, its context selectivity and its direction. The token clock is the first of them. The wiring clock is the second and third read together, and the paragraphs below establish that the two move on one schedule.

This appendix supports §3.1 and §3.2. It opens with the convergence timescales behind the two clocks and three controls on how they are read, a basis-free reading of direction (orthogonal Procrustes alignment), each clock against the ceiling an SAE's own identifiability sets, and each clock against its matched null. It then defines the token-set and context split and reads both parts over four populations, all persistent lineages, the seeded subset traceable to initialisation, the context-formed remainder, and the instrumented layers taken separately. It closes with what the early window predicts, which raw statistic alone ranks scaffold membership, the trackability-vs-fate control (matchability null), and the direction-matched-lineage control that rules out a chain-construction artefact.

Function and direction convergence. Function and direction are read from the same carrier chains at the same checkpoints. Function is the Pearson r of a carrier's per-token activation profile against its step-143k profile; direction is the cosine of its decoder column against the step-143k column. Averaged over each model's carriers, function stands at 0.53–0.58 at step 1k against 0.03–0.21 for direction, and at step 120k, the last checkpoint before the reference, at 0.78–0.82 against 0.42–0.72. Function therefore begins each trajectory already past its midpoint and direction begins it near zero, at every model and in the same direction.

Basis-rotation control (orthogonal Procrustes). Cross-checkpoint cosines are read in the residual basis of each checkpoint, so a global rotation of that basis between step 1k and step 143k would depress every feature's cosine at once. For each (model, layer, checkpoint t) we fit the orthogonal R minimising $\|X_t R - X_{143\text{k}}\|_F$ over the mean-centred activations of S^* (the same 20,000 positions at both checkpoints), map each carrier's step-t decoder column through R, and recompute its cosine to the step-143k column. Alignment lifts the pooled step-1k carrier cosine from 0.03–0.21 to 0.23–0.37, and at step 120k from 0.42–0.72 to 0.59–0.73. The fitted map is far from the identity at step 1k ($\mathrm{tr}(R)/d$ = 0.24–0.46) and reaches it by step 120k (1.00 at the Pythia scales, 0.97 at OLMo-2-1B), so the early residual basis does rotate, most at OLMo-2-1B, where the near-zero raw cosine in Fig. 1b is partly a coordinate effect. What the rotation cannot do is close the gap: the aligned step-1k direction remains below the step-1k function (0.53–0.58) at every model, and at the Pythia scales the gain from alignment is under 0.01 by step 120k.

Each clock against its own identifiability ceiling. The three readings are not equally identifiable by an SAE, so a gap between them could be a gap in measurement rather than in learning. Two dictionaries initialised independently and trained on the same Pythia-160M activations agree, at one checkpoint, at $r = 0.85$ on the token set, 0.40 on context selectivity and cosine 0.42 on direction at step 1k, and 0.83, 0.56 and 0.47 at step 143k. Read against those ceilings, the step-1k lineage readings of 0.74, 0.22 and 0.27 stand at 87%, 55% and 65% of what the instrument allows, so the token clock leads by a factor of 1.6 rather than the 3.3 the raw readings give. The ordering survives the correction and the margin does not close, but the raw gap overstates it. At step 120k the lineage readings exceed the same-checkpoint ceilings, which the two measurements running over different populations is enough to produce, so no normalised figure is read there, and the claim that context selectivity and direction are still moving at 84% of training rests on their trajectory rather than on a distance from a ceiling.

Cross-run universality of the early clock. Sibling runs of Pythia-160M (`pythia-160m-weight-seed{1,2}`, initialisation varied with the data order fixed; `pythia-160m-data-seed1`, data order varied with the initialisation fixed) are captured at steps 1k and 143k on the S^* positions, their SAEs trained with the recipe of App. A, and features matched across runs by the lineage rule of §2.2 (Hungarian one-to-one on activation-profile r, accepted at $r \geq 0.5$). Of the 437 canonical-run lineages alive at step 1k, 85–88% have a step-143k partner above the gate in each of the three sibling runs, so the functions the run has committed to by step 1k are functions every sibling run also forms. The decoder directions implementing matched pairs agree far less. After orthogonal Procrustes alignment of the two runs' step-143k bases, converged matched pairs have mean cosine 0.39–0.49 at L2/L5/L8 across the three runs, and step-1k-to-converged pairs 0.27–0.42 (L11 is the collapsed substrate of App. F.6). Split by token, features matched across runs agree far better on their token-set part than on their context part or their direction, at both steps. Read against the ceiling dictionary noise alone allows, which is the same-run agreement of two independently initialised dictionaries above, the cross-run deficit is at most 0.05 on every reading: the token map is the one property identifiable across dictionaries, and cross-run matching cannot separate run-specificity of context selectivity or direction from dictionary non-identifiability.

Both clocks against their own nulls. Function is read as a Pearson r over sparse activation profiles and direction as a cosine in d dimensions, and the two statistics have different chance levels: a top-k profile correlation can be inflated by the zeros the two profiles share, and a cosine between random directions has standard deviation near $1/\sqrt{d}$. To place the readings on one scale, every lineage at every checkpoint is paired with randomly drawn other converged features of its layer (the pool is every persistent lineage endpoint in that layer) and the reading is recomputed against each: the step-t profile against the other feature's step-143k profile, which holds both marginals and the top-k sparsity pattern fixed, and the step-t decoder column against the other feature's step-143k column. A third null pairs the column with a random unit vector and is also the null of the Procrustes-aligned reading, since an orthogonal map leaves the angle to an isotropic direction distributed as before. Table 6 reports each Table 2 reading as its distance from its own null in null standard deviations and as the fraction of lineages above the null's 95th percentile. Every null mean lies within 0.01 of zero. At step 1k function stands far above its own null at every model and direction stands much closer to its own, and the ordering survives on the direction-matched lineages, which are built to favour

direction. The one reading that fails to clear its null is the raw carrier direction at OLMo-2-1B, at 1.0 standard deviations, the basis-sensitive case the Procrustes control addresses. By step 120k the two readings are comparably far from chance, so the gap between them belongs to early training rather than to the two statistics. A scale-free reading agrees: over all persistent lineages, function at step 1k has reached 71–75% of its step-120k value and direction 17–42%.

Table 6: Step-1k readings of Table 2 against their own matched nulls (20 draws per lineage, App. B). For each model, sd is the standard deviation of the null draws, z the reading's distance above the null mean in those units, and % the percentage of lineages whose own reading exceeds the null's 95th percentile. Direction is scored against the other-column null.

	Pythia-160M			Pythia-410M			Pythia-1B			OLMo-2-1B		
Reading at step 1k	sd	z	%	sd	z	%	sd	z	%	sd	z	%
Firing-matched: function r	0.086	6.6	100	0.056	10.0	97	0.070	7.5	100	0.077	7.4	100
Firing-matched: direction cos	0.042	3.6	70	0.038	4.3	72	0.035	6.1	83	0.027	1.0	26
Direction-matched: function r	0.045	11.6	96	0.030	19.1	100	0.028	24.2	99	0.026	26.2	100
Direction-matched: direction cos	0.046	8.3	96	0.040	10.7	98	0.031	15.3	100	0.022	9.0	97
All persistent: function r	0.063	9.3	99	0.055	10.7	99	0.033	17.3	99	0.039	14.0	99
All persistent: direction cos	0.042	4.7	84	0.039	6.2	90	0.031	9.7	93	0.024	4.4	71

The token-set and context split. Every activation profile over S^* is split exactly into a token-set part, the feature's mean activation over the positions carrying each token, and a context remainder, the deviation from that mean at each position. The remainder sums to zero within every token, so it is orthogonal to any profile that depends on the token alone, and the covariance between a step-t profile and its step-143k profile is the sum of a token-set and a context covariance with no cross term. Positions whose token occurs fewer than five times in S^* are left out, keeping 57–58% of them. Two readings follow for every persistent lineage alive at step 1k. The token clock reads the correlation of the token-set parts, and context selectivity the correlation of the remainders. Table 7 reads both, together with the basis-aligned decoder direction, over every population this appendix uses, and the paragraphs below take each population in turn.

At step 1k the token set leads context selectivity in every population at every model, by 0.30–0.52 over all persistent lineages, and both stand far above the chance a random other converged feature of the same layer gives (95th percentile of the pooled null draws 0.12–0.14 for the token set and 0.02–0.03 for context). Read per lineage rather than as a mean, the token clock leads in 86–95% of them. Dividing the raw correlation into its two contributions locates the movement. From step 1k to step 120k the token-set contribution gains at most 0.06 while the context contribution gains 0.17–0.20, so everything the raw clock gains after the first checkpoint is contextual. The token set carries 34–59% of a converged feature's variance.

Three further readings bear on how the split should be read. Estimating the per-token means on one half of each token's occurrences and taking the remainder on the other half leaves step-1k context selectivity within 0.01 of Table 7, so in-sample means do not produce it. Context selectivity and the aligned direction move together rather than merely ranking alike, which is what licenses reading them as one wiring clock. The largest gap between them at any chain checkpoint stays at or below 0.10, and their across-lineage correlation runs 0.34–0.71 at step 1k and 0.43–0.68 at step 120k. Alignment matters most at OLMo-2-1B, whose raw step-1k cosine of 0.10 becomes 0.40 once the basis is mapped. Two earlier observations are special cases of context selectivity, the context-sensitive features of Ge et al. (2026), which form last in their emergence hierarchy, and the polysemous token features of Xu et al. (2024), which keep their token and narrow to one sense over training.

All persistent lineages. Fig. 1a–b and Table 7 read the clocks over every persistent lineage rather than over the carrier chains alone. To build that population, every feature alive at step 143k is traced backward with the lineage rule of §2.2 itself (sparse top-32 SAE encoding of S^*, Hungarian one-to-one assignment on Pearson r between consecutive checkpoints, match accepted at $r \geq 0.5$, columns active on at least 0.5% of positions, persistent if present in at least 4 of the 6 post-initialisation snapshots), and the clocks are read along each one exactly as for the carriers. Of 837 to 1,409

Table 7: The clocks by population. Mean readings over the persistent lineages alive at step 1k, for the four populations this appendix uses: all of them, the canonical carriers among them, and the split by whether the step-1k feature has a step-0 partner under the lineage rule (seeded) or forms its token set during the first 1k steps (context-formed). Token and Context are the token set and context selectivity, Dir. the decoder cosine after the orthogonal map of the step-t basis onto the step-143k basis. All three are read against the step-143k endpoint of the same lineage.

			Step 1k			Step 120k		
Model	Population	n	Token	Context	Dir.	Token	Context	Dir.
Pythia-160M	All persistent	437	0.74	0.22	0.27	0.90	0.64	0.62
	Canonical carriers	37	0.71	0.22	0.23	0.91	0.69	0.66
	Seeded	182	0.84	0.21	0.24	0.94	0.66	0.61
	Context-formed	255	0.67	0.23	0.30	0.87	0.63	0.62
Pythia-410M	All persistent	489	0.78	0.27	0.33	0.91	0.69	0.66
	Canonical carriers	29	0.77	0.32	0.29	0.89	0.68	0.63
	Seeded	201	0.87	0.22	0.32	0.94	0.65	0.64
	Context-formed	288	0.71	0.31	0.34	0.89	0.72	0.68
Pythia-1B	All persistent	528	0.71	0.32	0.42	0.88	0.75	0.74
	Canonical carriers	12	0.76	0.32	0.37	0.95	0.78	0.74
	Seeded	113	0.89	0.26	0.49	0.96	0.76	0.76
	Context-formed	415	0.66	0.34	0.40	0.86	0.75	0.73
OLMo-2-1B	All persistent	586	0.71	0.41	0.40	0.86	0.68	0.65
	Canonical carriers	23	0.87	0.37	0.36	0.93	0.66	0.59
	Seeded	13	0.93	0.34	0.41	0.95	0.66	0.62
	Context-formed	573	0.70	0.41	0.40	0.86	0.68	0.65

persistent lineages per model, 437 to 586 reach step 1k. The rebuilt lineages recover every canonical carrier with identical chain histories, and restricted to them reproduce the means of Fig. 1 exactly.

Read whole rather than split, function stands at 0.54–0.59 at step 1k against direction at 0.10–0.30, still leads by step 120k, and exceeds direction in 97.5–99.8% of individual lineages. Being a carrier adds nothing to the early function reading. Pairing each carrier with the persistent non-carrier nearest in step-143k firing breadth (greedy, as for the load control) gives a step-1k function reading within noise of the carriers' own at every model (Mann-Whitney $p \geq 0.20$), and 0.56 against 0.53 pooled ($p = 0.44$). On direction the controls sit at or above the carriers. The early window is therefore a property of persistent features in general, and the carrier subset is where it becomes predictive of mature position rather than where it is strongest.

The seeded subset at step 0. The chain rule starts at step 1k, so Fig. 1 opens with function already past its midpoint. To read the clocks at initialisation, the same one-to-one assignment (Hungarian on Pearson r over S^*, columns active on at least 0.5% of positions, accepted at $r \geq 0.5$) is applied once more between the step-0 and step-1k dictionaries, and every persistent lineage whose step-1k feature receives a step-0 partner is read at step 0 against its step-143k endpoint exactly as at the later steps. Lineages without a partner are absent at step 0 rather than scored zero, so the step-0 point describes the seeded rows of Table 7 (13 to 201 lineages per model, of which 2 to 20 are carriers).

On that subset the two halves of function part company before training begins. The token clock already reads 0.77–0.90 at step 0, while context selectivity reads 0.02–0.14 and the direction 0.01–0.06, against chance on every reading for a random active step-0 feature scored against the same endpoint. Read whole, function is at 0.55–0.58 and exceeds direction in 97.8–100% of the matched lineages. For a lineage traceable to the random network, then, the token set it will answer to is already visible in the initialisation, and everything else about it starts at chance.

The assignment never sees a decoder column, so the direction reading involves no selection. What the firing match selects is the subject of the next paragraph. The precursor rule of App. C (top-k values clamped at zero, columns active on more than 0.1% of positions) recovers the committed precursor carriers (2 to 20 per model) and moves every step-0 mean by at most 0.02. As in App. C, a step-0 correlation may reflect token and positional structure visible to a random network, and OLMo-2-1B's step 0 is its own initialisation checkpoint.

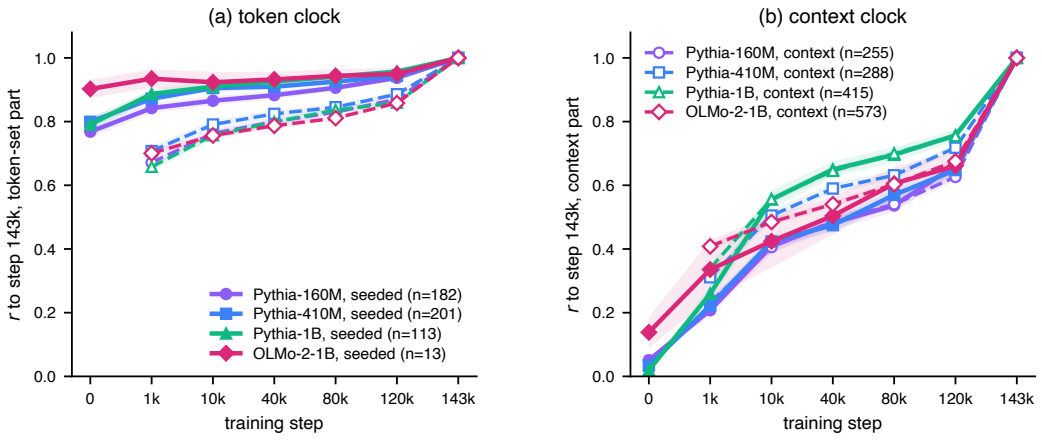


Figure 5: Two routes to an early token set, one population per curve. **(a)** The token clock and **(b)** context selectivity for the persistent lineages whose step-1k feature has a step-0 partner under the lineage rule (seeded; solid, from step 0) and for those without one (context; dashed, from step 1k), all four models, means with ± 1 SEM. Seeded lineages carry their token set from initialisation; context lineages form theirs by step 1k; both read context selectivity alike.

Seeded against context-formed lineages. The step-0 partners are token detectors (Fig. 5). Token-selective features at random initialisation have been reported before (Heap et al., 2025; Ge et al., 2026). Ge et al. (2026) see their initialisation features recover from an early dip in crosscoder decoder norm and then decay over training, while the seeded lineages here persist to convergence. The two readings hold different things fixed. A crosscoder feature is one code shared across snapshots, and its per-snapshot norm falls once a later snapshot describes the same positions with a different unit, whereas a lineage is re-matched by profile at every checkpoint with an independent SAE and so follows the seed through that re-description.

Reading the token at every S^* position, the most frequent token among a step-0 feature's firing positions accounts for a median 62–87% of them at the Pythia scales (32% at OLMo-2-1B, 13 lineages), and for 98–99% of them for the seeded carriers at the two smaller scales. The converged partner keeps that seed and grows around it. It still fires on 73–94% of the seed token's occurrences, while the seed has fallen to 38–65% of its own firings. Training widens a token detector into a category containing the token rather than re-assigning it.

The persistent lineages with no step-0 partner are a different kind of feature, their top token accounting for 9–12% of their firings and their token set forming from context between steps 0 and 1k. Table 7 separates the two routes. The seeded group enters step 1k with the higher token clock at every model (0.84–0.93 against 0.66–0.71), the two groups read context selectivity within 0.09 of each other, and direction lags both. The ordering belongs to the split rather than to the token set, then, holding in the group whose token set is inherited from the embedding and in the group whose token set is learned. At OLMo-2-1B, whose fraction-matched first checkpoint lies far more tokens from initialisation than Pythia's step 1k, almost every lineage takes the second route.

Depth. Table 8 reads the same clocks per instrumented layer. Read whole, function at step 1k is highest in the shallowest layer at every model, the bottom-up ordering of Raghu et al. (2017). Context selectivity at step 1k is lowest in the shallowest layer at Pythia-160M, -410M and OLMo-2-1B and in the two shallow layers at Pythia-1B, where the shallowest trails the deepest by 0.07 (Mann-Whitney $p = 0.04$, bootstrap CI $[-0.14, 0.00]$) against 0.17–0.22 elsewhere ($p \leq 2.3 \times 10^{-9}$ at each). At step 120k the shallowest layer still trails on context selectivity by 0.06–0.12 at Pythia-160M, -410M and OLMo-2-1B and leads by 0.02 at Pythia-1B. The token clock stays within 0.08 across the usable layers of a model at step 1k, and the token variance share and the seeded fraction both fall with depth at every model. Pythia-160M L11 (3 lineages) is listed and excluded from the comparisons.

Table 8: The clocks by depth: persistent features alive at step 1k, per instrumented layer. Share is the token variance share of the converged feature, seeded the percentage traced to a step-0 partner; the four step-1k columns are the raw function r, the token clock, context selectivity and the basis-aligned direction cosine to step 143k, and the last two are context selectivity and direction at step 120k.

					Step 1k				Step 120k	
Model	Layer	n	Share	Seeded (%)	Raw	Token	Context	Dir.	Context	Dir.
Pythia-160M	L2	193	0.72	52	0.66	0.76	0.15	0.23	0.59	0.60
	L5	140	0.50	36	0.57	0.72	0.26	0.30	0.69	0.64
	L8	101	0.48	30	0.58	0.75	0.31	0.33	0.68	0.63
	L11	3	0.21	0	0.32	0.47	0.27	0.23	0.90	0.19
Pythia-410M	L4	171	0.68	51	0.65	0.78	0.18	0.27	0.64	0.67
	L11	129	0.45	40	0.62	0.82	0.29	0.35	0.71	0.64
	L17	103	0.43	36	0.56	0.74	0.32	0.37	0.73	0.69
	L23	86	0.37	28	0.57	0.74	0.39	0.39	0.70	0.65
Pythia-1B	L3	137	0.61	25	0.65	0.73	0.29	0.47	0.75	0.76
	L7	147	0.48	17	0.53	0.68	0.27	0.37	0.76	0.74
	L11	150	0.41	22	0.57	0.71	0.37	0.44	0.78	0.75
	L15	94	0.36	22	0.53	0.72	0.36	0.39	0.72	0.69
OLMo-2-1B	L3	81	0.50	7	0.58	0.73	0.32	0.37	0.58	0.58
	L7	130	0.38	2	0.51	0.72	0.35	0.33	0.68	0.65
	L11	182	0.39	1	0.52	0.70	0.37	0.38	0.68	0.67
	L15	193	0.21	2	0.57	0.69	0.51	0.48	0.70	0.66

Univariate cross-check. Ranking features by a single unfitted step-1k statistic already separates future carriers. The Containment Index and firing breadth both rank them well above chance (Table 9), while the three activation statistics alone rank at or near chance (0.30–0.63, and all three below chance only at Pythia-160M). Their large positive coefficients in the fitted model are therefore conditional, a suppressor structure: activation magnitude distinguishes future carriers among features of similar breadth and containment rather than marginally. Leave-one-layer-out AUC is $0.74/0.68/0.73$ holding out L2/L5/L8 at Pythia-160M (L11 contributes no carriers, so no held-out AUC is defined there) and 0.65–0.90 across the Pythia-410M layers. At Pythia-1B the held-out positive counts (1–5 per layer) are too small for stable estimates.

Recomputing the predictor from the token-set profiles alone splits the two univariate readings cleanly. Breadth reads as high from the token set as it does on the raw profiles, while the Containment Index falls to 0.54–0.64 (Table 9). Only at OLMo-2-1B does the fitted token-set AUC exceed the raw one, on a population that lost 5 of its 23 carriers. The breadth signal lives in the token set, and the containment signal needs the within-token selectivity that nests one feature's firing set inside another's.

Label decomposition. A step-1k feature is a future carrier when its lineage persists to step 143k and its converged member is a root (CI ≥ 10). Scoring the two parts separately, with nothing fitted, splits the label (Table 9). Survival is near chance from either statistic. Among the persistent features the same index ranks root status at convergence far above chance, and its Spearman correlation with the converged index is 0.27–0.48. Restricting the carrier ranking to (carrier, non-carrier) pairs from the same breadth decile leaves the index above chance, while breadth within index deciles stays near it. At the very top, the k highest-index features, k the carrier count, contain 1 to 4 carriers (4.7–$38.7\times$ the base rate), and 5 to 11 of the carriers are already roots at step 1k against a population root rate of 1.5–8.7%.

Matchability null (trackability vs. fate). The persistence label is awarded by an $r \geq 0.5$ matching gate, so a feature our instrument renders crisply re-matches more reliably and clears the persistence criterion more readily regardless of its developmental fate. Because the predictor's inputs also correlate with crispness, part of the AUC could reflect trackability rather than fate. We bound this at all four models. For each (model, layer) we train a second SAE on the *same* step-1k activations from an independent initialisation, trained at each model's canonical budget, and match it against the canonical dictionary under the same Hungarian assignment and $r \geq 0.5$ gate that builds the

Table 9: The step-1k predictor, reading by reading. Rankings are AUC over the step-1k features of the classifier's population, with nothing fitted except the two *Fitted AUC* rows, which are the five-fold cross-validated logistic fit of §2.3. A feature is a future carrier when its lineage persists to step 143k and its converged member is a root. Decile rows score only (carrier, non-carrier) pairs drawn from the same decile of the other statistic. For the token-set rows each profile is reduced to its per-token means, so a feature fires on a token when it fires on at least half of that token's occurrences at its mean activation there. A feature left with an empty profile drops out, and that costs some carriers as well. k is each model's carrier count.

Reading	Pythia-160M	Pythia-410M	Pythia-1B	OLMo-2-1B
Population at step 1k				
Active features	3,198	3,286	3,031	5,114
Persistent lineages	436	485	527	586
Future carriers	37	29	12	23
Unfitted ranking of the carrier label (AUC)				
Containment Index	0.75	0.86	0.77	0.92
Firing breadth	0.74	0.77	0.69	0.87
Containment Index, within breadth deciles	0.64	0.79	0.73	0.84
Firing breadth, within index deciles	0.54	0.45	0.51	0.61
The label scored in two parts (AUC)				
Persistence, from the index	0.54	0.55	0.53	0.58
Persistence, from breadth	0.46	0.46	0.47	0.61
Root given persistence, from the index	0.74	0.84	0.76	0.89
Root given persistence, from breadth	0.81	0.85	0.72	0.82
Spearman, index at step 1k against step 143k	0.40	0.44	0.27	0.48
Token-set profiles alone				
Features with a non-empty token set	2,797	2,757	2,550	3,204
Carriers retained	34	28	12	18
Fitted AUC, raw profiles	0.80	0.83	0.69	0.81
Fitted AUC, token set alone	0.77	0.74	0.60	0.88
Containment Index alone, token set	0.55	0.60	0.54	0.64
Firing breadth alone, token set	0.76	0.82	0.72	0.90
Precision at the top				
Carriers among the top k by index	2	2	1	4
Carriers already roots at step 1k	11	11	5	7
Population root rate at step 1k (%)	8.7	6.1	2.0	1.5

lineages. Each active step-1k feature thereby receives a re-match reliability r, measured entirely at step 1k and independent of training time.

Two tests follow. *Incremental AUC* compares a reliability-only model against reliability plus the six firing-space predictors. *Stratified AUC* scores only (carrier, non-carrier) pairs drawn from the same reliability decile, where trackability is near-constant and so cannot discriminate; the interval is a 1,000-fold bootstrap over features. Table 10 reports both. Carriers are more reliably re-matched than non-carriers at every model, confirming the covariate is real; but the prediction survives holding it fixed at Pythia-160M, -410M and OLMo-2-1B, with an incremental lift of $+0.086$ to $+0.103$. At Pythia-1B, with 12 positives, the two tests disagree (stratified 0.746 with a CI reaching 0.53; incremental $+0.002$) and the control is underpowered; we report it as inconclusive rather than supporting. Three further observations: the recomputed full AUCs reproduce Table 3 exactly, so the control runs on the published classifier; carriers that an independent dictionary fails to recover *at all* ($r = 0$) still rank at mean percentile 60 or above at every model where such a carrier occurs, on 1 to 5 of them; and the association between a carrier's rank and its reliability weakens and turns negative as the stratified estimate strengthens, so Pythia-160M is the worst case rather than the typical one. Because reliability partly reflects genuine feature strength and not only instrument quality, stratifying on it removes some real signal; the stratified figures are therefore a lower bound on the developmental component.

Table 10: Matchability null for the step-1k predictor. Re-match reliability is measured against an independently initialised SAE trained on the same step-1k activations. *Stratified* scores only (carrier, non-carrier) pairs within a reliability decile (95% bootstrap CI); *incremental* is the AUC lift of reliability+firing over reliability alone. Pythia-1B (12 positives) is underpowered and reported as inconclusive.

	Pythia-160M	Pythia-410M	Pythia-1B	OLMo-2-1B
Full AUC (= Table 3)	0.805	0.833	0.686	0.814
Reliability alone	0.727	0.766	0.670	0.741
Incremental over reliability	+0.086	+0.103	+0.002	+0.092
Stratified AUC	0.685	0.820	0.746	0.881
95% CI	[0.59, 0.78]	[0.73, 0.89]	[0.53, 0.87]	[0.76, 0.95]
Carrier re-match r	0.71	0.75	0.65	0.67
Non-carrier re-match r	0.41	0.40	0.39	0.33
Carriers alive at step 1k	37	29	12	23

Direction-matched lineages. The early stability of firing identity in §3.1 could in principle be a selection artefact: chains are built by firing-pattern matching, so the same metric may appear stable at step 1k by construction. Fig. 6 repeats the convergence analysis on lineages built from decoder directions. The rule is that of §2.2 with only the similarity exchanged: between consecutive checkpoints, columns active on at least 0.5% of S^* at both steps are assigned one-to-one by Hungarian assignment on decoder cosine, a match is accepted at cosine ≥ 0.5, and a lineage is persistent if present in at least 4 of the 6 post-initialisation snapshots. This yields 485 to 1,108 persistent lineages per model, of which 24 to 164 reach step 1k.

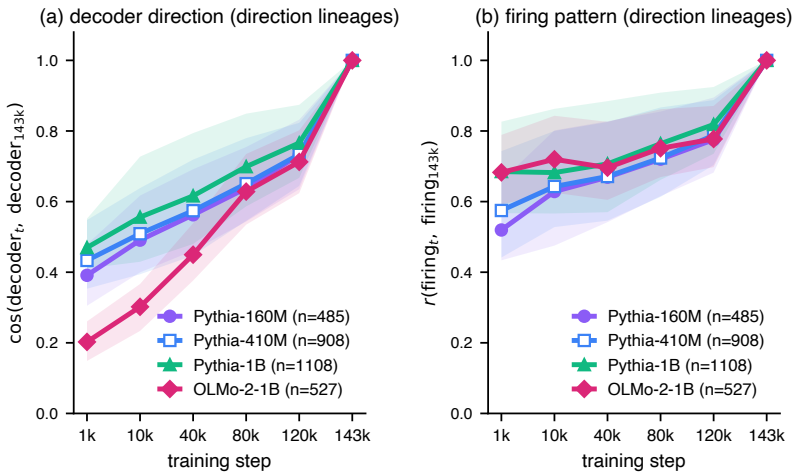


Figure 6: Convergence along direction-matched lineages (Hungarian assignment on decoder cosine, accepted at ≥ 0.5; persistent lineages, $n = 485/908/1{,}108/527$). **(a)** Mean decoder cosine to the step-143k column. **(b)** Mean firing-profile r to the step-143k profile on the same lineages. Bands: interquartile range. At step 1k function stands at 0.52–0.69 and direction at 0.20–0.47, so function leads on lineages built to favour direction.

Under this exchange of tracking criterion the ordering of Fig. 1a–b persists. Direction matching lifts step-1k direction to 0.39–0.47 at the Pythia scales, the counterpart of the firing lift on firing-matched chains, and function still leads in 79–100% of individual lineages (Wilcoxon signed-rank $p \leq 3 \times 10^{-5}$ at every model). The two similarities mostly select the same partner, 81–94% of direction-matched links coinciding with the firing-matched link. Few direction lineages reach back to step 1k (5–15% of persistent lineages against 37–52% under firing matching), because decoder columns rarely clear cosine 0.5 between steps 1k and 10k. By step 120k both have risen and function still leads (Fig. 6).

C Formation: Is the Opening-Phase Burst Real?

This appendix supports §3.2. It gives the noise-floor correction that establishes the 36–38× collapse against a dictionary-noise reading and the pre-existing structure the collapse builds on.

Feature-level event rates and the noise-floor correction. The released per-transition lineage counts inherited a 200-pair storage cap that miscounted every match beyond 200 per (layer, transition) as one birth plus one death; an uncapped recomputation gives Phase-I feature-level birth rates 21.7–29.4× the later-window mean at the three Pythia scales. Because a pure dictionary-noise null, in which two independently initialised dictionaries at the *same* checkpoint already fail to re-match at rate $1 - m$ ($m = 0.49$–0.66 at 160M; App. F), would contribute a roughly constant per-transition spurious-birth count regardless of window width, we subtract the estimated floor $(1 - m)\, N_{\text{active}}$ and recompute the Phase-I:later-mean ratio, bootstrapping over test-retest seed pairs. The corrected ratio is 36–38× across the three scales, at or above the uncorrected values, because removing the near-constant noise background sharpens the contrast rather than flattening it. The first post-warmup window (steps 1k–10k) still exceeds the later mean by 25× at 160M, so the concentration is not a warm-up transient.

Pre-existing structure at initialisation. At the two smaller scales, Phase I substantially amplifies pre-existing structure rather than building the scaffold from scratch: 39% of eventual Pythia-160M carriers and 31% of Pythia-410M carriers are already weakly active at step 0 (Fig. 2c–e, hatched bars). At Pythia-1B the fraction is 7% (2 of 27), consistent with the later genesis reported in §3.2. These step-0 correlations may partly reflect token/positional structure visible to a random network rather than latent learned representations; we treat the precursor fractions as descriptive.

Table 11: The control ladder at step 143k. Joint ΔCE in nats from projecting each population's decoder contributions out of the unmodified model across the four instrumented layers, with 95% bootstrap intervals over 1,000 resamples of 32 held-out chunks. Ratio is carrier ΔCE over control ΔCE, with whether the two intervals are disjoint. The last two columns give the achieved match, the carrier-to-control ratio of mean firing breadth and of mean total activation energy.

Model	Population	ΔCE [95% CI]	Ratio (disjoint)	Breadth	Energy
Pythia-160M	Carriers ($n = 51$)	+0.48 [+0.44, +0.52]	–	–	–
	Breadth-matched	+0.15 [+0.14, +0.17]	3.17× (yes)	1.01	1.31
	Energy-matched	+0.20 [+0.19, +0.22]	2.36× (yes)	0.89	0.99
	Breadth + energy matched	+0.27 [+0.25, +0.29]	1.77× (yes)	1.01	1.07
Pythia-410M	Carriers ($n = 49$)	+0.52 [+0.49, +0.54]	–	–	–
	Breadth-matched	+0.15 [+0.13, +0.16]	3.48× (yes)	1.00	1.44
	Energy-matched	+0.56 [+0.53, +0.59]	0.92× (no)	1.01	1.08
	Breadth + energy matched	+0.29 [+0.27, +0.31]	1.78× (yes)	0.99	1.18
Pythia-1B	Carriers ($n = 27$)	+0.33 [+0.30, +0.35]	–	–	–
	Breadth-matched	+0.14 [+0.13, +0.15]	2.33× (yes)	1.01	1.35
	Energy-matched	+0.23 [+0.21, +0.25]	1.40× (yes)	0.77	1.04
	Breadth + energy matched	+0.19 [+0.17, +0.21]	1.71× (yes)	0.99	1.05
OLMo-2-1B	Carriers ($n = 43$)	+0.36 [+0.33, +0.38]	–	–	–
	Breadth-matched	+0.17 [+0.15, +0.19]	2.13× (yes)	0.99	1.59
	Energy-matched	+0.16 [+0.14, +0.18]	2.24× (yes)	0.75	1.00
	Breadth + energy matched	+0.19 [+0.17, +0.21]	1.85× (yes)	0.99	1.01

D LOAD: DOES THE CARRIER POPULATION COST MORE TO REMOVE?

This appendix supports §2.3 and §3.3. It gives the two confound tests for the load-bearing claim (activation-energy covariate balance, independent-seed replication), the decoder-transplantation control, and why the projection measure rather than the SAE-roundtrip readout is primary.

Per-firing ΔCE does not isolate the carrier tier. Carriers sit in the high-containment root tier, and per-firing single-feature ablation impact does not separate that tier from either of the others: pooled across the instrumented layers, root-versus-leaf and root-versus-mid are both statistically indistinguishable (two-sample KS, $p = 0.51$ and 0.22 at Pythia-160M; $p = 0.57$ and 0.95 at Pythia-410M). The population-level load gap is therefore invisible to this readout and emerges only under joint cross-layer intervention.

Activation energy and the control ladder. At matched breadth, carriers fire harder than the non-persistent-root control (1.2–1.4× mean conditional activation) and more variably (1.4–2.4×) at the two audited Pythia scales, and at Pythia-410M the breadth-matched control carries about half the carriers' total activation energy (median carrier-to-control ratio 2.06×, against 1.02 at Pythia-160M). We therefore ran two further count-matched controls at all four models under the same projection harness: non-carriers matched on total activation energy alone, and non-carriers matched jointly on energy and firing breadth (Table 11). Energy alone does not settle the question. Against the energy-matched control the ratio reaches 2.36×, and at Pythia-410M it falls to 0.92× with overlapping intervals. At that scale the energy-matched and the jointly matched controls both lie within 2% of the carriers' mean breadth and within 8–18% of their energy, yet one costs nearly twice as much as the other to remove, so the choice of control set there moves the estimate by as much as the carrier gap itself. Holding breadth and energy fixed together restores the gap at every model, at 1.71–1.85× with disjoint intervals and both covariates matched (Table 11). This jointly matched control is the one the abstract and §3.3 quote, and its ratio sits near the 1.6× that independent re-initialisation gives below.

Matched leaves. The third control of §2.3, non-carriers at the leaf tier matched per layer on count, isolates hierarchical position rather than persistence. Under the SAE-roundtrip harness the matched leaves cost a fraction of what the carriers cost, about half at Pythia-410M and less at the other

two scales, where the leaf estimate at Pythia-160M straddles zero. The leaf tier is the cheapest population to remove at every scale tested. The roundtrip baseline is dictionary-dependent (below), and this control was not run at OLMo-2-1B.

Non-persistent roots. The first control of §2.3, roots (CI ≥ 10) that fail the persistence criterion, count-matched to the carriers per layer (23 to 49 features), costs less to remove than the carriers at all three audited Pythia scales under the same harness, by 1.7–$5.1\times$, with disjoint intervals at Pythia-410M and -1B and overlapping ones at Pythia-160M. This control holds hierarchical tier fixed and varies persistence; it is count-matched rather than breadth-matched, its baseline is the roundtrip harness, and it was not run at OLMo-2-1B.

Direction calibration carries real cost, but no specific direction is special. Replacing each carrier's step-143k decoder column with its step-1k value costs $+0.96$ (160M) and $+0.61$ (410M) nats. A random-direction control at the same angular distance and norm as the step-1k column (three seeds) costs a statistically indistinguishable $+0.86$ (mean, sd 0.08) at 160M and $+0.45$ at 410M (ΔR^2 within 0.013 of the real transplant at Pythia-410M, where the reconstruction-level audit was run): the transplant cost is dominated by displacement away from the calibrated direction, not by residual early signal. Median cosine between step-1k and step-143k columns is 0.15/0.14.

Prospective selection from step 1k alone. §3.2 ranks future carriers from step-1k containment and §3.3 shows carriers cost more to remove; this test attempts to close the loop without the persistence label. At step 1k, every alive feature is ranked by its Containment Index and the top k per layer are taken, k being that layer's canonical carrier count; each is followed forward along its lineage under the rule of §2.2 (App. B), and the survivors' decoder contributions are projected out of the unmodified step-143k model jointly across the four layers, on the same chunks and bootstrap as the main load measure. Few selected features survive: 10 of 51 at Pythia-160M (6 of them canonical carriers) and 7 of 49 at Pythia-410M (3 carriers). What the survivors cost to remove sits inside the spread of count-matched random step-1k survivors at Pythia-160M and above it at Pythia-410M, and above the breadth-matched survivors at both. Per feature the prospective set is the cheaper population at Pythia-160M, at 2.22 millinats against 9.37 for a carrier, and comparable at Pythia-410M. Selecting on step-1k breadth instead of containment leaves fewer survivors and costs no more. With seven to ten surviving features and random controls this variable, the test cannot separate step-1k containment from chance, and we treat it as inconclusive.

Why the projection measure is primary. An alternative harness substitutes the full SAE reconstruction at all four layers instead of projecting decoder directions out of the clean model; its baseline is dictionary-dependent (roundtrip baseline 5.8–10.7 nats depending on the dictionary, vs. the clean-model CE of 2.4–2.9). The carrier-vs-control ordering agrees with the projection measure, but the roundtrip baseline itself does not replicate across dictionary seeds (own-carrier cost -0.16 and $+0.17$ nats at two 160M replicate dictionaries whose baselines differ by 1.4–2.0 nats). Any claim about load *trends across training* measured under the roundtrip harness (e.g. Fig. 3b) inherits this instability and should be read as suggestive, not as a replicated trajectory.

Independent-initialisation replication. The checkpoint SAEs share an initialisation seed, so persistence could in principle reflect optimisation anchoring rather than stable model structure. Retraining all 24 Pythia-160M checkpoint SAEs from 24 distinct initialisations and rebuilding the lineage end-to-end: the scaffold persists, with 39 carriers against 51, the birth fraction, the predictor AUC and the carrier load all within a tenth of their canonical values, and the L11 extinction reproducing. The one quantity that moves materially is the carrier-to-control load ratio, which falls from $3.2\times$ to $1.6\times$ (independent-init lineages are noisier and the control pool retains some genuinely persistent features). Shared initialisation modestly inflates the carrier count and effect size; it does not create the scaffold. This is the replication Table 3's caption quotes. App. F reports a different one, end-to-end re-derivation of carriers under three independent dictionary seeds, whose breadth-matched ratios stay at 2.2–$3.2\times$. The two differ in what is reseeded and both are reported.

E CAUSAL MEDIATION: RULING OUT ALTERNATIVE EXPLANATIONS FOR THE CONTAINMENT EDGE EFFECT

This appendix supports §2.3 and §3.4. It gives five tests, each ruling out one alternative explanation for why ablating a containment parent cuts its child's firing, plus a harness-validity check.

Which edges the test can use. Containment is ordered by firing breadth rather than by depth (App. A), and the three cases that ordering produces behave differently under ablation. One encoding pass fixes every latent of a dictionary, so removing a parent's decoder contribution from the reconstruction leaves a same-layer child untouched and a within-layer test reads zero by construction. Where the parent lies deeper than its child the model runs from the child to the parent, so that pair is tested with its roles swapped, ablating the shallower member and reading the deeper one. That arm transmits a median 0.72% at Pythia-160M and 0.35% at Pythia-410M against 0.00% for its own matched controls, so a containment relation pointing against the residual stream carries almost nothing, and the coupling reported below follows the model's computation rather than the relation alone. The headline test is the remaining case, the $1{,}249$ Pythia-160M edges whose parent is the shallower feature.

Harness validity: placebo and determinism. A placebo arm ablates one member of a matched control pair and measures the effect on the other (control-vs-control): the paired median effect is 0.00000 (95% CI $[-0.00006, +0.00007]$, sign test $p = 0.86$), confirming the harness carries no baseline bias. A bit-identical re-run of the full pipeline reproduces the reported effect exactly.

Edges share parents. The harness caps each arm at 250 distinct ablated parents, drawn at random from the 402 available at Pythia-160M and the 601 at Pythia-410M, which leaves the 791 and 907 edges reported here (median 2 edges per parent, at most 32 and 34). Because those edges share parents, a sign test over edges overstates the independent evidence. Taking the parent as the unit instead, the median paired firing difference over a parent's edges is positive for about 94% of parents at both scales (sign test $p = 6.8\times10^{-61}$ and 1.6×10^{-64}), and a bootstrap that resamples parents with replacement and pools their edges gives a 95% interval on the mean firing difference that overlaps the edge-level interval at both.

The control pool is covariate-balanced. A containment DAG excludes every broad, high-CI feature from "non-parent" contention by construction, so the control pool is matched per source's own descendant set. It balances breadth (0.0402 against 0.0393 at Pythia-160M, 0.0307 against 0.0310 at Pythia-410M) and matches the Containment Index exactly, so the effects of §3.4 are not an artefact of an unmatched control.

Co-occurrence alone is very nearly inert. Two controls isolate statistical association from containment. Non-parents matched on raw co-firing overlap with the same child (median overlap 0.601, i.e. *more* overlap than the containment threshold itself) transmit a median 0.99% drop where containment transmits an order of magnitude more (sign $p = 3.3 \times 10^{-14}$, $n = 183$). A stricter control matched on *lift* (association strength) rather than raw overlap transmits 0.00% (sign $p = 4.2\times10^{-24}$). Co-occurrence, however measured, transmits at most a small fraction of what containment does.

What the readout measures, and why the control is exactly zero. Firing is the fraction of the $8{,}192$ evaluation tokens on which the child's encoder activation exceeds 10^{-6}, so the effect is discrete in steps of $1/8{,}192$: a control median of exactly 0.00% means ablating the source flips no token at all, rather than a small effect rounding to zero. This follows from containment itself, since the parent's decoder contribution is present on almost all of the child's firing tokens while the breadth- and CI-matched control fires on a largely disjoint set. Children with few firing tokens are not what carries the effect. Stratified by firing count, the median relative drop falls from 13.5% for children firing on 31–100 tokens to 7.5% for those above 300, while the median absolute loss rises across the same strata, and restricting to the 716 of 791 children firing on at least 31 tokens leaves the median at 10.8%.

Not reducible to geometry. A graph built instead from cross-layer decoder-cosine alignment, matched to the containment DAG's edge count, is also load-transmitting (8.9% median drop, 95%

of edges parent-over-control) but selects an almost entirely disjoint set of edges from containment (containment edges sit at median cosine 0.11). Read-write geometric alignment and token-set containment are two complementary causal channels, not the same one measured twice.

Not one instrument's artefact. Repeating the full pipeline end to end on an independently initialised SAE (different seed, same activations) reproduces both the causal effect (10.21% $[8.45, 11.36]$ vs. the canonical 11.11% $[10.17, 12.24]$, control 0.00% both) and the L11 substrate failure (App. F.6). The graph is a property of the model, not the dictionary. That is the seed-level form of the objection from the sparse-coding literature, where splitting and absorption would place the hierarchy in the decomposition. App. F.2 takes that objection at the level of the decomposition itself.

Not in-sample overfitting. Evaluating on token chunks held out from those used to define containment reproduces the effect (10.64% vs. the in-sample 11.11%; 471 edges, sign $p = 6.0 \times 10^{-76}$); the in-sample estimate carries a modest ($\sim 4\%$ relative) winner's-curse inflation and no more.

When does the effect appear? Repeating the edge test at each Pythia-160M checkpoint traces the graph from step 1k (6.82% of 878 edges, 82.5% parent-over-control) through a step-10k peak (18.10%) to its converged value (11.11% at step 143k); all adjacent-window differences exclude zero by bootstrap. Pythia-410M reproduces the step-1k gateway, at 71% of converged strength against Pythia-160M's 61%, but not the step-10k peak, which there sits at half of convergence rather than above it: the assemble-then-relax *curve shape* is 160M-specific and is not claimed beyond that scale. What generalises across both scales is the gateway result itself: the causal graph is substantially wired within the first 1% of training.

F Robustness, Generalisation, and Two Anomalies

This appendix collects checks that support the pipeline as a whole rather than one specific claim: whether apparent life history is dictionary noise, whether containment is an artefact of sparse decomposition, whether the persistence threshold is doing the work, whether persistence adds coverage beyond containment alone, replication in a second model family, and two notable single-layer/qualitative anomalies the main results surface.

F.1 Is it dictionary noise?

Retraining SAEs from independent seeds on the same checkpoints quantifies how much apparent life history is dictionary noise. Same-checkpoint dictionaries match (Hungarian, $r \geq 0.5$) at only 45–66% of active features across the two audited scales, so population-level event counts carry a substantial noise floor, which App. C's correction accounts for. The structures this paper is about are far more stable. Root-tier features match at 53–100% and the collapsed 160M-L11 population at 86–92%, and every carrier at Pythia-160M and all but one at Pythia-410M have a counterpart at $r \geq 0.5$ in an independently initialised dictionary trained on the same step-143k activations, at median best r 0.74 and 0.76. End-to-end re-derivation under three independent dictionary seeds gives 52 to 57 carriers against the 51 canonical ones, and their load replicates under the projection measure at ratios of 2.2–$3.2\times$ against breadth-matched controls. App. D's reseeding of all 24 checkpoint SAEs rebuilds the lineage as well and is the experiment that gives $1.6\times$. Individual membership, meanwhile, overlaps only partially (Jaccard 0.12–0.17). The scaffold's statistical signature is dictionary-robust, and its membership list is not.

What the dictionary-relativity pattern means. A one-to-one identity match compounds across six checkpoints: even a reliable per-transition match rate falls well below its own value once carried through the whole chain, enough on its own to produce a Jaccard in the observed range. The more informative comparison is what does and does not survive: firing-pattern correlation survives at 98–100%, while the exact dictionary atom assigned to carry that function does not. This is the paper's function/direction split appearing one level up, as a property of the measurement instrument as well as of training. App. B reads each clock against its own instrument ceiling, where the training gap narrows and survives. It also suggests where SAE non-canonicality concentrates: carriers, being broad

and high-containment, plausibly occupy a less crowded region of representation space than narrow, rarely-firing features, which would predict lower cross-seed superposition and higher recoverability for exactly the population this paper studies. We have not tested this directly. Read together, the pattern reinforces this paper's population-level framing: carrier claims describe a statistical stratum that every independently trained dictionary recovers.

F.2 Is containment an artefact of sparse decomposition?

Feature splitting decomposes one underlying feature into a broad latent and several narrow ones, and absorption leaves a parent latent silent on its child's tokens whenever sparsity prefers it (Chanin et al., 2025; Bussmann et al., 2025; Grandien et al., 2026). Either would manufacture the nesting this paper measures. Both also work among the latents of a single dictionary, which compete to reconstruct one activation vector, so the within-layer edges of the DAG are the channel they can reach. This section deletes that channel and reports what survives.

Within-layer edges are a minority of the reduced DAG. At Pythia-160M their share rises from 25% of edges at step 1k to 36% at convergence, at Pythia-410M it stays between 26% and 29% across the six snapshots, and the independently initialised Pythia-160M dictionary gives 35%. Recomputing the Containment Index on the cross-layer subgraph alone leaves its ordering in place, at Spearman 0.94–0.97 against CI on the full graph over the Pythia-410M snapshots and 0.95 in the independent dictionary. What moves is the tier boundary. The $\mathrm{CI} \geq 10$ gate that selects carrier endpoints keeps 73–80% of its members at Pythia-410M and 66% in the independent dictionary, while most of the 50 highest-CI features stay in the top 50. The ranking that the step-1k predictor reads (§3.2) is therefore close to unchanged under deletion, and the AUC of that predictor is reported on the full graph alone.

The signature of hard absorption can also be read straight off the edge population. Because $s(u, v)$ is the share of the child's firing tokens the parent covers, $1 - s(u, v)$ is the share on which the parent is silent, which is what absorption produces. Its median runs 0.17–0.28 over the Pythia-410M snapshots, and within-layer edges sit below their cross-layer counterparts at every one of them, as they do at Pythia-160M. The channel open to a dictionary-local mechanism thus carries less of the signature than the channel closed to it. The direction of the bias points the same way. Absorption lowers $s(u, v)$ on exactly the pairs that stand in a hierarchy, and the edge rule thresholds on $s(u, v)$ at the calibrated τ (0.6–0.7 at these snapshots), so the mechanism costs true edges rather than inventing false ones.

Two limits bound this test. Per-edge layer tags ship for Pythia-410M at all six snapshots, for the independently initialised Pythia-160M dictionary and for the step-0 dictionary, whereas Pythia-1B and OLMo-2-1B ship per-node CI alone, so extending the recomputation to them needs the activation store. The stronger condition of Cao et al. (2026) is a training objective written over each child's true concept direction, so a containment DAG built under it means training a different dictionary architecture at every layer and checkpoint rather than re-reading ours.

F.3 Does the persistence threshold do the work?

No: stricter thresholds yield a smaller but more coherent scaffold. Tightening the canonical criterion (≥ 4 of 6 post-init snapshots) to ≥ 5 moves the 160M carrier count from 51 to 43 while the born-by-step-1k fraction rises from 73% to 86%; at ≥ 6 (present at every checkpoint) the count falls to 37, all born by step 1k by construction (Table 12). The core finding strengthens, rather than weakens, under stricter thresholds.

Table 12: 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%

A higher-budget SAE control at Pythia-160M ($5\times$ the training samples, $2.5\times$ the steps, two instrumented layers) reproduces reconstruction fidelity and alive-feature counts at carrier-relevant breadth thresholds.

F.4 Does persistence add organisational reach beyond containment?

A persistence-neutral control seeds the same sub-DAG breadth-first search from non-persistent high-CI roots (CI ≥ 10 but not carriers), matched to the carrier set per layer on firing breadth and count. This null already covers $52.5\%/35.8\%/38.7\%$ of active features at 160M/410M/1B, against carrier coverage $63.7\%/52.6\%/29.3\%$: persistence adds reach beyond high-CI selection at the two smaller scales ($1.21\times/1.47\times$) but not at Pythia-1B ($0.76\times$, where the 27 carriers concentrate in the deepest instrumented layer). A non-Pythia model at the same scale (OLMo-2-1B, below) retains the increment at $1.48\times$. Coverage is therefore largely a property of high containment; the persistence-specific increment is modest and scale-inconsistent within Pythia. Features inside the sub-DAG fire more narrowly than the residual, $4.4\times$ at Pythia-160M, weakening with scale and absent at OLMo-2-1B (Fig. 7).

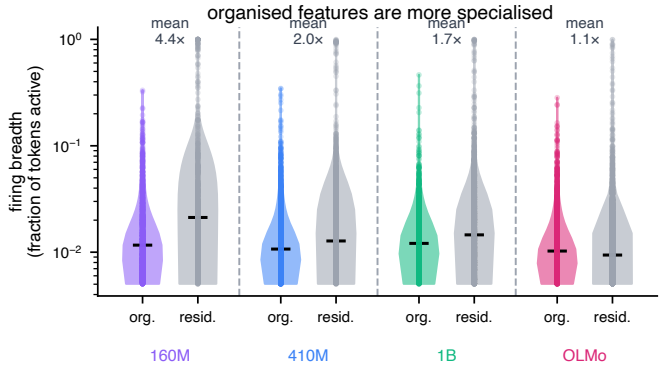


Figure 7: Firing breadth at step 143k for scaffold-organised features against unorganised residuals (grey), at all four models, as full distributions with medians marked in black. Mean breadth is $4.4\times$ lower inside the sub-DAG than outside at Pythia-160M, $2.0\times$ at Pythia-410M, $1.7\times$ at Pythia-1B, and $1.1\times$ at OLMo-2-1B, so the narrowness contrast weakens monotonically with scale and is absent at OLMo-2-1B.

F.5 Cross-Architecture Replication: OLMo-2-1B

To test whether the carrier scaffold is specific to Pythia or the Pile, we ran the identical pipeline on OLMo-2-1B (Team OLMo, 2025; Allen Institute for AI, 2025) (SwiGLU, RMSNorm, sequential blocks; Dolma-based corpus), instrumenting four layers at matching fractional depths (L3/L7/L11/L15 of 16) across six checkpoints fraction-matched to the Pythia schedule, with the identical carrier definition, load measure, coverage null, and predictor. All four signatures reproduce (Table 13); on the two signatures that were a boundary case at Pythia-1B (coverage and predictability), OLMo-2-1B behaves like the *smaller* Pythia scales, indicating that boundary is Pythia-1B-specific rather than a property of the 1B parameter scale.

Table 13: Four-claim comparison of OLMo-2-1B against Pythia-1B under an identical pipeline. OLMo-2-1B reproduces all four signatures; its coverage and predictability do not exhibit the Pythia-1B boundary, indicating that boundary is specific to Pythia-1B rather than a property of the 1B scale.

Signature	Pythia-1B	OLMo-2-1B
Formation (carriers; % born by first snapshot)	27; 44%	43; 53%
Load (carrier dCE vs. breadth-matched control)	$2.3\times$ (CIs disjoint)	$2.1\times$ (CIs disjoint)
Coverage (carrier sub-DAG vs. CI-matched null)	$0.76\times$ (boundary)	$1.48\times$
Predictability (step-1k pooled CV AUC)	0.69 (boundary)	0.81

OLMo-2-1B's carriers concentrate in the deepest instrumented layer as at Pythia-1B, with familiar broad-structural content: a year detector, a person-name and title feature, legal boilerplate, subword fragments. All four signatures reproduce, and on coverage and predictability OLMo-2-1B sits with the smaller Pythia scales rather than at Pythia-1B's boundary (Table 13). Its step-1k predictor holds up layer by layer as well, at leave-one-layer-out AUC 0.72–0.93. Three robustness controls were repeated at this scale. Birth concentration is at least as strong as at the Pythia scales, at $32.7\times$ the later-window mean against 22–29$\times$ on the same uncorrected reading. Persistence survival is a weak, positive function of step-1k firing breadth, at 6–18% across quartiles. And on direction-matched lineages function leads direction at step 1k, at $r = 0.68$ against cosine 0.20 over 61 lineages (App. B). OLMo-2-1B behaves as a genuine same-scale, different-family instance of the phenomenon.

F.6 Two anomalies, briefly

The L11 extinction (Pythia-160M). Between steps 80k and 120k, Pythia-160M's L11 loses its scaffold capacity entirely (zero carriers), invisible to the monotonically improving validation loss. Active SAE features collapse from $\sim$612 to $\sim$39, the residual stream's effective rank collapses to $\sim$13 dimensions with the surviving variance absorbed into the unembedding subspace, and the SAE reconstruction becomes harmful (loss recovery -0.33, against $+0.33$ at the depth-matched Pythia-410M-L23 control); a $10\times$ larger SAE training budget does not change it. The same relative depth integrates normally at larger scale (Pythia-410M-L11 and Pythia-1B-L15 both contribute carriers, the latter with effective rank > 150), so the extinction is a small-model event rather than a generic deepest-layer phenomenon.

Qualitative anatomy of the carrier scaffold. Of Pythia-160M's 51 carriers, 39 are labelable from top-activating tokens; 82% of those receive high-or-medium-confidence auto-interpretation labels (Table 14), spanning morphological/discourse content at L2/L5 and sentence-initial framing/modal predicates at L8, with a residual low-confidence tail of Unicode/CJK/HTML-boundary artefacts concentrated at L2. A blind second-rater validation, in which Claude Opus 4.8 saw only top-activating tokens with GPT-5.6 Sol's label hidden, agrees with the auto-interp concept label on 18 of 20 randomly sampled carriers (90%). Confidence-level agreement is lower (50%), driven mainly by the second rater reading several byte-boundary carriers as coherent encoding-artefact detectors where auto-interp scored them low-confidence. This validates using confidence tier as a covariate in App. D's legibility-vs-load check. The highest-CI labelable carrier, L8 fid 4795 (sentence-initial framing connectives: *Although, Despite, However*; CI $= 243$), has 200+ descendants spanning discourse-structural nested roots (e.g. a concessive-pronoun feature, paragraph-continuation markers) down to narrow lexical leaves; the content gradient is broad-to-narrow but not strictly taxonomic (99.6% of scaffold edges are subset-ratio *predictive* rather than taxonomic, vs. 95.2% for non-scaffold edges).

Table 14: Auto-interpretation confidence for the 51 persistent carriers at Pythia-160M (step 143k, expansion-factor-64 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, above)
Total	51	20	12	19	82% high/medium of 39 labelable