GEOMETRIC PRECURSORS OF LEARNING TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A model can begin reorganizing internally before a new capability becomes meaningfully detectable in its output behavior. After that point, intervention may require expensive rollbacks during training, so efficient steering requires cheap early detection methods. To this end, we develop a geometric framework for understanding and detecting such changes by tracking latent representations on relevant inputs. By constructing Gram matrices out of latent representations, we are able to give an invariant decomposition of motion during learning into growth, spectral allocation, internal rotation, and support transport. Our central theoretical result relates the amplitude of an emerging response to the rate of its geometric motion, explaining when normalized geometry can change before amplitude-dependent measurements reveal it. Using this description, we derive scalar early warning signals of learning. We then test this method in three settings (deep linear networks, induction heads, and the Pythia suite), each with decreasing prior knowledge about future learning events. In deep linear networks, support transport predicts the next acquired mode. For induction, alarms based on representations alone precede behavioral expression and remain silent on controls. Across Pythia scales, we identify recurring internal reorganizations and a shift from early growth to later internal rotation.

1 INTRODUCTION

Learning curves are a standard way to track capability development in large language models (Evanson et al., 2023). They show when a capability becomes expressed, but internal reorganization can begin much earlier. This separation can make development appear more sudden than it is (Schaeffer et al., 2023; Kangaslahti et al., 2026). It appears in silent alignment (Atanasov et al., 2022), deep-linear dynamics (Saxe et al., 2014; 2019), hidden progress on algorithmic tasks (Power et al., 2022; Barak et al., 2022; Nanda et al., 2023; Singh et al., 2024), and stagewise transformer development (Hoogland et al., 2025; Chen et al., 2024; Edelman et al., 2024). Once a capability is expressed, intervening may require an expensive rollback of training, so the internal change is worth detecting while it begins. This raises two questions: what internal change should be measured, and when can that change precede behavioral expression?

Existing early-warning methods use changes in per-sample loss correlations and their spectra (Hennick & Corlouer, 2026), connecting training to critical-slowing-down diagnostics based on fluctuations near critical points (Scheffer et al., 2009; Kuehn, 2011). Various detectors have also been built for specific transitions such as grokking and in-context learning (Nanda et al., 2023; Reddy, 2024). Representation similarity measures have long been used to describe how representations develop over training (Raghu et al., 2017; Kornblith et al., 2019; Saphra & Lopez, 2019), but they compare pairs of snapshots, often against a final checkpoint that an online observer cannot yet see. Theories of learning from small initialization describe staged acquisition as a sequence of saddle escapes and rank increases (Saxe et al., 2014; Gidel et al., 2019; Li et al., 2021; Pesme & Flammarion, 2023; Simon et al., 2023; Boix-Adsera et al., 2023), and developmental interpretability tracks stagewise structure through loss-landscape geometry (Lau et al., 2025; Hoogland et al., 2025; Wang et al., 2025). What is missing is an account of how the representations themselves move during training: what kind of change is occurring, when it can precede behavior, and which scalar an online observer should monitor.

We address these questions through the geometry of the representations themselves. We describe representations measured on fixed inputs through their Gram matrices, whose motion separates into growth, spectral allocation, internal rotation, and support transport (Section 2). An emerging response

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.

can reorganize while it is still too weak to produce a visible change in behavior. We derive a law for this separation between structure and strength, explaining how internal geometry can provide advance warning of a learning transition (Section 3). Declaring what counts as no relevant change then turns this geometry into an online signal: distance from the no-change set is the largest stable score (Section 4). We test this geometric account across a range of learning transitions (Section 5). We show that the method predicts known transitions in deep linear networks and induction heads (Olsson et al., 2022), and separates two stages within grokking (Appendix E). We then apply it without any event label to the Pythia suite (Biderman et al., 2023), where it finds recurring internal reorganizations. Appendix A.1 reviews related work in more detail, Appendix A.3 compares our detector with the loss-correlation signal on the same induction task, and Appendix A.4 relates the geometry to similarity measures.

2 Geometry of representational learning

We start by fixing a set of input probes $x_1, \dots, x_N$. At a timestep t of training, we record representations $h_t(x_i) \in \mathbb{R}^d$ (for example, from the residual stream or an intermediate layer) into a matrix $H_t \in \mathbb{R}^{N\times d}$. The observed state is

$$K_t = H_t H_t^\top, \qquad a_t = \operatorname{tr} K_t, \qquad \rho_t = K_t/a_t, \qquad P_t = K_t K_t^+. \tag{1}$$

Here a_t is the total response energy, ρ_t is its normalized shape, and P_t projects onto the represented probe subspace. Unlike H_t, the Gram state K_t is unchanged by an orthogonal change of feature basis, as in standard representation comparisons (Kriegeskorte et al., 2008; Kornblith et al., 2019; Williams et al., 2021).

Because the feature coordinates are arbitrary, we want a notion of representational motion that does not change under an orthogonal change of feature basis. We also want the same motion to have the same size if the representation and its velocity are rescaled together. These requirements essentially fix the local geometry. In particular, among isotropic feature metrics, normalizing pure scale growth to have unit speed gives

$$g_H(Z,Z) = \frac{4\|Z\|_F^2}{\|H\|_F^2}. \tag{2}$$

Thus motion is measured relative to the current size of the representation, rather than in arbitrary activation units.

We observe representations through the Gram state $K = HH^\top$, so feature velocities that leave K unchanged should not contribute to the observed motion. For a Gram velocity $\Xi = \dot K$ induced by a feature velocity $Z = \dot H$, we therefore assign the length of the shortest feature-space velocity that produces it:

$$g_K(\Xi,\Xi) = \min_{Z:\, ZH^\top + HZ^\top = \Xi} \frac{4\|Z\|_F^2}{\operatorname{tr} K}. \tag{3}$$

This quotient construction (Edelman et al., 1998; Absil et al., 2008) depends only on K, not on the particular factor H, and is the scale-normalized Bures geometry (Uhlmann, 1976; Bhatia et al., 2019; Williams et al., 2021; Harvey et al., 2024). It therefore gives an invariant notion of how quickly the observed representation is changing. The formal uniqueness statement, closed form, and proof are given by Lemma B.1 in Appendix B.

We write $\|\Xi\|_K^2 := g_K(\Xi,\Xi)$. A representation can now change in several qualitatively different ways: it can grow in overall magnitude, redistribute weight among existing directions, rotate those directions, or move into previously unrepresented directions. The next result shows that these are not merely useful descriptions: they form a complete orthogonal decomposition of every smooth Gram-state motion (Fig. 1). Write $\rho = \sum_{i\le k} p_i u_i u_i^\top$ with $p_i > 0$, and complete the u_i by u_α ($\alpha > k$) to an orthonormal basis. Then $\omega_{ij} = u_i^\top \dot u_j$ and $\omega_{i\alpha} = u_\alpha^\top \dot u_i$ are the angular velocities of this basis within the support and toward its complement.

Theorem 1 (Four motions of a Gram state; informal)**.** At any rank-k Gram state, every tangent splits uniquely into four mutually orthogonal pieces: growth, allocation, internal rotation, and support transport. Along a smooth constant-rank Gram path the squared speed is therefore the sum of four squared speeds,

$$\left\|\dot{K}\right\|_K^2 = \underbrace{\left(\frac{\dot{a}}{a}\right)^2}_{\text{growth}} + \underbrace{\sum_{i=1}^{k} \frac{\dot{p}_i^2}{p_i}}_{\text{allocation}} + \underbrace{4\sum_{i<j} \frac{(p_i - p_j)^2}{p_i + p_j}\,\omega_{ij}^2}_{\text{internal rotation}} + \underbrace{4\sum_{i\le k<\alpha} p_i \omega_{i\alpha}^2}_{\text{support transport}}. \tag{4}$$

The four pieces vary independently, exhaust the tangent space, and depend only on K, not on the factor H or on the basis used to describe the occupied subspace.

Growth changes overall scale; allocation moves energy among represented directions, at a speed given by the Fisher metric on the weights p_i (Amari, 2016); internal rotation turns unequal-weight directions inside the current support; and support transport moves represented energy toward previously unrepresented directions. On unit-trace states g_K is the quantum Fisher metric, and the last three terms are the spectral form of the quantum Fisher information of ρ (Braunstein & Caves, 1994; Liu et al., 2020). Our split separates eigenvector motion inside the support from motion out of it, and adds growth as an orthogonal log-scale direction. We denote the square roots of the four terms by v_{grow}, v_{alloc}, v_{rot}, and v_{trans}. The optimizer and parameterization determine which of these motions is realized. Appendix B.4 gives the formal response-operator description. Appendix A.2 relates it to the 2-datapoint reduced density matrix introduced by Hennick & Corlouer (2026).

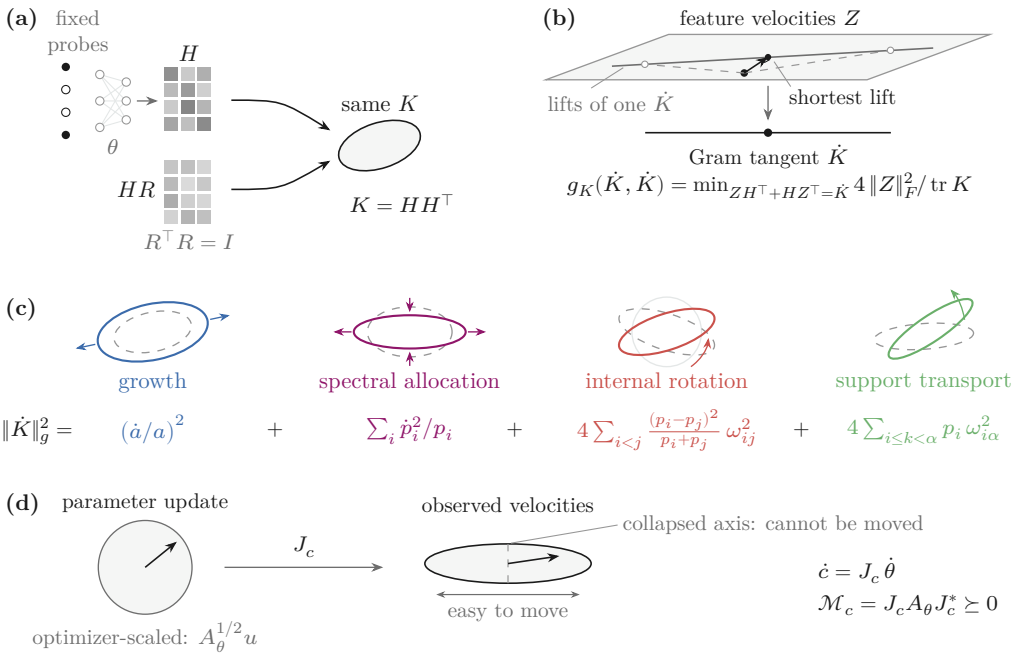

Figure 1: **Invariance defines the geometry of observed learning.** Fixed probes define a Gram state; scale and basis invariance fix its local metric; every smooth tangent decomposes into the four motions in Eq. (4); and the parameter update determines which motion is realized.

3 When geometry precedes expression

Section 2 identifies the motions available to a Gram state. For early warning, we also need to know how soon those motions can begin when the response is still small. Consider a family of emerging probe responses whose Gram states satisfy

$$K_\varepsilon(0) = \varepsilon^2 \rho_0, \qquad 0 < \varepsilon \ll 1, \qquad \rho_0 \succeq 0, \quad \operatorname{tr} \rho_0 = 1. \tag{5}$$

The response is small but nonzero, so its normalized geometry is well defined. The matrix K_ε may describe a full probe response or a residual after removing structure fixed earlier. We suppress the subscript below and write $r = \sqrt{\operatorname{tr} K}$ for response size and $\rho = K/\operatorname{tr} K$ for normalized geometry.

Near a small response, the leading term of the training dynamics is homogeneous of degree p in r: $p = 1$ when the linear term is nonzero, and $p > 1$ otherwise. Because $K = HH^\top$ is quadratic in the feature response, a feature velocity of order r^p induces a Gram velocity of the form

$$\dot{K} = r^{p+1}V + o(r^{p+1}), \qquad p \geq 1, \tag{6}$$

where the finite symmetric matrix V gives the leading direction of Gram motion and may vary along the training path. Differentiating r and ρ now gives

$$\dot{r} = \frac{1}{2} r^p \operatorname{tr} V + o(r^p), \qquad \dot{\rho} = r^{p-1}\big[V - (\operatorname{tr} V)\rho\big] + o(r^{p-1}). \tag{7}$$

The bracket in the second equation has zero trace, so it changes normalized geometry without changing response size. Equation (7) contains the central separation: response size moves at order r^p, whereas normalized geometry can move at order r^{p-1}. Pure amplification is the special case in which the bracket vanishes and the geometry does not move.

Theorem 2 (Response-order law)**.** Let the Gram path follow Eq. (6), and suppose a scalar measurement has leading form

$$c(K) = r^\nu \phi(\rho) + o(r^\nu), \qquad \nu \geq 0.$$

If the leading Gram motion changes this measurement, then

$$|\dot{c}| \asymp r^{p+\nu-1}. \tag{8}$$

A scale-free geometric measurement has $\nu = 0$ and responds at order r^{p-1}. Response amplitude has $\nu = 1$ and responds at order r^p, while a raw Gram quantity has $\nu = 2$ and responds at order r^{p+1}.

The response-order law ranks measurements using two quantities. The training dynamics set p, and the measurement sets ν. No solution of the full training trajectory is needed. Appendix C gives the precise assumptions, leading coefficient, proof, and extensions.

Response order is a local statement about speed, not yet a statement about when a threshold is crossed. Let T_{geom} be the first time ρ travels a fixed geometric distance, and let T_{expr} be the first time an amplitude-dependent readout reaches a fixed expression level. If the geometric motion persists toward its crossing and the response size grows, then two timing regimes follow:

$$\begin{aligned} p = 1: &\quad T_{\mathrm{geom}} = O(1), && T_{\mathrm{expr}} = \Theta(\log(1/\varepsilon)), \\ p > 1: &\quad T_{\mathrm{geom}}, T_{\mathrm{expr}} = \Theta(\varepsilon^{1-p}), && \text{with a possible finite fractional lead.} \end{aligned} \tag{9}$$

When $p = 1$, normalized geometry can move on an order-one training clock, while a response starting at size ε needs order $\log(1/\varepsilon)$ time to reach a fixed expression level, as in the plateaus of two-layer linear networks (Saxe et al., 2014) and their silent alignment (Atanasov et al., 2022). When $p > 1$, the whole emerging response slows down. Geometry may still lead, but the lead is generally a finite fraction of the shared timescale rather than a separate asymptotic clock. The crossing-time bounds are proved in Appendix C.

An early geometric change may still be too small to measure. Estimating $\rho = K/\operatorname{tr} K$ becomes unstable when observation noise is large relative to $\operatorname{tr} K$ (Davis & Kahan, 1970; Wedin, 1972). Let T_{vis} be the first time the geometry can be resolved, and let δ_{obs} be the delay from checkpoint spacing and persistence. The first possible alarm is then

$$T_{\mathrm{alarm}} = \max\{T_{\mathrm{geom}}, T_{\mathrm{vis}}\} + \delta_{\mathrm{obs}}. \tag{10}$$

Thus response order determines the available latent lead, while resolution and sampling determine how much of that lead can be observed. A fixed observation floor can erase the relative warning interval. Appendix C.4 derives the corresponding perturbation, visibility, and sampling limits.

4 From motion to a causal event

The previous section tells us when geometric change can become visible before behavioral expression. To build an online detector, however, we still need to decide *which* geometric change matters. We do this by specifying what would count as "no relevant change" and measuring departure from that set.

Definition 1 (No-change family)**.** Let $(\mathcal{X}, d_g)$ be the metric space in which the chosen observation is measured. A *no-change family* is a nonempty closed set $\mathcal{N} \subset \mathcal{X}$ containing exactly those observations regarded as unchanged for the question of interest. We define the corresponding departure score by

$$S_{\mathcal{N}}(x) := d_g(x, \mathcal{N}) = \inf_{y \in \mathcal{N}} d_g(x, y). \tag{11}$$

The construction is deliberately question dependent. If we care about growth, for example, $\mathcal{N}$ contains tangents with no growth component; if we care about support transport, $\mathcal{N}$ contains tangents whose occupied subspace is fixed. For a path break, $\mathcal{N}$ instead contains velocities consistent with continuation of the recent path, so the distance is an innovation in the sense of filtering (Kailath, 1968).

Distance to $\mathcal{N}$ is a natural scalar score because it has a simple optimality property. Suppose we require a score to vanish whenever no relevant change has occurred and to be stable to observation error, in the sense that a perturbation of size δ can change the score by at most δ. Then no such scalar can exceed the distance to the no-change family:

$$|f(x)| \le d_g(x, \mathcal{N}), \qquad f|_{\mathcal{N}} = 0, \qquad \mathrm{Lip}_g(f) \le 1. \tag{12}$$

Equality is achieved by $f(x) = d_g(x, \mathcal{N})$, the McShane extension of the zero function on $\mathcal{N}$ (McShane, 1934). Thus, once the metric and the notion of "no change" have been specified, distance gives the largest scalar signal compatible with this stability requirement. The formal statement and proof are given in Appendix D.1. Near a smooth no-change family, this distance has the same leading order as the component normal to that family. It therefore preserves the earliest response order carried by the relevant geometric motion.

For the four motions this becomes especially simple. Let Π_B denote the g_K-orthogonal projection onto one geometric block $B \in \{\mathrm{grow}, \mathrm{alloc}, \mathrm{rot}, \mathrm{trans}\}$, and define

$$\mathcal{N}_B = \{\Xi : \Pi_B \Xi = 0\}. \tag{13}$$

Then

$$d_K(\Xi, \mathcal{N}_B) = \|\Pi_B \Xi\|_K = v_B. \tag{14}$$

Hence the four block speeds are not arbitrary summaries of the Gram velocity: each is exactly the distance from the family in which that type of motion is absent (Table 1).

Table 1: **From a proposed change to its geometric measurement.**

Change of interest	No change means	Measurement selected by the theory
Instantaneous motion		
Any motion	$\Xi = 0$	Total speed $\|\Xi\|_K$
Growth	No growth component	$v_{\mathrm{grow}} = \|\dot{a}/a\|$
Allocation	No redistribution among the p_i	Fisher speed v_{alloc}
Internal rotation	No turning within the support	v_{rot}
Support transport	The occupied support is fixed	v_{trans}
Change along a path		
Net departure	The state remains in a reference family	Geometric displacement
Accumulated motion	Zero velocity along the path	Path length
Path break	The current velocity follows recent continuation	Innovation distance
Support or rank entry	The state remains inside the old support or rank family	Support or rank distance

The preceding sections now give a complete recipe for early detection. Fixed probes turn the evolving network into a Gram-state trajectory K_t. Its motion can be decomposed into interpretable geometric changes, and these changes can become visible while the corresponding response is still too small to be expressed behaviorally. The distance rule then turns a proposed change into a scalar measurement: specify the family $\mathcal{N}$ in which that change has not occurred, and measure distance from it.

The resulting picture is simple. Rather than waiting for a behavioral quantity to become large, we ask whether the internal representation has already begun to move in a way that is inconsistent with "no relevant change." At checkpoint t_j, the observer therefore computes

$$c_j \longmapsto s_j = d_g(c_j, \mathcal{N}_j), \tag{15}$$

where both the observation c_j and the no-change family $\mathcal{N}_j$ are constructed using only information available up to t_j. A large s_j means that the current representation has departed from the specified no-change regime.

To turn this geometric departure into an online event, we first record a *watch* when the score crosses a causal threshold,

$$T_{\text{watch}} = \min\{t_j : s_j > h_j\}, \tag{16}$$

and confirm an *alarm* only when that departure persists for n_{pers} checkpoints,

$$T_{\text{alarm}} = \min\{t_j : s_{j-\ell} > h_{j-\ell} \text{ for } \ell = 0, \dots, n_{\text{pers}} - 1\}\,. \tag{17}$$

This is the operational form of the alarm time in Eq. (10), with persistence contributing to δ_{obs}. Thus a watch marks the first detectable geometric departure, while persistence distinguishes a sustained internal reorganization from an isolated fluctuation, as in sequential change detection (Page, 1954). A later behavioral crossing is recorded separately. The theory therefore provides an *early detector of internal change*.

5 Experiments: decreasing prior knowledge

The experiments relax prior knowledge in stages. Deep linear networks supply a known future direction and a matched geometric prediction. Induction keeps the capability label but gives the online detector no access to its mechanism or behavioral readout. Pythia finally removes the event label and asks what a fixed observer finds over a long language-model run. We also test whether the geometry can resolve multiple stages within a single long transition. In modular-arithmetic grokking (Power et al., 2022; Nanda et al., 2023), it separates an early post-interpolation reorganization from a later spectral-concentration stage, both of which are absent in no-weight-decay controls despite memorization. Because this experiment concerns stage resolution rather than blind precursor detection, we defer the full results to Appendix E. Full protocols, thresholds, controls, and robustness analyses are in Appendices F–H.

5.1 Deep linear networks: matched transport precedes acquisition

We first test the framework in a setting where the next learned feature is known exactly. Geometric support transport anticipates acquisition of the next teacher mode in all 12 evaluation runs (Fig. 2), with a median lead of 1,750 training steps (range 50–11,300). Moreover, the residual direction present at the alarm later becomes the acquired teacher direction in every run, while none of eight rank-one control runs produce an alarm. Thus the geometric signal predicts not only *when* a new mode is beginning to form, but also *which direction* will subsequently be learned.

We train a five-layer linear student on a rank-three teacher with unequal singular values, so that the teacher modes are acquired sequentially (Saxe et al., 2014; 2019). After the leading mode has stabilized at time t_0, the online observer removes that established component and tracks the strongest remaining probe-space direction. Let $P_2(t)$ project onto the two strongest student probe directions and let $\Pi_2^\star$ project onto the leading two teacher directions. The teacher is used only offline to define the acquisition event through

$$S_2(t) = \operatorname{tr}(\Pi_2^\star P_2(t))\,. \tag{18}$$

The online detector, by contrast, has no access to the teacher direction. If $P^{\text{res}}(t)$ denotes the leading residual projector, it measures

$$D_{\text{res}}(t) = \frac{1}{\sqrt{2}} \left\| P^{\text{res}}(t) - P^{\text{res}}(t_0) \right\|_F, \tag{19}$$

the distance from the no-transport family in which the residual support remains fixed.

An independently constructed weight-space projector produces an alarm in all 12 runs as well, with a median absolute difference of only 37.5 steps from the representation-space alarm. Together with the

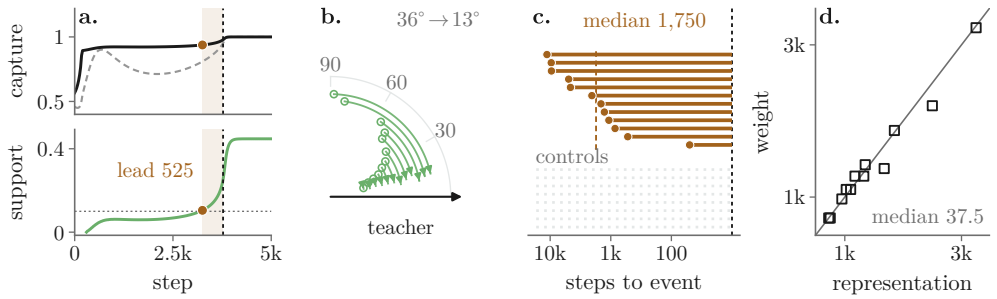

Figure 2: **Conditioned support transport warns of a new deep-linear mode. a,** In one run, residual support displacement (lower) crosses its alarm before cumulative top-two teacher-subspace capture (upper) reaches the acquisition event. **b,** Across runs, the residual direction at the alarm (open circles) turns toward the next teacher direction by the event (arrowheads). **c,** Alarm-to-event intervals for all 12 positive runs; the eight rank-one controls below remain quiet. **d,** Alarm times from representation and weight projectors lie near the identity line.

rank-one controls, this provides a particularly clean precursor test: a causally observed geometric direction appears before acquisition, later becomes the learned task direction, disappears when no additional mode exists, and is recovered independently in both representation and weight space.

5.2 Induction heads: a blind path break precedes expression

We next remove knowledge of the mechanism and ask whether representation geometry can blindly detect the onset of induction (Olsson et al., 2022; Singh et al., 2024; Reddy, 2024). The resulting path-break alarm precedes induction (Fig. 3) in all 28 transition runs and remains quiet in all 12 controls. Using four earlier runs to calibrate a single alarm-to-induction latency, the same rule predicts induction in 24 later runs with a median absolute error of 30 steps, compared with 50 steps for a fixed training-time baseline. Every forecast is issued before induction is expressed.

The key constraint is that the online observer cannot directly see the behavior it is meant to anticipate. We train a two-layer autoregressive transformer on a mixture of random sequences and repeated sequences $[x, x]$, but the detector records representations only on fixed random *first-half* prefixes. It receives no repeated suffix, attention-head identity, induction score, token-lag statistic, loss, or behavioral readout. Its no-change family is simply continuation of the recent normalized Gram path.

For probe bank b, let $\rho_j^{(b)}$ denote the normalized Gram state at checkpoint j and let $\mathcal{C}_j^{(b)}$ contain velocities predicted by causal continuation of its earlier path. We measure the innovation

$$I_j^{(b)} = d_g\left(\dot{\rho}_j^{(b)}, \mathcal{C}_j^{(b)}\right), \tag{20}$$

and pool the standardized departures across fixed probe banks,

$$s_j^{\mathrm{ind}} = \mathrm{median}_b\, \mathcal{Z}_{<j}\left(I_j^{(b)}\right), \tag{21}$$

where $\mathcal{Z}_{<j}$ standardizes each value using earlier checkpoints only. Thus an alarm indicates that several representation paths depart coherently from what their own past motion would predict, without using induction itself to define the event.

After the blind detector localizes this early path break, a separate repetition-conditioned observation reveals a later mechanistic sequence. Subtracting responses to token-matched permutations and removing the established first-half support gives

$$C_t^{\pi} = (I - P_t)\,(B_t - B_t^{\pi}), \qquad K_t^{\mathrm{rel}} = \mathbb{E}_{\pi}\left[C_t^{\pi} C_t^{\pi\top}\right]. \tag{22}$$

The spectrum of K_t^{rel} first concentrates and then broadens as induction is expressed. Heads selected by the broadening stage have the larger joint-ablation effect in all 12 independent replication runs. The blind path break therefore identifies an early internal transition, while the conditioned geometry resolves a later sequence associated with the emerging induction mechanism.

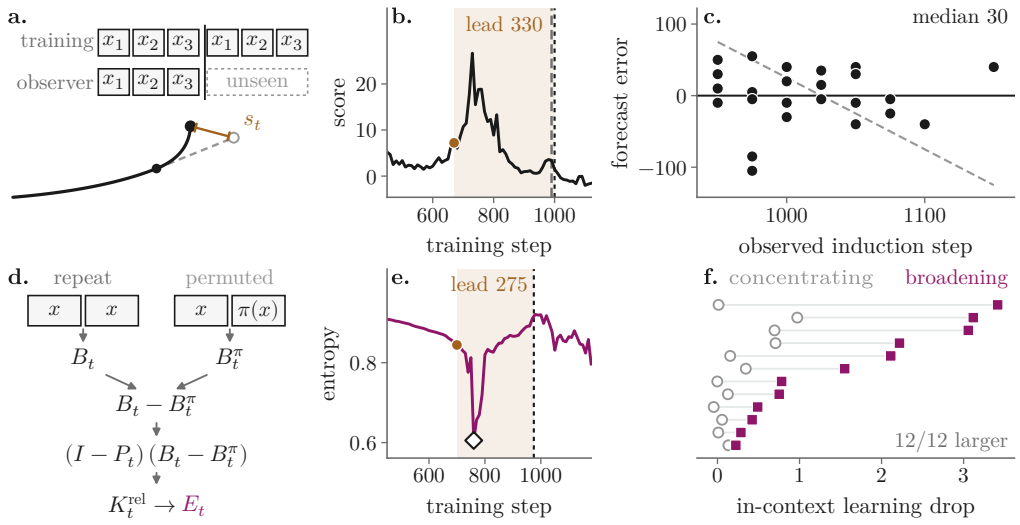

Figure 3: **A blind path break precedes induction, and conditioned geometry resolves its internal sequence. a,** The blind observer sees only fixed first-half prefixes and measures departure from causal path continuation. **b,** Its alarm precedes induction in a representative run. **c,** The event-locked clock predicts later induction times more closely than a fixed schedule. **d,** A separate repeated-minus-permuted observation removes established support. **e,** Its entropy first concentrates and then broadens. **f,** Broadening-aligned heads have the larger joint-ablation effect in all 12 replication runs.

5.3 Pythia: blind event discovery across scale

Finally, we remove the event label itself and ask what a fixed geometric observer discovers over long language-model training runs. Applying the same observer to Pythia 160M, 410M, and 1B (Biderman et al., 2023) reveals a common qualitative progression: early training is dominated by growth, while later training is dominated by internal rotation (Fig. 4a). After step 2,000, the median rotation fractions are 0.750, 0.796, and 0.666 for the three model sizes, respectively. Thus later learning is primarily reorganization of already represented directions rather than uniform amplification.

At each checkpoint, the observer uses three fixed views of the model: representations on natural text, hidden-state responses to repeated sequences relative to token-matched permutations, and the corresponding attention responses,

$$H^{\mathrm{nat}} = h(x), \qquad H^{\mathrm{rel}} = h(x^{\mathrm{rep}}) - h(x^{\mathrm{perm}}), \qquad H^{\mathrm{att}} = A(x^{\mathrm{rep}}) - A(x^{\mathrm{perm}}). \tag{23}$$

Each view defines a normalized Gram path. At every layer and fixed probe split, we measure departure from causal continuation of the recent path and combine these fixed scores into one observer. No behavioral measurement is used to define an event.

For Pythia 1B, this geometry-only observer identifies seven persistent event intervals, at approximately 8k, 12k, 18k, 24k, 29k, 107k, and 122k training steps (Fig. 4b). Every observation bank contributing to these intervals is dominated by internal rotation. Behaviors are examined only after this event catalogue has been fixed.

The 24k interval illustrates what such an unlabeled event can localize. Among 45 same-cadence comparison intervals, it contains the largest observed change in induction attention, the largest change in gradient norm, and the second-largest change in repetition gain. These quantities move at different checkpoints within the interval, so we interpret the event as a localized internal reorganization rather than a universal simultaneous behavioral jump.

The main conclusion is therefore deliberately modest: a fixed geometry-only observer can identify recurring internal reorganizations without specifying the target capability in advance, and some of these intervals contain unusually large downstream changes. Geometry localizes *when* the representation is reorganizing, but does not by itself determine the task meaning or sign of that change. The full

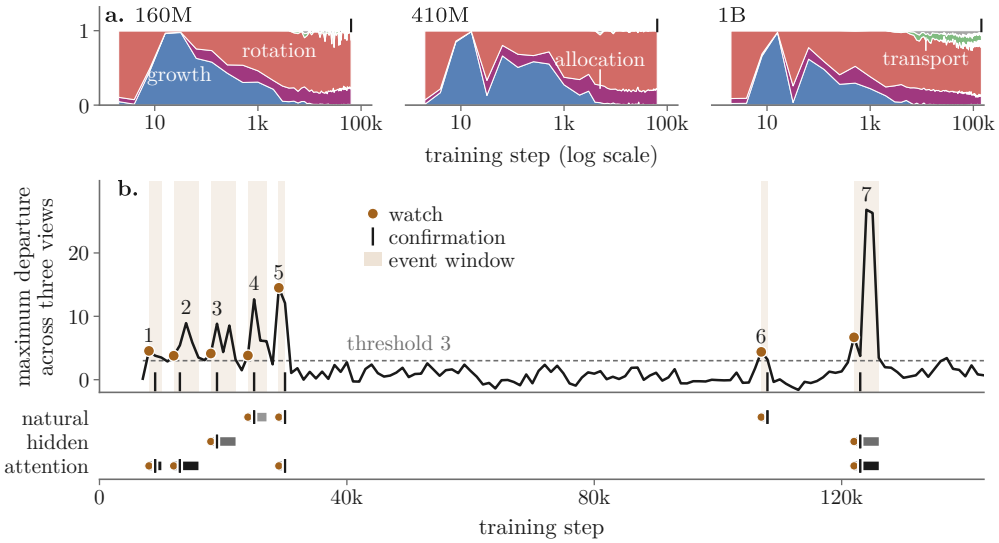

Figure 4: **One observer reveals a common Pythia progression and a blind event catalogue.** Natural text and repeated-minus-permuted contrasts define hidden-state and attention Gram paths at fixed checkpoints (schematic in Appendix Fig. 11a). **a,** Fractions of squared geometric speed show all three sizes moving from early growth toward later internal rotation. **b,** The pooled 1B continuation-departure score localizes seven unlabeled events; watch dots sit on the curve, confirmation ticks on the time axis, and the lower rows identify the observation banks contributing to each interval.

seven-event catalogue, intrinsic event hierarchy, cross-size-window audit, layer propagation, 107k unresolved event, 122k case, and complete behavior battery are in Appendix H.

6 Discussion

The central idea of this work is that learning can become geometrically visible before it becomes behaviorally expressed. Fixed probes turn training into a trajectory of representation Gram states; the resulting motion decomposes into growth, spectral allocation, internal rotation, and support transport. Near a weak emerging response, normalized geometry can move one response order earlier than amplitude-dependent observables, creating the possibility of a temporal lead. Once a notion of "no relevant change" is specified, distance from that family converts the geometric motion into a stable scalar event signal. Together, these ingredients give a principled route from internal representation dynamics to online early detection.

The experiments test this idea under progressively weaker prior knowledge. In deep linear networks, support transport anticipates both the time and the direction of the next acquired mode. In induction, a detector that cannot see the repeated suffix or the eventual behavioral readout nevertheless identifies a path break before induction is expressed; a separate conditioned view then resolves a later mechanistic sequence. In Pythia, the event label itself is removed: one fixed observer identifies recurring internal reorganizations across scale, with later training dominated by internal rotation rather than uniform growth. These results support a family of geometry-based detectors matched to different notions of change, rather than a single universal warning statistic.

Several limitations remain. The framework depends on the choice of probes and observation space, and sufficiently weak responses may remain statistically unresolvable even when a latent geometric change has begun. The present experiments establish that geometric events can precede and localize meaningful learning transitions, but they do not imply that every geometric event corresponds to a capability change or that every capability must produce an early resolvable precursor. A central next step is therefore to understand which classes of learned mechanisms generate reliable geometric precursors, how these signals scale with model and data size, and whether they can be used to trigger targeted mechanistic analysis or intervention during training.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. doi: 10.1515/9781400830244.

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical Learning Periods in Deep Networks. In *International Conference on Learning Representations*, 2019.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer Japan, 2016. doi: 10.1007/978-4-431-55978-8.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural Networks as Kernel Learners: The Silent Alignment Effect. In *International Conference on Learning Representations*, 2022.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden Progress in Deep Learning: SGD Learns Parities near the Computational Limit. In *Advances in Neural Information Processing Systems*, 2022.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. doi: 10.1016/j.exmath.2018.01.002.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory 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 *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023.

Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers Learn Through Gradual Rank Increase. In *Advances in Neural Information Processing Systems*, 2023.

Samuel L. Braunstein and Carlton M. Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439–3443, 1994. doi: 10.1103/PhysRevLett.72.3439.

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 *International Conference on Learning Representations*, 2024.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a High-Dimensional Abstraction Phase in Language Transformers. In *International Conference on Learning Representations*, 2025.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a Similarity Measure in Deep Learning. In *International Conference on Learning Representations*, 2023.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding Representation Similarity Through Statistical Testing. In *Advances in Neural Information Processing Systems*, 2021.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding Emergent Abilities of Language Models from the Loss Perspective. In *Advances in Neural Information Processing Systems*, 2024.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/s0895479895290954.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, and Surbhi Goel. The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains. In *Advances in Neural Information Processing Systems*, 2024.

Linnea Evanson, Yair Lakretz, and Jean Rémi King. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12205–12218. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.773.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep Learning Versus Kernel Learning: An Empirical Study of Loss Landscape Geometry and the Time Evolution of the Neural Tangent Kernel. In *Advances in Neural Information Processing Systems*, 2020.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the Downstream Performance of Pretrained Self-Supervised Representations by Their Rank. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974, 2023.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 3202–3211, 2019.

Sarah E. Harvey, Brett W. Larsen, and Alex H. Williams. Duality of Bures and Shape Distances with Implications for Comparing Neural Representations. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pp. 11–26, 2024.

Max Hennick and Guillaume Corlouer. From Density Matrices to Phase Transitions in Deep Learning: Spectral Early Warnings and Interpretability. *arXiv preprint arXiv:2603.29805*, 2026. URL `https://arxiv.org/abs/2603.29805`.

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.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

T. Kailath. An innovations approach to least-squares estimation–Part I: Linear filtering in additive white noise. *IEEE Transactions on Automatic Control*, 13(6):646–655, 1968. doi: 10.1109/tac.1968.1099025.

Sara Kangaslahti, Elan Rosenfeld, and Naomi Saphra. Hidden Breakthroughs in Language Model Training. In *International Conference on Learning Representations*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Christian Kuehn. A mathematical framework for critical transitions: Bifurcations, fast–slow systems and stochastic dynamics. *Physica D: Nonlinear Phenomena*, 240(12):1020–1035, 2011. doi: 10.1016/j.physd.2011.02.012.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the Transition from Lazy to Rich Training Dynamics. In *International Conference on Learning Representations*, 2024.

Richard D Lange, Devin Kwok, Jordan Kyle Matelsky, Xinyue Wang, David Rolnick, and Konrad Kording. Deep Networks as Paths on the Manifold of Neural Representations. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pp. 102–133, 2023.

Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The Local Learning Coefficient: A Singularity-Aware Complexity Measure. In *International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 244–252, 2025.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards Resolving the Implicit Bias of Gradient Descent for Matrix Factorization: Greedy Low-Rank Learning. In *International Conference on Learning Representations*, 2021.

Jing Liu, Haidong Yuan, Xiao-Ming Lu, and Xiaoguang Wang. Quantum Fisher information matrix and multiparameter estimation. *Journal of Physics A: Mathematical and Theoretical*, 53(2):023001, 2020. doi: 10.1088/1751-8121/ab5d4d.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards Understanding Grokking: An Effective Theory of Representation Learning. In *Advances in Neural Information Processing Systems*, 2022.

E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40 (12):837–842, 1934. doi: 10.1090/s0002-9904-1934-05978-0.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on Representational Similarity in Neural Networks with Canonical Correlation. In *Advances in Neural Information Processing Systems*, pp. 5727–5736, 2018.

Boris Muzellec and Marco Cuturi. Generalizing Point Embeddings Using the Wasserstein Space of Elliptical Distributions. In *Advances in Neural Information Processing Systems*, pp. 10237–10248, 2018.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress Measures for Grokking via Mechanistic Interpretability. In *International Conference on Learning Representations*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads. *arXiv preprint arXiv:2209.11895*, 2022. URL https://arxiv.org/abs/2209.11895.

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1-2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Scott Pesme and Nicolas Flammarion. Saddle-to-Saddle Dynamics in Diagonal Linear Networks. In *Advances in Neural Information Processing Systems*, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

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*, pp. 6076–6085, 2017.

Gautam Reddy. The Mechanistic Basis of Data Dependence and Abrupt Learning in an In-Context Classification Task. In *International Conference on Learning Representations*, 2024.

Olivier Roy and Martin Vetterli. The Effective Rank: A Measure of Effective Dimensionality. In *Proceedings of the 15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Naomi Saphra and Adam Lopez. Understanding learning dynamics of language models with SVCCA. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3257–3267, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1329. URL https://aclanthology.org/N19-1329/.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks. In *International Conference on Learning Representations*, 2014.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019. doi: 10.1073/pnas.1820226116.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? In *Advances in Neural Information Processing Systems*, 2023.

Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009. doi: 10.1038/nature08227.

James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the Stepwise Nature of Self-Supervised Learning. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31852–31876, 2023.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie C.Y. Chan, and Andrew M Saxe. What Needs to Go Right for an Induction Head? a Mechanistic Study of In-Context Learning Circuits and Their Formation. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45637–45662, 2024.

Shuai Tang, Wesley J. Maddox, Charlie Dickens, Tom Diethe, and Andreas Damianou. Similarity of Neural Networks with Gradients. *arXiv preprint arXiv:2003.11498*, 2020. URL `https://arxiv.org/abs/2003.11498`.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. doi: 10.1016/j.sigpro.2019.107299.

Armin Uhlmann. The transition probability in the state space of a *-algebra. *Reports on Mathematical Physics*, 9(2):273–279, 1976. doi: 10.1016/0034-4877(76)90060-4.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and Specialization of Attention Heads via the Refined Local Learning Coefficient. In *International Conference on Learning Representations*, 2025.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT*, 12(1): 99–111, 1972. doi: 10.1007/BF01932678.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized Shape Metrics on Neural Representations. In *Advances in Neural Information Processing Systems*, 2021.

Kayo Yin and Jacob Steinhardt. Which Attention Heads Matter for In-Context Learning? In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72428–72461, 2025.

STATEMENT ON GENERATIVE AI

Generative AI tools were used to assist with code development, literature review, figure design, and prose editing. The authors remain responsible for the derivations, source verification, empirical analyses, and final text.

REPRODUCIBILITY STATEMENT

The appendix records assumptions, proof details, experiment definitions and all relevant details for reproduction of the main results and findings of the work.

A RELATED WORK AND COMPARISON WITH OTHER METHODS

This section places the paper in the literature and compares the observer with its two closest relatives: the 2-datapoint reduced density matrix, analytically and on the same induction runs, and representation-similarity measures.

A.1 RELATED WORK

Emergence and hidden progress. Some capabilities of language models are absent in smaller models and present in larger ones (Wei et al., 2022). Part of this apparent sharpness can come from the choice of evaluation metric (Schaeffer et al., 2023), and downstream performance tracks pre-training loss across model and data sizes (Du et al., 2024). Over the course of training, abrupt transitions can rest on continuous progress that the loss does not show. Stochastic gradient descent gradually amplifies sparse parity solutions while loss and error stay flat (Barak et al., 2022); mechanistic progress measures underlie grokking (Power et al., 2022; Nanda et al., 2023); the abrupt acquisition of syntactic attention structure in masked language models precedes their later linguistic capabilities (Chen et al., 2024); and decomposing the loss reveals breakthroughs hidden in the aggregate curve (Kangaslahti et al., 2026). The observer developed here addresses the same gap between internal change and expression, but reads it from representation geometry on fixed probes.

Learning dynamics from small initialization. Exact solutions for deep linear networks show long plateaus followed by the sequential acquisition of modes (Saxe et al., 2014; 2019). Related analyses find learning of gradually increasing rank in linear networks and matrix factorization (Gidel et al., 2019; Li et al., 2021), saddle-to-saddle dynamics in diagonal networks (Pesme & Flammarion, 2023), stepwise learning of eigenmodes in self-supervised learning (Simon et al., 2023), and gradual rank increase in transformers (Boix-Adsera et al., 2023). The kernel can align early in training and then grow only in scale (Atanasov et al., 2022), the neural tangent kernel changes rapidly during an early transient (Fort et al., 2020), and the Fisher information of the weights rises early in training and then decreases (Achille et al., 2019). Grokking has been analyzed through an effective theory of representation learning (Liu et al., 2022) and as a transition from lazy to rich training dynamics (Kumar et al., 2024). The response-order law of Section 3 asks a different question of such staged processes: which observations of an emerging response move first.

Developmental interpretability and in-context learning. The local learning coefficient is a loss-landscape complexity measure from singular learning theory (Lau et al., 2025). It tracks stagewise development in transformers (Hoogland et al., 2025), and refined variants follow the differentiation and specialization of individual attention heads (Wang et al., 2025). In-context learning emerges abruptly through induction heads (Olsson et al., 2022; Singh et al., 2024), passes through distinct phases on Markov-chain data (Edelman et al., 2024), and can be preceded by measurable progress (Reddy, 2024). Ablation studies separate the roles of induction heads and function-vector heads (Yin & Steinhardt, 2025). These approaches measure the loss landscape or identified circuits; the observer here measures the representation geometry of fixed probes and requires no circuit to be identified in advance.

Early warning and change detection. Critical slowing down predicts transitions in dynamical systems from rising fluctuations (Scheffer et al., 2009; Kuehn, 2011), and spectral statistics of per-sample loss correlations bring this idea to training (Hennick & Corlouer, 2026); Appendices A.2 and A.3 relate and compare the two constructions. Sequential detectors threshold a running statistic online (Page, 1954), offline change-point methods segment a completed series (Truong et al., 2020), and innovations measure the part of an observation not predicted from its past (Kailath, 1968). The watch and alarm rule of Section 4 is a causal threshold of this kind applied to a geometric departure score.

Representation geometry. Beyond the similarity measures discussed in Appendix A.4, geometric analyses of trained language models find a phase of high intrinsic dimension in which representations first carry a full abstraction of the input (Cheng et al., 2025). Spectral summaries such as effective rank (Roy & Vetterli, 2007; Garrido et al., 2023) describe how widely a representation spreads its energy; the allocation speed of Section 2 measures how fast that spread changes.

A.2 Relationship to the 2-datapoint reduced density matrix

The 2-datapoint reduced density matrix (2RDM) of Hennick & Corlouer (2026) and the response operator of Appendix B.4 are the same construction with different weights on parameter space. Four results make this precise. The two objects agree to first order when the parameter covariance is proportional to the optimizer operator (Lemma A.1). The normalized 2RDM is a normalized Gram state in loss-response space, so it keeps allocation, direction, and support and drops only scale (Lemma A.2). Its motion therefore obeys the four-motion decomposition and the response-order law (Lemma A.3). Over a short window it reduces to the direction of the loss velocity (Lemma A.4).

Setup. Work in the orthonormal observation coordinates of Appendix B. For an observation $c\colon \Theta \to \mathbb{R}^N$ with Jacobian $J_c = Dc_\theta$ and a positive-semidefinite operator B on parameter space, define the covariance-response pushforward

$$\mathcal{R}_c[B] := J_c B J_c^*, \tag{24}$$

where $J_c^* = J_c^\top$ in these coordinates. The response operator is $\mathcal{M}_c = \mathcal{R}_c[A_\theta]$. The dynamical 2RDM uses instead the covariance of parameter displacements within a training window, observed through the per-probe losses $\ell(\theta) = (\ell_1(\theta), \dots, \ell_N(\theta))^\top$.

Lemma A.1 (First-order equivalence)**.** Let $\theta_\varepsilon = \bar\theta + \varepsilon\xi$ with $\mathbb{E}[\xi] = 0$ and $\mathrm{Cov}(\xi) = B$, and suppose c is differentiable at $\bar\theta$ in the L^2 sense,

$$c(\bar\theta + \varepsilon\xi) = c(\bar\theta) + \varepsilon J_c \xi + r_\varepsilon, \qquad \|r_\varepsilon\|_{L^2} = o(\varepsilon). \tag{25}$$

Then $\mathrm{Cov}[c(\theta_\varepsilon)] = \varepsilon^2 \mathcal{R}_c[B] + o(\varepsilon^2)$. For the loss observation this gives the local 2RDM $C_\varepsilon^{(\ell)} = \varepsilon^2 J_\ell B J_\ell^* + o(\varepsilon^2)$. If $B = \gamma A_\theta$ with $\gamma > 0$ and $\mathcal{M}_\ell \neq 0$, then

$$\frac{C_\varepsilon^{(\ell)}}{\operatorname{tr} C_\varepsilon^{(\ell)}} = \frac{\mathcal{M}_\ell}{\operatorname{tr} \mathcal{M}_\ell} + o(1). \tag{26}$$

Proof. Centering Eq. (25) gives $c(\theta_\varepsilon) - \mathbb{E}[c(\theta_\varepsilon)] = \varepsilon J_c \xi + \tilde r_\varepsilon$ with $\tilde r_\varepsilon = r_\varepsilon - \mathbb{E}[r_\varepsilon]$. The covariance is therefore $\varepsilon^2 J_c B J_c^*$ plus two cross terms and $\mathbb{E}[\tilde r_\varepsilon \tilde r_\varepsilon^*]$. By Cauchy–Schwarz and $\|\tilde r_\varepsilon\|_{L^2} = o(\varepsilon)$, all three are $o(\varepsilon^2)$. Taking $c = \ell$ and $B = \gamma A_\theta$ gives $\mathcal{R}_\ell[B] = \gamma \mathcal{M}_\ell$, and trace normalization is continuous on nonzero positive-semidefinite matrices. □

The structural relation does not need the proportionality: without it, the 2RDM and the response operator are two positive-semidefinite forms pushed through the same Jacobian. The remaining results hold for any $B \succeq 0$. Write $C = \mathcal{R}_\ell[B] \neq 0$, $a_C = \operatorname{tr} C$, $\widehat C = C/a_C$, and $R = J_\ell B^{1/2}$, so that $C = RR^*$.

Lemma A.2 (What the normalized 2RDM retains)**.** The normalized 2RDM is the normalized Gram state of the covariance-weighted loss-response matrix, $\widehat C = RR^* / \|R\|_{\mathrm{HS}}^2$. For a unit probe direction u, the quantity $u^* \widehat C u = \left\|B^{1/2} J_\ell^* u\right\|^2 / \|R\|_{\mathrm{HS}}^2$ is the fraction of linearized loss-response variance along u. If $R = U \operatorname{diag}(s_1, \dots, s_r) V_R^*$ is a compact singular value decomposition, then

$$\widehat C = \sum_{i=1}^r p_i u_i u_i^*, \qquad p_i = \frac{s_i^2}{\sum_j s_j^2}, \qquad P_C = CC^+ = \sum_{i=1}^r u_i u_i^*. \tag{27}$$

The normalized 2RDM therefore keeps the response directions u_i, their relative strengths p_i, and the response support, and discards only the total scale a_C.

Proof. $C = J_\ell B J_\ell^* = RR^*$ and $\operatorname{tr} C = \|R\|_{\mathrm{HS}}^2$. For any u, $u^*Cu = \left\|B^{1/2}J_\ell^* u\right\|^2$. The spectral form follows by substituting the singular value decomposition. □

Lemma A.3 (Motion of the normalized 2RDM)**.** Along a C^1 constant-rank path $C(t)$ with four-motion decomposition $\dot{C} = \dot{C}_{\mathrm{grow}} + \dot{C}_{\mathrm{alloc}} + \dot{C}_{\mathrm{rot}} + \dot{C}_{\mathrm{trans}}$, where $\dot{C}_{\mathrm{grow}} = (\dot{a}_C/a_C)C$,

$$\dot{\widehat{C}} = \frac{1}{a_C}\big(\dot{C}_{\mathrm{alloc}} + \dot{C}_{\mathrm{rot}} + \dot{C}_{\mathrm{trans}}\big), \tag{28}$$

so normalization removes pure growth exactly and keeps the other three motions. If moreover $r_C = \sqrt{\operatorname{tr} C}$ and $\dot{C} = r_C^{p+1}V + o(r_C^{p+1})$ with $p \geq 1$, then

$$\dot{\widehat{C}} = r_C^{p-1}\big[V - (\operatorname{tr} V)\widehat{C}\big] + o(r_C^{p-1}), \tag{29}$$

the same scale-free response-order law as for a representation Gram state (Eq. (7)).

Proof. Differentiating $\widehat{C} = C/a_C$ gives $\dot{\widehat{C}} = a_C^{-1}\big(\dot{C} - (\dot{a}_C/a_C)C\big)$. The subtracted term is $\dot{C}_{\mathrm{grow}}$, and the other three components are trace-free, which gives Eq. (28). For the second statement, $a_C = r_C^2$ gives $\dot{r}_C = \operatorname{tr}\dot{C}/(2r_C) = \frac{1}{2}r_C^p \operatorname{tr} V + o(r_C^p)$. Substituting this into the derivative gives Eq. (29). □

Lemma A.4 (Small-window limit)**.** Let $\ell(t)$ be a C^3 loss trajectory, and let $C_W(t)$ be the covariance of $\ell(t+s)$ for s uniform on $[-W/2, W/2]$, the continuous-time 2RDM over a symmetric window of width W. Then

$$C_W(t) = \frac{W^2}{12}\dot{\ell}(t)\dot{\ell}(t)^* + O(W^4), \tag{30}$$

and if $\dot{\ell}(t) \neq 0$ the normalized 2RDM is the projector onto the loss velocity up to $O(W^2)$: $C_W/\operatorname{tr} C_W = \dot{\ell}\dot{\ell}^*/\left\|\dot{\ell}\right\|^2 + O(W^2)$.

Proof. Write $v = \dot{\ell}(t)$, $q = \ddot{\ell}(t)$, and $\overline{\ell}_W(t)$ for the window mean. Taylor expansion gives $\ell(t+s) - \overline{\ell}_W(t) = sv + \frac{1}{2}(s^2 - W^2/12)q + O(|s|^3)$ uniformly on the window. The odd moments of s vanish and its second moment is $W^2/12$, so the cross term between the linear and quadratic parts averages to zero, the linear part contributes $(W^2/12)vv^*$, and everything else is $O(W^4)$. Dividing by the trace, whose leading value is $(W^2/12)\|v\|^2$, gives the normalized statement. □

Scalar statistics. Since $C = a_C\widehat{C}$, the participation ratio and the spectral heat capacity satisfy $\mathrm{PR}(C) = 1/\operatorname{tr}(\widehat{C}^2)$ and $\mathrm{SHC}(C) = a_C^2\,\mathrm{SHC}(\widehat{C})$. With spectral heat capacity taken as the population variance of the N eigenvalues, and using $\operatorname{tr}\widehat{C} = 1$,

$$\mathrm{SHC}(\widehat{C}) = \frac{1}{N}\operatorname{tr}(\widehat{C}^2) - \frac{1}{N^2} = \frac{1}{N\,\mathrm{PR}(C)} - \frac{1}{N^2}. \tag{31}$$

After normalization the two scalars carry the same spectral information. Eigenvector motion and support motion are visible only in the full path $\widehat{C}(t)$.

Interpretation. When $B = A_\theta$, write $\mathcal{M}_\ell = \sum_i \lambda_i u_i u_i^*$. The response equation gives

$$\dot{\ell} = -\mathcal{M}_\ell\phi_\ell = -\sum_i \lambda_i\langle u_i, \phi_\ell\rangle u_i. \tag{32}$$

The normalized 2RDM identifies the accessible loss-response directions u_i and their relative responsiveness $\lambda_i/\sum_j \lambda_j$, while the drive ϕ_ℓ selects which combination of them is realized. In the rank-one case every nonzero loss response is parallel to its single eigenvector. The 2RDM is therefore a geometric state in loss-response space, not the representation Gram state itself: representation changes annihilated by the loss Jacobian are invisible to it. Subject to that bottleneck, it retains exactly the scale-free allocation, direction, and support of the covariance-weighted loss-response map.

A.3 Experimental comparison with the spectral heat capacity

Protocol. We retrain 28 transition runs of the two-layer transformer of Sec. 5.2 (2,000 steps; batches half repeated $[x, x]$ and half random), 8 no-repetition controls trained on random sequences only, and 4 slow controls with the learning rate divided by ten. Every five steps we record the final residual stream averaged over the positions of 256 fixed random prefixes, the input of the blind observer of Appendix G, and the per-sample losses on the 512 repeated sequences used by Hennick & Corlouer (2026) and on the 256 prefixes. The geometric score s_t^{ind} is computed as in Appendix G. The 2RDM uses checkpoints every ten steps and $W = 20$, the published induction configuration. Every score receives the same causal rule: the standardization applied to the score, threshold 4.5 for two consecutive checkpoints, a 16-observation refractory period, at most three alarms, and no alarms during optimizer warmup. The offline induction event is the first checkpoint after which the mean induction-attention score of the two strongest heads stays above 0.5 (median step 990, range 945–1,145). The first 4 transition runs calibrate one alarm-to-event latency per score and one fixed step; both are scored on the remaining 24 runs. Leads are measured from the first post-warmup alarm.

Results. Figure 5 and Table 2 summarize the comparison. The blind path break warns before induction in all 28 runs with median lead 330 steps (interquartile range 292–362) and is silent in all 12 controls. Under the identical rule, the heat capacity on the repeated bank warns in all 28 runs with median lead 202 steps (195–221), reaches its largest pre-event value 95 steps before the event (range 70–155), and rises again after the event in every run as the loss covariance reorganizes while the circuit settles. It also alarms in 6 of 12 controls, at steps 1,540–1,830, where the heat capacity of a model that learns nothing drifts as the learning rate decays; applying the continuation-departure rule of Eq. (99) to the heat-capacity path instead gives median lead 218 steps and alarms in 9 of 12 controls. On the repetition-free prefixes seen by the geometric observer, the heat capacity leads by only 45 steps (25 of 28 runs), and the loss-vector speed on the repeated bank, an amplitude quantity, leads by 45 steps. The order is therefore representation shape, then loss-fluctuation shape, then loss amplitude. The heat-capacity alarm is later but more tightly locked to expression: event-locked forecasts have median absolute error 15 steps from its alarm against 38 from the geometric alarm and 38 from a fixed step, and the geometric alarm nearest the event has median lead 242 steps. Under the same rule the raw geometric speed warns in 27 of 28 runs with median lead 270 steps and no control alarms, so on this task the acceleration of the representation is itself an early signal, about 60 steps after the path break. The eigenvalue-variance form of the heat capacity alarms with median lead 205 steps and no control alarms, but its largest pre-event value falls at the event (median 8 steps before it), because it is dominated by loss amplitude. Evaluating the geometric score at the 2RDM cadence gives median lead 280 steps; evaluating the heat capacity at the geometric cadence with $W = 20$ or $W = 40$ gives 195 and 215 steps; moving the event threshold to 0.25 or 0.75 moves the geometric median lead to 318 and 345 steps and the heat-capacity median lead to 190 and 220 steps. The rolling z-score detector released with the 2RDM code has no persistence rule and flags 63 percent of all post-warmup checkpoints on transition runs and controls alike, so it is not evaluated as a detector.

What the comparison shows. The 2RDM summaries measure how loss fluctuations are allocated among probe directions; the path-break score measures whether the normalized representation path departs from its own continuation. On the same runs the representation departs first, and most of the heat capacity's lead comes from probes that already contain the repetition the model is about to exploit. The comparison is confined to the induction task and to the alarm rule of this paper; the 2RDM summaries apply wherever per-sample losses exist, including settings in which no representation is exposed.

A.4 Relationship to representation similarity measures

Representation-similarity measures compare two representations evaluated on the same probes. The observer developed here instead treats one representation as a path and asks how its geometry is changing at the current checkpoint. The closest connection is through the fixed-probe quotient geometry. Alignment methods such as CCA and generalized shape distances act on feature matrices and optimize over a map between feature coordinates, whereas summary methods such as RSA, CKA, and normalized Bures similarity compare probe-by-probe summaries that are already invariant to the declared feature symmetries (Kriegeskorte et al., 2008; Kornblith et al., 2019; Harvey et al., 2024). Once the alignment is optimized out, shape distances also have Gram-matrix expressions. Below we

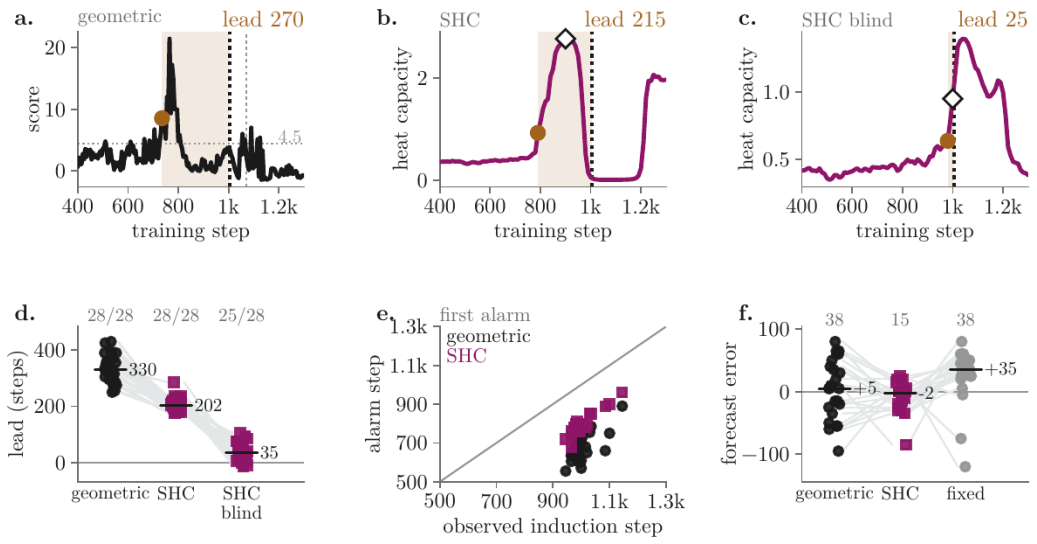

Figure 5: **The blind path break and the spectral heat capacity on the same runs. a,** Representative evaluation run (seed 4): the blind path-break score crosses its threshold at the gold dot, 270 steps before the induction event (dashed ink); the tint spans alarm to event and the dashed gray vertical is the event-clock forecast. **b,** The spectral heat capacity of the windowed loss covariance on the 512 repeated probe sequences of Hennick & Corlouer (2026), with the same alarm rule (gold dot, lead 215) and its largest pre-event value (open diamond, lead 105). **c,** The same statistic on the repetition-free prefix bank seen by the geometric observer (lead 25). **d,** Lead from the first post-warmup alarm to the event across the 28 transition runs; gray lines pair the same run, bars are medians, and counts above each column give runs alarmed before the event. **e,** First alarm step against the observed induction step; the line is the identity. **f,** Event-locked forecast errors on the 24 evaluation runs for latencies calibrated on the 4 development runs, and the fixed-step baseline; counts above each column give median absolute errors.

Table 2: Detection and forecasting on 28 retrained transition runs and 12 controls. Leads run from the first post-warmup alarm to the induction event; the median and interquartile range are over runs with an alarm. Forecasts use one latency calibrated on 4 development runs and are scored on 24 evaluation runs.

Score and rule	Before event	Lead (IQR)	Controls	Forecast MAE
Geometric path break, blind prefixes	28/28	330 (292–362)	0/12	38
checkpoint cadence 10	28/28	280 (265–318)	0/12	30
Raw geometric speed	27/28	270 (245–298)	0/12	25
SHC, repeated bank	28/28	202 (195–221)	6/12	15
continuation-departure rule	28/28	218 (205–238)	9/12	25
eigenvalue-variance SHC	28/28	205 (190–225)	0/12	15
Participation ratio, repeated bank	28/28	195 (185–205)	5/12	15
SHC, blind prefixes	25/28	45 (20–60)	5/12	38
Loss-vector speed, repeated bank	28/28	45 (40–50)	0/12	5
Fixed training step	–	–	–	38

record that correspondence, show why the weightings differ near entry of new support, and separate our causal path statistics from prior uses of similarity along training.

A.4.1 Finite distances and the Bures–shape duality

For direct comparison with standard linear CKA and the formulation of Harvey et al. (2024), assume in this subsection that the representations have been centered over probes. Let $H, G \in \mathbb{R}^{N\times d}$, with

centered linear Gram matrices

$$K = HH^\top, \qquad L = GG^\top, \qquad \rho = \frac{K}{\operatorname{tr} K}, \qquad \sigma = \frac{L}{\operatorname{tr} L}. \tag{33}$$

The two summary similarities used below are

$$\mathrm{CKA}(K, L) = \frac{\langle K, L\rangle_{\mathrm{F}}}{\|K\|_{\mathrm{F}}\|L\|_{\mathrm{F}}}, \qquad \mathrm{NBS}(K, L) = \frac{F(K, L)}{\sqrt{\operatorname{tr} K \operatorname{tr} L}} = F(\rho, \sigma), \tag{34}$$

where

$$F(K, L) := \operatorname{tr}\left[\left(K^{1/2} L K^{1/2}\right)^{1/2}\right] \tag{35}$$

is the fidelity. Linear CKA is the cosine between the centered Gram matrices in Frobenius space (Kornblith et al., 2019); NBS is the normalized fidelity used in the Bures–Wasserstein literature (Muzellec & Cuturi, 2018; Tang et al., 2020).

The duality of Harvey et al. (2024) identifies these summary and alignment views. For an orthogonal alignment $Q \in O(d)$,

$$\min_{Q^\top Q = I} \|H - GQ\|_{\mathrm{F}}^2 = \operatorname{tr} K + \operatorname{tr} L - 2F(K, L) = d_{\mathrm{BW}}^2(K, L), \tag{36}$$

so the squared Euclidean shape distance is the squared Bures–Wasserstein distance. They also show that

$$\mathrm{NBS}(K, L) = \cos\theta_{\mathrm{shape}}(H, G), \tag{37}$$

where θ_{shape} is the Riemannian shape distance obtained after optimal feature alignment. Generalized shape metrics place this construction in a broader family of metric representation spaces (Williams et al., 2021).

This duality locates the finite geometry used in this paper. The distance derived in Appendix B is the product metric

$$d_g^2(K, L) = \log^2\left(\frac{\operatorname{tr} L}{\operatorname{tr} K}\right) + 4\arccos^2 F(\rho, \sigma). \tag{38}$$

On a fixed-trace slice, $d_g = 2\arccos \mathrm{NBS}$. Infinitesimally, the quotient metric is four times the Bures metric divided by the current trace. The normalization therefore retains growth as an orthogonal log-scale coordinate rather than discarding it as a nuisance transformation.

CKA and NBS both discard this scale coordinate and depend only on normalized shape, but they place different weights on that shape. If ρ and σ commute, with eigenvalues p_i and q_i in a common basis, then

$$\mathrm{CKA}(\rho, \sigma) = \frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2}\sqrt{\sum_i q_i^2}}, \qquad F(\rho, \sigma) = \sum_i \sqrt{p_i q_i}. \tag{39}$$

More generally, CKA is a normalized Hilbert–Schmidt inner product while NBS is a normalized fidelity, equivalently a nuclear-norm quantity on the cross-covariance (Harvey et al., 2024). The finite normalized-shape factor is therefore inherited from the representation-similarity literature.

The difference in weighting matters for weak emerging responses. Ding et al. (2021) show that CKA can be dominated by the largest singular values and miss meaningful differences among smaller ones, where orthogonal Procrustes distance behaves more consistently, and Davari et al. (2023) show that CKA can change substantially under transformations that roughly preserve function. Here the relevant weak motion is often entry into a previously unrepresented direction, for which the Bures and Frobenius weightings differ in response order.

A.4.2 Similarity along training

Similarity measures have long been used to describe representation development. Raghu et al. (2017) compared layers across checkpoints with their final representations and found bottom-up convergence; Morcos et al. (2018) found a similar pattern in recurrent language models with projection-weighted CCA; and Saphra & Lopez (2019) used SVCCA to follow the emergence of syntactic, semantic, and topical structure during language-model training.

The closest geometric precedent is Lange et al. (2023), who treat the ordered layers of a fixed network as a path through representation space and use metric versions of CKA and generalized shape distance to study its step lengths, geodesics, tangents, and interior angles. Their path parameter is depth rather than training time, but the construction shows that representation metrics support path-level quantities, not only endpoint comparisons. Our approach differs in the reference state, in the information the scalar retains, and in what counts as an event.

Reference state. Similarity to the final checkpoint measures displacement toward a state an online observer cannot yet see, so it suits retrospective analysis but is anticausal. Our displacement, path length, sampled speed, and innovation use only checkpoints already observed. Consecutive-checkpoint similarity is causal, and it becomes a sampled speed when it comes from a length metric: raw CKA is not a distance, but angular CKA and the shape metrics of Williams et al. (2021) are.

What the scalar retains. Every similarity measure quotients out some changes and keeps others (Table 3). Plain CCA correlations are cosines of principal angles between the two represented probe subspaces, so CCA sees relative support only and is blind to growth, allocation, and internal rotation. SVCCA truncates each spectrum before CCA, so allocation enters through which subspace is kept; PWCCA reweights the canonical correlations and loses the full invariances of CCA (Kornblith et al., 2019). CKA and NBS keep allocation, internal rotation, and support transport, with the weightings above, but discard growth. Euclidean shape distance also keeps growth but merges all four motions into one endpoint distance; d_g keeps the same information while separating growth from shape, and its tangent decomposition separates the three shape motions. RSA is a family whose invariances depend on the RDM dissimilarity and the comparison statistic (Kriegeskorte et al., 2008; Harvey et al., 2024).

Measure	Invariance and effective object	Information retained
CCA mean correlation	Invertible maps of each feature space; principal angles between represented supports	Relative support only; discards growth, allocation, and internal rotation
SVCCA	PCA truncation followed by CCA; invertible invariance only if the retained subspaces are unchanged	Relative position of selected supports; allocation enters through truncation
PWCCA	Weighted canonical correlations; isotropic-scale invariant, but not generally orthogonally or invertibly invariant	Support relation with coordinate-dependent importance weighting
Linear CKA	Orthogonal feature maps and independent isotropic rescaling; Frobenius geometry on normalized Gram states	Allocation, internal rotation, and support transport; discards growth
NBS / angular shape	Orthogonal feature maps and independent isotropic rescaling; fidelity/Bures geometry on normalized Gram states	Allocation, internal rotation, and support transport; discards growth
Euclidean shape	Orthogonal feature maps; unnormalized Bures geometry on Gram states	All four motions, combined in one endpoint distance
This work, d_g	Orthogonal feature maps and common rescaling of activation units; product geometry on $(\log \mathrm{tr}\, K, \rho)$	All four motions, separated along the training path

Table 3: Representation-similarity measures as partial views of the Gram path. RSA is omitted as a separate row because it denotes a family whose invariances depend on both stages of the analysis; the linear-Gram, scale-normalized case is represented by the CKA row.

The distinction matters for convergence. Saturation of a similarity-to-final curve shows only that the retained information has stabilized, not that the representation has stopped moving. In Pythia, later training is dominated by internal rotation, which plain CCA cannot see and which CKA and NBS see with weights that suppress rotations between nearly degenerate directions (CKA also emphasizes high-energy directions).

Events. The training studies above read convergence from a similarity curve; they do not define a causal null, a time-local threshold, or a persistence rule. Our score is distance from a declared no-change family, the largest unit-stable scalar that vanishes on it (Theorem D.1), and a persistent

causal crossing defines the event. The closest path statistic in Lange et al. (2023) is the interior angle θ between adjacent depth-wise steps; a straight path has $\theta = \pi$, so $\pi - \theta$ measures local turning. Innovation extends this from two steps to a continuation fitted to a window of earlier tangents, measured in the local metric and standardized against its own causal history. A new transverse component then enters innovation at first order but changes total speed only at second order (Eq. (86)). Conventional similarity measures can register a transition; the difference is that this observer keeps the motions relevant to the question, uses only information available at the current checkpoint, and turns departure from an explicit null into a persistent event.

In summary, the representation-similarity literature supplies the finite normalized-shape geometry used here. We add scale as an orthogonal coordinate, differentiate the training path, separate its tangent motions, select statistics by no-change families, and construct events causally. To our knowledge, prior representation-similarity work does not combine these with the response-order analysis that explains why a weak internal reorganization can become geometrically visible before behavioral expression.

B GEOMETRY AND PROOFS

This section proves the geometric results of Section 2: the completeness of the Gram matrix as an orthogonal invariant, the invariant metric and its finite distance, and the four-motion decomposition of Theorem 1. It ends with the response operator that links these motions to training. Throughout, probes are fixed over each comparison interval, width normalization is absorbed into H when immaterial, and $K = HH^\top \neq 0$.

B.1 THE GRAM MATRIX IS THE COMPLETE ORTHOGONAL INVARIANT

If $G = HR$ with $R^\top R = I$, then $GG^\top = HH^\top$. Conversely, choose compact SVDs $H = U\Sigma V_H^\top$ and $G = U\Sigma V_G^\top$, using common left singular vectors in repeated eigenspaces. Equality of the Gram matrices fixes U and Σ, and the map $V_H \mapsto V_G$ extends orthogonally to the ambient feature space, giving $G = HR$. Thus K removes exactly the orthogonal feature gauge. Under general invertible feature changes, only $P = HH^+ = KK^+$ remains invariant.

For targets y and ridge parameter $\lambda > 0$, the fixed-probe ridge predictions are

$$\widehat{y}_\lambda = K(K + \lambda I)^{-1} y. \tag{40}$$

Hence K retains the metric information available to an isotropically regularized linear readout.

B.2 THE INVARIANT METRIC AND FINITE DISTANCE

Let $\mathcal{S}_k$ denote the set of positive semidefinite $N \times N$ matrices of rank exactly k. It is a smooth manifold of dimension $Nk - k(k-1)/2$, and its tangent space at K is

$$T_K\mathcal{S}_k = \{\Xi = \Xi^\top : \; u^\top \Xi v = 0 \text{ for all } u, v \in \ker K\}, \tag{41}$$

the symmetric matrices with vanishing null–null block. Rank changes cross between strata $\mathcal{S}_k$ and are not tangent directions.

We now state the invariant-metric result precisely.

Lemma B.1 (Invariant metric on Gram states)**.** Let $H \in \mathbb{R}^{N\times d}$ with $K = HH^\top \neq 0$ and write $a = \operatorname{tr} K = \|H\|_{\mathrm{F}}^2$. Consider isotropic feature metrics of the form $g_H(Z,Z) = w(a)\,\|Z\|_{\mathrm{F}}^2$ for $Z \in \mathbb{R}^{N\times d}$, where $w : (0,\infty) \to (0,\infty)$.

(i) *Uniqueness.* Every such metric is invariant under feature-basis changes $(H,Z) \mapsto (HR, ZR)$, $R^\top R = I$. It is in addition invariant under common rescaling $(H,Z) \mapsto (cH, cZ)$ for all $c > 0$ if and only if $w(a) = C/a$ for a constant $C > 0$, and the normalization $g_H(\dot H, \dot H) = 1$ along $H(s) = e^{s/2}H(0)$ forces $C = 4$. Hence the unique such metric is

$$g_H(Z,Z) = \frac{4\,\|Z\|_{\mathrm{F}}^2}{\|H\|_{\mathrm{F}}^2}. \tag{42}$$

(ii) *Quotient.* Let K have rank k and let $\Xi \in T_K\mathcal{S}_k$. The set $\{Z : ZH^\top + HZ^\top = \Xi\}$ is nonempty, and the minimum

$$g_K(\Xi,\Xi) = \min_{Z:\, ZH^\top + HZ^\top = \Xi} \frac{4\,\|Z\|_{\mathrm{F}}^2}{\operatorname{tr} K} \tag{43}$$

is attained at exactly one Z, the *horizontal lift* $Z = BH$, where $B = B^\top$ solves the Lyapunov equation

$$\Xi = BK + KB. \tag{44}$$

In an orthonormal eigenbasis of K with eigenvalues $\lambda_1, \dots, \lambda_N$ and Ξ_{ij} the entries of Ξ in that basis,

$$g_K(\Xi,\Xi) = \frac{2}{\operatorname{tr} K} \sum_{i,j:\ \lambda_i+\lambda_j>0} \frac{\Xi_{ij}^2}{\lambda_i + \lambda_j}. \tag{45}$$

Consequently g_K depends on K alone, not on the factor H, and it is a positive-definite quadratic form on $T_K\mathcal{S}_k$ varying smoothly with K. It is four times the Bures metric divided by the current scale.

Proof. *(i)* Basis invariance holds because $\|ZR\|_{\mathrm{F}} = \|Z\|_{\mathrm{F}}$ and $\|HR\|_{\mathrm{F}} = \|H\|_{\mathrm{F}}$ for orthogonal R. Under $(H,Z) \mapsto (cH, cZ)$ the metric becomes $w(c^2a)\,c^2\,\|Z\|_{\mathrm{F}}^2$, so invariance for all Z is equivalent to

$$w(c^2a)\,c^2 = w(a) \qquad \text{for all } a, c > 0. \tag{46}$$

Setting $a = 1$ gives $w(c^2) = w(1)/c^2$, i.e. $w(a) = C/a$ with $C = w(1)$; conversely every $w(a) = C/a$ satisfies Eq. (46). Along $H(s) = e^{s/2}H(0)$ we have $\dot H = \frac{1}{2}H$, so $g_H(\dot H, \dot H) = (C/\|H\|_{\mathrm{F}}^2)\cdot\frac{1}{4}\|H\|_{\mathrm{F}}^2 = C/4$, and unit speed forces $C = 4$.

(ii) Write $\mathcal{L}_H(Z) = ZH^\top + HZ^\top$, the differential at H of $\pi(H) = HH^\top$, and let $\mathcal{V}_H = \ker \mathcal{L}_H$ be the vertical space.

Existence. Each $\mathcal{L}_H(Z)$ is symmetric, and its null–null block vanishes because $\ker K = \ker H^\top$, so the image of $\mathcal{L}_H$ lies in $T_K\mathcal{S}_k$. Conversely, given $\Xi \in T_K\mathcal{S}_k$, work in an orthonormal eigenbasis of K and set $B_{ij} = \Xi_{ij}/(\lambda_i + \lambda_j)$ when $\lambda_i + \lambda_j > 0$ and $B_{ij} = 0$ otherwise. Then $B = B^\top$, and $BK + KB = \Xi$ because both sides agree whenever $\lambda_i + \lambda_j > 0$ and vanish on the null–null block by Eq. (41). Hence $\mathcal{L}_H(BH) = BK + KB = \Xi$, and the constraint set in Eq. (43) is nonempty.

Minimality and uniqueness. For symmetric B and $Z_0 \in \mathcal{V}_H$,

$$\langle BH, Z_0\rangle_{\mathrm{F}} = \operatorname{tr}(BZ_0H^\top) = \tfrac{1}{2}\operatorname{tr}\big(B(Z_0H^\top + HZ_0^\top)\big) = 0.$$

Every Z in the constraint set satisfies $Z - BH \in \mathcal{V}_H$, so $\|Z\|_{\mathrm{F}}^2 = \|BH\|_{\mathrm{F}}^2 + \|Z - BH\|_{\mathrm{F}}^2$, and BH is the unique minimizer. (The symmetric B is determined only up to additions supported on $\ker K \times \ker K$, which do not change BH.)

Closed form. In the eigenbasis,

$$\|BH\|_{\mathrm{F}}^2 = \operatorname{tr}(BKB) = \sum_{i,j} B_{ij}^2\lambda_j = \tfrac{1}{2}\sum_{i,j} B_{ij}^2(\lambda_i + \lambda_j) = \tfrac{1}{2}\sum_{i,j:\ \lambda_i+\lambda_j>0} \frac{\Xi_{ij}^2}{\lambda_i + \lambda_j},$$

using the symmetry of B in the third step. Multiplying by $4/\operatorname{tr} K$ gives Eq. (45). The right-hand side involves only the spectral data of K, so g_K does not depend on the factor. It vanishes only if $\Xi_{ij} = 0$ whenever $\lambda_i + \lambda_j > 0$, which together with Eq. (41) forces $\Xi = 0$. Smooth dependence on K follows because Eq. (45) is the quotient of the smooth metric (42) by the fibers of the submersion π restricted to rank-k factors. Finally, the Bures metric at K is $\frac{1}{2}\sum_{i,j} \Xi_{ij}^2/(\lambda_i + \lambda_j)$, so g_K is $4/\operatorname{tr} K$ times the Bures metric. □

Finite distance. Writing $H = \sqrt{a}\,\widehat{H}$ with $\left\|\widehat{H}\right\|_{\mathrm{F}} = 1$ gives the product metric

$$\frac{4\,\|\mathrm{d}H\|_{\mathrm{F}}^2}{\|H\|_{\mathrm{F}}^2} = (\mathrm{d}\log a)^2 + 4\left\|\mathrm{d}\widehat{H}\right\|_{\mathrm{F}}^2. \tag{47}$$

Optimizing the right orthogonal gauge gives the fidelity angle. For $\rho = K/\operatorname{tr} K$, $\sigma = L/\operatorname{tr} L$, and $F(\rho,\sigma) = \operatorname{tr}(\rho^{1/2}\sigma\rho^{1/2})^{1/2}$, the induced geodesic distance is

$$d_g^2(K,L) = \log^2\frac{\operatorname{tr} L}{\operatorname{tr} K} + 4\arccos^2 F(\rho,\sigma). \tag{48}$$

The Bures chord $\sqrt{2-2F}$ has the same infinitesimal metric. We identify explicitly whether finite analyses use the chord or the geodesic angle.

B.3 The four motions

Fix $K \in \mathcal{S}_k$ with $a = \operatorname{tr} K$ and $\rho = K/a$. Let $\operatorname{range} K = E_1 \oplus \cdots \oplus E_M$ be the decomposition into the eigenspaces of the distinct positive eigenvalues of K, let $E_0 = \ker K$, and let Q_m be the orthogonal projector onto E_m for $m = 0, \dots, M$. Define four subspaces of $T_K\mathcal{S}_k$:

$$\begin{aligned} \mathcal{T}_{\mathrm{grow}} &= \{cK : c \in \mathbb{R}\}, & \mathcal{T}_{\mathrm{alloc}} &= \Big\{\Xi = \textstyle\sum_{m\ge 1} Q_m \Xi Q_m,\ \operatorname{tr}\Xi = 0\Big\}, \\ \mathcal{T}_{\mathrm{rot}} &= \Big\{\Xi = \textstyle\sum_{m\neq m'\ge 1} Q_m \Xi Q_{m'}\Big\}, & \mathcal{T}_{\mathrm{trans}} &= \big\{\Xi = Q_0\Xi(I-Q_0) + (I-Q_0)\Xi Q_0\big\}, \end{aligned} \tag{49}$$

where Ξ ranges over symmetric matrices. In words: growth is motion along K; allocation is trace-free motion that is block diagonal with respect to the eigenspaces of K; internal rotation couples distinct eigenspaces of the support; and support transport couples the support to its orthogonal complement.

The following is the formal version of Theorem 1.

Theorem B.1 (Canonical decomposition of fixed-rank Gram dynamics)**.** Let $K \in \mathcal{S}_k$ and let g_K be the metric of Lemma B.1, with $\|\Xi\|_K^2 = g_K(\Xi,\Xi)$.

(i) *Decomposition.* The four subspaces in Eq. (49) are mutually g_K-orthogonal and

$$T_K\mathcal{S}_k = \mathcal{T}_{\mathrm{grow}} \oplus \mathcal{T}_{\mathrm{alloc}} \oplus \mathcal{T}_{\mathrm{rot}} \oplus \mathcal{T}_{\mathrm{trans}}.$$

Hence every tangent splits uniquely as $\Xi = \Xi_{\mathrm{grow}} + \Xi_{\mathrm{alloc}} + \Xi_{\mathrm{rot}} + \Xi_{\mathrm{trans}}$ with $\Xi_B \in \mathcal{T}_B$ and $\|\Xi\|_K^2 = \sum_B \|\Xi_B\|_K^2$. The subspaces depend only on the spectral projectors of K, not on the factor H or on a basis chosen inside an eigenspace.

(ii) *Coordinate form.* Let $K(t)$ be a C^1 path in $\mathcal{S}_k$ with a C^1 spectral decomposition $\rho(t) = \sum_{i=1}^k p_i(t)u_i(t)u_i(t)^\top$, $p_i > 0$, $\sum_i p_i = 1$, completed by $\{u_\alpha\}_{\alpha>k}$ to a C^1 orthonormal basis of $\mathbb{R}^N$; such a decomposition exists whenever the positive spectrum is simple. Set $\omega_{ij} = u_i^\top \dot{u}_j$ for $i,j \le k$ and $\omega_{i\alpha} = u_\alpha^\top \dot{u}_i$ for $i \le k < \alpha$. Then

$$\begin{aligned} \left\|\dot{K}_{\mathrm{grow}}\right\|_K^2 &= \left(\frac{\dot{a}}{a}\right)^2, & \left\|\dot{K}_{\mathrm{alloc}}\right\|_K^2 &= \sum_{i=1}^k \frac{\dot{p}_i^2}{p_i}, \\ \left\|\dot{K}_{\mathrm{rot}}\right\|_K^2 &= 4\sum_{i<j}\frac{(p_i-p_j)^2}{p_i+p_j}\,\omega_{ij}^2, & \left\|\dot{K}_{\mathrm{trans}}\right\|_K^2 &= 4\sum_{i\le k<\alpha} p_i\omega_{i\alpha}^2. \end{aligned} \tag{50}$$

and the identity in Eq. (4) holds.

Proof. (i) Fix an orthonormal eigenbasis $\{u_i\}_{i=1}^N$ of K adapted to the eigenspaces $E_0, \dots, E_M$, and write $\Xi_{ij} = u_i^\top \Xi u_j$. Polarizing Eq. (45),

$$g_K(\Xi, \Psi) = \frac{2}{a} \sum_{i,j:\ \lambda_i+\lambda_j>0} \frac{\Xi_{ij}\Psi_{ij}}{\lambda_i + \lambda_j}. \tag{51}$$

This is an entrywise inner product whose weight depends only on the pair of eigenvalues (λ_i, λ_j). Two consequences follow. First, matrices supported on disjoint collections of blocks $E_m \times E_{m'}$ are g_K-orthogonal. Second, the value of g_K is unchanged by an orthonormal change of basis within any single eigenspace, so any statement phrased in terms of the projectors Q_m is basis independent.

Let $\Xi \in T_K\mathcal{S}_k$. Using $Q_0 \Xi Q_0 = 0$, write

$$\Xi = D + O + T, \qquad D = \sum_{m\geq 1} Q_m \Xi Q_m,$$

$$O = \sum_{m\neq m'\geq 1} Q_m \Xi Q_{m'}, \quad T = Q_0\Xi(I - Q_0) + (I - Q_0)\Xi Q_0.$$

Then $O \in \mathcal{T}_{\mathrm{rot}}$, $T \in \mathcal{T}_{\mathrm{trans}}$, and D, O, T are mutually orthogonal by the first consequence. Since $\operatorname{tr} O = \operatorname{tr} T = 0$, set $c = \operatorname{tr}\Xi/a = \operatorname{tr} D/a$ and split $D = cK + (D - cK)$. The matrix $D - cK$ is block diagonal with respect to $E_1, \dots, E_M$ and trace-free, so it lies in $\mathcal{T}_{\mathrm{alloc}}$; and because K is diagonal in the eigenbasis, Eq. (51) gives

$$g_K\big(K,\, D - cK\big) = \frac{2}{a} \sum_{i\leq k} \frac{\lambda_i\,(D - cK)_{ii}}{2\lambda_i} = \frac{1}{a}\operatorname{tr}(D - cK) = 0.$$

The same computation shows $\mathcal{T}_{\mathrm{grow}} \perp \mathcal{T}_{\mathrm{alloc}}$ in general, and $\mathcal{T}_{\mathrm{grow}}, \mathcal{T}_{\mathrm{alloc}}$ are orthogonal to $\mathcal{T}_{\mathrm{rot}}, \mathcal{T}_{\mathrm{trans}}$ by block support. Thus every tangent is a sum of four mutually orthogonal pieces, one from each subspace, and orthogonality makes the sum direct and the decomposition unique. All four subspaces are symmetric with vanishing null–null block, so they lie in $T_K\mathcal{S}_k$. Pythagoras gives $\|\Xi\|_K^2 = \sum_B \|\Xi_B\|_K^2$. When the positive spectrum is simple, $\mathcal{T}_{\mathrm{alloc}}$ consists of trace-free diagonal matrices on the support (dimension $k-1$), $\mathcal{T}_{\mathrm{rot}}$ of symmetric matrices with zero diagonal on the support ($k(k-1)/2$), and $\mathcal{T}_{\mathrm{trans}}$ of symmetric matrices coupling the k-dimensional support to its $(N-k)$-dimensional complement ($k(N-k)$); with the one-dimensional $\mathcal{T}_{\mathrm{grow}}$ these sum to $Nk - k(k-1)/2$.

(ii) Along the path, $\{u_i(t)\}$ is an eigenbasis of $K(t)$ adapted to its eigenspaces, with eigenvalues $\lambda_i = ap_i$ for $i \leq k$ and $\lambda_\alpha = 0$. Differentiating $K = a\rho$ and using orthonormality ($u_i^\top \dot u_j = -\dot u_i^\top u_j$), the entries of $\dot K$ in the moving basis are

$$\dot K_{ii} = \dot\lambda_i = \dot a\, p_i + a\dot p_i, \quad \dot K_{ij} = a(p_j - p_i)\,\omega_{ij} \quad (i \neq j \leq k),$$
$$\dot K_{i\alpha} = \dot K_{\alpha i} = a\, p_i\, \omega_{i\alpha}, \quad \dot K_{\alpha\beta} = 0 \qquad (i \leq k < \alpha,\ \alpha, \beta > k).$$

Entries (i, j) with $i \neq j$ and $p_i = p_j$ vanish, so the block-diagonal part D of $\dot K$ is the diagonal matrix with entries $\dot\lambda_i$ on the support, and $c = \operatorname{tr}\dot K/a = \dot a/a$. The four components are therefore

$$\dot K_{\mathrm{grow}} = \frac{\dot a}{a} K, \qquad (\dot K_{\mathrm{alloc}})_{ii} = \dot\lambda_i - \frac{\dot a}{a}\lambda_i = a\dot p_i,$$
$$(\dot K_{\mathrm{rot}})_{ij} = a(p_j - p_i)\omega_{ij}, \quad (\dot K_{\mathrm{trans}})_{i\alpha} = ap_i\omega_{i\alpha}.$$

Substitution into Eq. (45) now gives the four terms directly:

$$\left\|\dot K_{\mathrm{grow}}\right\|_K^2 = \frac{2}{a}\sum_i \frac{(\dot a/a)^2\lambda_i^2}{2\lambda_i} = \left(\frac{\dot a}{a}\right)^2,$$

$$\left\|\dot K_{\mathrm{alloc}}\right\|_K^2 = \frac{2}{a}\sum_i \frac{a^2\dot p_i^2}{2ap_i} = \sum_i \frac{\dot p_i^2}{p_i},$$

$$\left\|\dot K_{\mathrm{rot}}\right\|_K^2 = \frac{4}{a}\sum_{i<j} \frac{a^2(p_i - p_j)^2\omega_{ij}^2}{a(p_i + p_j)} = 4\sum_{i<j} \frac{(p_i - p_j)^2}{p_i + p_j}\omega_{ij}^2,$$

$$\left\|\dot K_{\mathrm{trans}}\right\|_K^2 = \frac{4}{a}\sum_{i\leq k<\alpha} \frac{a^2p_i^2\omega_{i\alpha}^2}{ap_i} = 4\sum_{i\leq k<\alpha} p_i\omega_{i\alpha}^2.$$

The factors of four in the last two lines arise because each unordered pair appears twice in the symmetric matrix. Summing the terms and applying (i) gives Eq. (4). Equivalently, the diagonal contribution is

$$\sum_i \frac{\dot{\lambda}_i^2}{a^2 p_i} = \left(\frac{\dot{a}}{a}\right)^2 + \sum_i \frac{\dot{p}_i^2}{p_i}, \tag{52}$$

using $\sum_i p_i = 1$ and $\sum_i \dot{p}_i = 0$, which is the growth-plus-allocation part of the identity. Rotations within a repeated-eigenvalue block ($p_i = p_j$) are unobservable, consistent with the vanishing coefficient $(p_i - p_j)^2$; splitting such a block is allocation. □

B.4 Training-induced observation motion and response operator

Let $c(\theta)$ be any differentiable observation, with Jacobian $J_c = D_\theta c$, and let $A_\theta \succeq 0$ describe the local optimizer action. Under gradient flow,

$$\dot{\theta} = -A_\theta \nabla_\theta L, \qquad \dot{c} = -J_c A_\theta \nabla_\theta L, \qquad \mathcal{M}_c = J_c A_\theta J_c^*. \tag{53}$$

Use orthonormal coordinates for the declared observation metric, and let $\Pi_B^{g_K}$ project onto a block in Eq. (4).

Since the ranges of $J_c A_\theta$, $J_c A_\theta^{1/2}$, and $J_c A_\theta J_c^*$ agree, $J_c A_\theta \nabla L \in \operatorname{range} \mathcal{M}_c$ and

$$\mathcal{M}_c \mathcal{M}_c^\dagger J_c A_\theta \nabla L = J_c A_\theta \nabla L. \tag{54}$$

The minimum-norm observation-space drive is

$$\phi_c = \mathcal{M}_c^\dagger J_c A_\theta \nabla L, \qquad \dot{c} = -\mathcal{M}_c \phi_c. \tag{55}$$

It is useful for attributing motion to responsiveness and drive, but is not required to monitor $\dot{c}$.

For one Gram-geometric component, define

$$f_\theta = A_\theta^{1/2} \nabla_\theta L, \qquad T_B = \Pi_B^{g_K} J_K A_\theta^{1/2}, \qquad \dot{K}_B = \Pi_B^{g_K} \dot{K}.$$

Then

$$\dot{K}_B = -T_B f_\theta. \tag{56}$$

Define

$$m_B = \|T_B\|_{\mathrm{HS}}^2 = \operatorname{tr}(\Pi_B^{g_K} \mathcal{M}_K \Pi_B^{g_K})$$

and $\|\mathrm{d}L\|_{A_\theta} = \|f_\theta\|$. Since $\|T_B\|_{\mathrm{op}} \le \|T_B\|_{\mathrm{HS}}$, $\left\|\dot{K}_B\right\|_K \le \sqrt{m_B}\, \|\mathrm{d}L\|_{A_\theta}$. For a nonzero denominator, define $\alpha_B = \left\|\dot{K}_B\right\|_K / (\sqrt{m_B}\, \|\mathrm{d}L\|_{A_\theta})$, and set $\alpha_B = 0$ otherwise. Then $0 \le \alpha_B \le 1$ and

$$\left\|\dot{K}_B\right\|_K = \sqrt{m_B}\ \|\mathrm{d}L\|_{A_\theta}\ \alpha_B. \tag{57}$$

The factors are local responsiveness, optimizer-weighted force, and their directional match. Computing the realized velocity requires one Jacobian–vector product; stochastic estimates of m_B require vector–Jacobian products but no dense Jacobian. Optimizers with memory require an augmented state or their actual finite-step response.

C Response order and visibility

This section proves the response-order law of Theorem 2 and the crossing-time comparison used in Section 3, then states when a weak response becomes statistically visible at a given checkpoint cadence.

C.1 The response-order law

Let $\mathcal{H} = \mathbb{R}^{N \times d}$ with the Frobenius inner product and let $\mathbb{S} = \{U \in \mathcal{H} : \|U\|_{\mathrm{F}} = 1\}$. For a nonzero feature response $H \in \mathcal{H}$, define

$$r = \|H\|_{\mathrm{F}}, \qquad U = H/r, \qquad K = HH^\top, \qquad \rho = K/\operatorname{tr} K = UU^\top.$$

The limit studied in the main text is a family with $H_\varepsilon(0) = \varepsilon U_0$, or equivalently $K_\varepsilon(0) = \varepsilon^2 \rho_0$, as $\varepsilon \downarrow 0$. The response remains nonzero on the interval under study, so U and ρ are defined. The following hypotheses make the local expansion and its remainders precise.

(A1) *Local dynamics.* $H(t)$ is a C^1 path in $\mathcal{H} \setminus \{0\}$ with $\dot{H} = F_p(H) + G(H)$, where $F_p : \mathcal{H} \setminus \{0\} \to \mathcal{H}$ is C^1 and positively homogeneous of degree $p \geq 1$, $F_p(\lambda H) = \lambda^p F_p(H)$ for $\lambda > 0$, and $\|G(H)\|_{\mathrm{F}} \leq C_G \|H\|_{\mathrm{F}}^{p+\delta}$ for some $\delta > 0$, $C_G < \infty$, and all $0 < \|H\|_{\mathrm{F}} \leq r_0$.

(A2) *Gram measurement.* The scalar c is C^1 and, for $0 < r \leq r_0$ and $U \in \mathbb{S}$, $c(rU) = r^\nu \phi(UU^\top) + \bar{c}(rU)$ with $\nu \geq 0$, where ϕ is C^1 on a neighborhood of the normalized Gram path. The remainder obeys $|\bar{c}(H)| \leq C_c \|H\|_{\mathrm{F}}^{\nu+\eta}$ and $\|D\bar{c}(H)\| \leq C_c \|H\|_{\mathrm{F}}^{\nu+\eta-1}$ for some $\eta > 0$, $C_c < \infty$.

Define the leading Gram direction

$$V(U) = F_p(U)U^\top + UF_p(U)^\top. \tag{58}$$

It is symmetric and satisfies $\operatorname{tr} V(U) = 2\langle U, F_p(U)\rangle_{\mathrm{F}}$. Since F_p and ϕ are C^1 on the compact normalized path, the bounds below can be taken uniformly.

The following is the formal version of Theorem 2.

Theorem C.1 (Invariant response-order law)**.** Assume (A1) and (A2), and set $\kappa_c = p + \nu - 1$ and $\theta = \min\{\delta, \eta\}$.

(i) *Gram, size, and shape flow.* For $r \leq r_0$,

$$\dot{K} = r^{p+1} V(U) + O_{\mathrm{F}}(r^{p+1+\delta}),$$

and hence

$$\begin{aligned} \dot{r} &= \frac{1}{2} r^p \operatorname{tr} V(U) + O(r^{p+\delta}), \\ \dot{\rho} &= r^{p-1}\big[V(U) - (\operatorname{tr} V(U))\rho\big] + O_{\mathrm{F}}(r^{p-1+\delta}). \end{aligned} \tag{59}$$

The matrix in brackets has zero trace.

(ii) *Response order.* Along the path,

$$\begin{aligned} \dot{c} = r^{\kappa_c}\Big\{ &\frac{\nu}{2}(\operatorname{tr} V(U))\phi(\rho) \\ &+ D\phi_\rho\big[V(U) - (\operatorname{tr} V(U))\rho\big]\Big\} + O(r^{\kappa_c+\theta}). \end{aligned} \tag{60}$$

Consequently, if the coefficient in braces is bounded away from zero along the path, then $|\dot{c}| \asymp r^{p+\nu-1}$. In particular, a scale-free Gram measurement has order $p-1$, response amplitude has order p, and a raw Gram measurement has order $p+1$.

Proof. By homogeneity, $F_p(H) = r^p F_p(U)$. Differentiating $K = HH^\top$ therefore gives

$$\dot{K} = r^{p+1}\big(F_p(U)U^\top + UF_p(U)^\top\big) + G(H)H^\top + HG(H)^\top.$$

The last two terms have Frobenius norm $O(r^{p+1+\delta})$, proving the first claim in part (i). Since $r^2 = \operatorname{tr} K$,

$$\dot{r} = \frac{\operatorname{tr} \dot{K}}{2r} = \frac{1}{2} r^p \operatorname{tr} V(U) + O(r^{p+\delta}).$$

Differentiating $\rho = K/r^2$ and substituting the two expansions gives

$$\dot{\rho} = \frac{\dot{K}}{r^2} - 2\frac{\dot{r}}{r}\rho = r^{p-1}\big[V(U) - (\operatorname{tr} V(U))\rho\big] + O_{\mathrm{F}}(r^{p-1+\delta}).$$

Its leading coefficient has trace $\operatorname{tr} V - (\operatorname{tr} V)\operatorname{tr}\rho = 0$.

For part (ii), apply the chain rule to $c = r^\nu \phi(\rho) + \bar{c}(H)$:

$$\begin{aligned} \dot{c} &= \nu r^{\nu-1}\dot{r}\,\phi(\rho) + r^\nu D\phi_\rho[\dot{\rho}] + D\bar{c}(H)[\dot{H}] \\ &= r^{p+\nu-1}\left\{\frac{\nu}{2}(\operatorname{tr} V)\phi(\rho) + D\phi_\rho[V - (\operatorname{tr} V)\rho]\right\} + O(r^{p+\nu-1+\theta}). \end{aligned} \tag{61}$$

The derivative bound in (A2) and $\left\|\dot{H}\right\|_{\mathrm{F}} = O(r^p)$ control the final remainder. If the coefficient in braces has absolute value at least $m > 0$, then the remainder is at most half the leading term for small enough r, which gives the stated two-sided order bound. □

The exponent also holds for a more general scalar response $c(rU) = r^\nu \chi(U) + \bar{c}(rU)$ that need not be a function of the Gram state. Writing $q(U) = \langle U, F_p(U) \rangle_{\mathrm{F}}$ and $F_p^\perp(U) = F_p(U) - q(U)U$ gives the leading coefficient

$$\nu q(U)\chi(U) + D\chi_U[F_p^\perp(U)]. \tag{62}$$

The main text uses the invariant Gram specialization because those are the measurements used by the proposed observer.

The response-order theorem concerns local velocity. A lower response order implies an earlier fixed-distance crossing only when the motion persists toward the crossing. The next result states those extra conditions.

Theorem C.2 (Fixed-distance crossing times)**.** Let $\varepsilon > 0$ be a small parameter, and let $x_\varepsilon(t)$ be a C^1 real function of $t \geq 0$. Fix $D_x > 0$, $\kappa_x \in \mathbb{R}$, and $0 < c_x \leq C_x < \infty$, all independent of ε. Define $T_x(\varepsilon) = \inf\{t \geq 0 : x_\varepsilon(t) - x_\varepsilon(0) \geq D_x\}$ and assume $c_x \varepsilon^{\kappa_x} \leq \dot{x}_\varepsilon(t) \leq C_x \varepsilon^{\kappa_x}$ for $0 \leq t \leq T_x(\varepsilon)$. Then

$$\frac{D_x}{C_x \varepsilon^{\kappa_x}} \leq T_x(\varepsilon) \leq \frac{D_x}{c_x \varepsilon^{\kappa_x}}. \tag{63}$$

If y_ε satisfies the same conditions with constants D_y, κ_y, c_y, C_y, then

$$T_x = \Theta(\varepsilon^{-\kappa_x}), \qquad T_y = \Theta(\varepsilon^{-\kappa_y}), \qquad \frac{T_x}{T_y} \longrightarrow 0 \quad \Longleftrightarrow \quad \kappa_x < \kappa_y. \tag{64}$$

Proof. Integrating the lower speed bound shows that the distance D_x is reached no later than $D_x/(c_x \varepsilon^{\kappa_x})$. Integrating the upper bound shows that it cannot be reached before $D_x/(C_x \varepsilon^{\kappa_x})$. This proves Eq. (63). Applying the same bounds to y_ε and taking ratios gives

$$\frac{D_x c_y}{C_x D_y}\, \varepsilon^{\kappa_y - \kappa_x} \leq \frac{T_x}{T_y} \leq \frac{D_x C_y}{c_x D_y}\, \varepsilon^{\kappa_y - \kappa_x},$$

which proves Eq. (64). □

Remarks. If the coefficient in Eq. (60) vanishes, the first later nonzero term sets the realized response order. The crossing theorem is a separate statement about sustained motion across a fixed distance. The response-order law supplies its speed bounds with $\varepsilon = r(0)$ when r remains comparable to ε during the crossing, as in the higher-order case $p > 1$. For $p = 1$, response size changes exponentially, and the logarithmic timing law is derived next.

C.2 Timing regimes in intrinsic time

This subsection derives Eq. (9). Let τ be the intrinsic time $\mathrm{d}\tau = r^{p-1}\mathrm{d}t$, in which normalized Gram geometry moves at order one. From part (i) of Theorem C.1, the radial solution is

$$r(\tau) = \varepsilon \exp\left(\frac{1}{2}\int_0^\tau \operatorname{tr} V(U(s))\mathrm{d}s\right)(1 + o(1)). \tag{65}$$

For $p = 1$, intrinsic and wall time agree: normalized geometry crosses in $O(1)$, whereas fixed-radius expression requires logarithmic growth. For $p > 1$, $\mathrm{d}t = r^{1-p}\mathrm{d}\tau$, and

$$t(\tau) = \varepsilon^{1-p} \int_0^\tau \exp\left(-\frac{p-1}{2}\int_0^s \operatorname{tr} V(U(v))\mathrm{d}v\right)\mathrm{d}s\,(1 + o(1)). \tag{66}$$

Thus both crossings scale as ε^{1-p}, possibly with a limiting ratio in $(0, 1)$.

C.3 Anisotropic residuals and time-dependent local dynamics

We record two extensions of the response-order law, with sketches rather than full proofs. First, a single radius is unnecessary. In a fixed local chart, let weak entries or blocks scale as $\zeta_i = \varepsilon^{w_i}\bar{\zeta}_i$, $w_i > 0$, and expand $F_i(\zeta) = \sum_\alpha f_{i\alpha}\zeta^\alpha$ and $c(\zeta) = \sum_\beta c_\beta \zeta^\beta$. Every nonzero monomial in $Dc[F]$ has weighted exponent

$$\langle w, \alpha + \beta - e_i \rangle\,. \tag{67}$$

The smallest noncancelling exponent is the realized response order, and Eq. (64) applies unchanged.

Second, for a time-dependent linearization along the surrounding training path, suppose $\dot{H}_\varepsilon = \mathcal{L}_t[H_\varepsilon] + O_{\mathrm{F}}(\|H_\varepsilon\|_{\mathrm{F}}^{1+\delta})$ and $H_\varepsilon(0) = \varepsilon H_0$. On each fixed time interval, variation of constants gives

$$H_\varepsilon(t) = \varepsilon H^{\mathrm{lin}}(t) + o_{\mathrm{F}}(\varepsilon). \tag{68}$$

Hence the normalized response converges to $H^{\mathrm{lin}}(t)/\left\|H^{\mathrm{lin}}(t)\right\|_{\mathrm{F}}$, while its Frobenius norm is $\varepsilon \|H_0\|_{\mathrm{F}} e^{\Gamma(t)}$ to leading order; a transverse geometric crossing occurs at $t_F + o(1)$.

If the linear comparison remains accurate up to constant factors until the response reaches fixed size r_*, then expression satisfies

$$\Gamma(T_{\mathrm{expr}}) = \log \frac{r_*}{\varepsilon \|H_0\|_{\mathrm{F}}} + O(1). \tag{69}$$

If $\Gamma(t) = \bar{\gamma} t + O(1)$, inversion gives the main-text logarithmic law. Uniformly well-conditioned linear reparameterizations change the log Frobenius norm by only $O(1)$ and preserve its leading coefficient. For piecewise-smooth networks the argument holds within activation regions; boundary crossings are discrete events.

C.4 Statistical visibility and checkpoint cadence

Let a weak factor be $X = rG$ and observe $\widehat{X} = X + W$. If $\|W\|_{op} \le r\sigma_k(G)/2$, Wedin's perturbation bound (Wedin, 1972) gives

$$\left\|\widehat{P} - P\right\|_{op} \le \frac{2\|W\|_{op}}{r\sigma_k(G)}. \tag{70}$$

For direct Gram observations, the eigengap is order r^2 (Davis & Kahan, 1970), changing the dependence on additive noise. Thus probe count, conditioning, and the smallest active singular value determine T_{vis}.

For a two-point Gaussian experiment with m replicates, templates G_0, G_1, and noise variance σ^2, the joint divergence is

$$D_{\mathrm{KL}} = \frac{mr^2}{2\sigma^2} \|G_0 - G_1\|_{\mathrm{F}}^2 . \tag{71}$$

Pinsker's inequality bounds total variation by $O(r\sqrt{m}/\sigma)$, so below this scale no method reliably identifies the support or produces a reliable geometric warning interval.

For linear growth, $T(a;\varepsilon) = \gamma_*^{-1}\log(a/\varepsilon) + O(1)$. Substituting $a = r_{\mathrm{vis}} = \varepsilon^\beta L(\varepsilon)$ with slowly varying L and normalizing by the expression time gives

$$\frac{T_{\mathrm{alarm}}}{T_{\mathrm{expr}}} \longrightarrow [1-\beta]_+, \qquad [x]_+ = \max\{x, 0\}. \tag{72}$$

A fixed observation floor has $\beta = 0$ and removes the leading relative warning interval. Improving resolution increases β and moves the alarm toward the latent geometric crossing.

Finite differences introduce a separate cadence tradeoff. If a Gram-state estimate has error e_t, then a backward tangent estimate over interval Δ satisfies locally

$$\left\|\widehat{\Xi}_t - \Xi_t\right\| \le C_{\mathrm{curv}}\Delta + C_{\mathrm{obs}}\frac{e_t + e_{t-\Delta}}{\Delta}. \tag{73}$$

Long intervals incur curvature bias; short intervals amplify noise. Sparse gaps are therefore reported as intervals or censored observations.

D Stable measurements and causal events

This section proves that distance from a declared no-change family is the largest stable score (Section 4), gives its likelihood and support forms, and specifies the causal event construction used in every experiment.

D.1 STABLE MEASUREMENTS

Let $(\mathcal{X}, d_g)$ be a metric space. For $f : \mathcal{X} \to \mathbb{R}$ define the sensitivity $\mathrm{Lip}_g(f) = \sup_{u \neq v} |f(u) - f(v)|/d_g(u, v) \in [0, \infty]$; f is *unit stable* when $\mathrm{Lip}_g(f) \leq 1$. For a nonempty $\mathcal{N} \subseteq \mathcal{X}$ set $S_{\mathcal{N}}(x) = d_g(x, \mathcal{N}) = \inf_{z \in \mathcal{N}} d_g(x, z)$, as in Eq. (11).

The following is the formal version of the optimality statement in Eq. (12).

Theorem D.1 (Distance is the maximal stable measurement)**.** Let $\mathcal{N} \subseteq \mathcal{X}$ be nonempty and closed.

(i) *Stability and null.* $S_{\mathcal{N}}$ is unit stable,

$$|S_{\mathcal{N}}(x) - S_{\mathcal{N}}(z)| \leq d_g(x, z) \qquad \text{for all } x, z \in \mathcal{X}, \tag{74}$$

and $S_{\mathcal{N}}(x) = 0$ if and only if $x \in \mathcal{N}$.

(ii) *Extremality.* For every $x \in \mathcal{X}$,

$$S_{\mathcal{N}}(x) = \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\ f|_{\mathcal{N}}=0 \\ \mathrm{Lip}_g(f)\leq 1}} |f(x)|, \tag{75}$$

and the supremum is attained by $f = S_{\mathcal{N}}$. Equivalently, every f vanishing on $\mathcal{N}$ with $0 < \mathrm{Lip}_g(f) < \infty$ satisfies $|f(x)|/\mathrm{Lip}_g(f) \leq S_{\mathcal{N}}(x)$.

(iii) *Block specialization.* Let $(V, \langle\cdot,\cdot\rangle)$ be a finite-dimensional inner-product space with norm $\|\cdot\|$, let Π be an orthogonal projection on V, and let $\mathcal{N}_\Pi = \{W \in V : \Pi W = 0\}$. Then for every $\Xi \in V$,

$$d(\Xi, \mathcal{N}_\Pi) := \inf_{W \in \mathcal{N}_\Pi} \|\Xi - W\| = \|\Pi\Xi\| = \sup_{\substack{\Pi Y = Y \\ \|Y\| \leq 1}} \langle Y, \Xi\rangle\,. \tag{76}$$

In particular, take $V = T_K\mathcal{S}_k$ with the inner product g_K of Lemma B.1 and $\Pi = \Pi_B$ the g_K-orthogonal projection onto one of the four subspaces of Theorem B.1. Then $\mathcal{N}_B = \{W : \Pi_B W = 0\}$ is the family of tangents with no motion of type B, and $d_K(\Xi, \mathcal{N}_B) = \|\Pi_B\Xi\|_K = v_B$.

Proof. *(i)* For $x, z \in \mathcal{X}$ and any $y \in \mathcal{N}$, the triangle inequality gives $d_g(x, y) \leq d_g(x, z) + d_g(z, y)$; taking the infimum over y yields $S_{\mathcal{N}}(x) \leq d_g(x, z) + S_{\mathcal{N}}(z)$, and exchanging x and z gives Eq. (74). If $x \in \mathcal{N}$ then $S_{\mathcal{N}}(x) \leq d_g(x, x) = 0$. Conversely, if $S_{\mathcal{N}}(x) = 0$ there are $y_n \in \mathcal{N}$ with $d_g(x, y_n) \to 0$, so x lies in the closure of $\mathcal{N}$, which is $\mathcal{N}$ itself.

(ii) Let f vanish on $\mathcal{N}$ with $\mathrm{Lip}_g(f) \leq 1$. For any $y \in \mathcal{N}$,

$$|f(x)| = |f(x) - f(y)| \leq d_g(x, y),$$

and taking the infimum over y gives $|f(x)| \leq S_{\mathcal{N}}(x)$. By (i), $S_{\mathcal{N}}$ is itself admissible and $|S_{\mathcal{N}}(x)| = S_{\mathcal{N}}(x)$, so the supremum is attained. For the equivalent form, apply the bound to $f/\mathrm{Lip}_g(f)$, which is admissible; rescaling f leaves $|f|/\mathrm{Lip}_g(f)$ unchanged, and an observation error of size δ changes f by at most $\mathrm{Lip}_g(f)\,\delta$.

(iii) Write $\Xi = \Pi\Xi + (I - \Pi)\Xi$. For $W \in \mathcal{N}_\Pi$ the difference $\Xi - W = \Pi\Xi + \big((I - \Pi)\Xi - W\big)$ splits into orthogonal pieces, so $\|\Xi - W\|^2 = \|\Pi\Xi\|^2 + \|(I - \Pi)\Xi - W\|^2 \geq \|\Pi\Xi\|^2$, with equality at $W = (I - \Pi)\Xi \in \mathcal{N}_\Pi$. For the supremum, if $\Pi Y = Y$ then $\langle Y, \Xi\rangle = \langle Y, \Pi\Xi\rangle \leq \|Y\|\,\|\Pi\Xi\|$ by Cauchy–Schwarz, with equality at $Y = \Pi\Xi/\|\Pi\Xi\|$ when $\Pi\Xi \neq 0$ (and trivially at $Y = 0$ otherwise). The Gram-state case follows by taking $V = T_K\mathcal{S}_k$, $\langle\cdot,\cdot\rangle = g_K$, and $\Pi = \Pi_B$, and using $\|\Pi_B\Xi\|_K = v_B$ from Eq. (50). □

Thus the block norm is both the minimum correction removing that motion and the maximum unit-sensitivity linear response blind to other blocks. With no preferred within-block direction, the four block norms are the complete invariants under independent blockwise isometries.

D.2 Likelihood, response-order, and structural forms of the distance

For $\widehat{\Xi} = \Xi + \sigma Z$ in g_K-orthonormal coordinates, with standard Gaussian Z, the generalized likelihood ratio for $\Xi \in \mathcal{N}_B$ versus an unrestricted tangent is

$$2 \log \Lambda_B = \frac{\inf_{W \in \mathcal{N}_B} \left\| \widehat{\Xi} - W \right\|_{g,K}^2}{\sigma^2} = \frac{\left\| \Pi_B \widehat{\Xi} \right\|_{g,K}^2}{\sigma^2}. \tag{77}$$

An independently specified within-block direction gives the corresponding matched one-dimensional projection. For anisotropic error covariance, replace the squared norm by

$$\inf_{W \in \mathcal{N}_B} (\widehat{\Xi} - W)^\top \Sigma^{-1} (\widehat{\Xi} - W). \tag{78}$$

The intrinsic norm measures effect size; the whitened residual measures resolvability.

Assume $\mathcal{N}$ has a smooth nearest-point projection near $x_0 \in \mathcal{N}$, and write a nearby family in normal coordinates as

$$x_r = \mathrm{Exp}_{y_r} \left(r^\ell n + o(r^\ell) \right), \qquad y_r \in \mathcal{N}, \quad n \perp T_{x_0} \mathcal{N}. \tag{79}$$

Then the tubular-neighborhood expansion gives

$$S_{\mathcal{N}}(x_r) = r^\ell \left\| n \right\|_g + o(r^\ell). \tag{80}$$

By Eq. (75), no equally sensitive scalar with this null has a lower power of r or larger normalized coefficient. Combining with Eq. (63) yields the main-text response-order claim.

For the null of unit-trace states supported in P_0, set $s = \mathrm{tr}(P_0 \rho)$. The measurement $\{P_0, I - P_0\}$ gives $F(\rho, \sigma) \le \sqrt{s}$, attained by $P_0 \rho P_0 / s$ via purification projection (Uhlmann, 1976); $s = 0$ follows by continuity. Matching scale in Eq. (48) gives

$$S_{P_0}(\rho) = 2 \arccos \sqrt{\mathrm{tr}(P_0 \rho)}. \tag{81}$$

The rank-at-most-k null is the union of these sets over all rank-k projectors. Ky Fan's principle gives $\max_{\mathrm{rank}\, P_0 = k} \mathrm{tr}(P_0 \rho) = \sum_{i=1}^k p_i$, hence

$$S_{\le k}(\rho) = 2 \arccos \sqrt{\sum_{i=1}^k p_i}. \tag{82}$$

For outside mass q, $2 \arccos \sqrt{1 - q} = 2\sqrt{q} + O(q^{3/2})$, placing the distance on the feature-amplitude scale. A hard rank event still requires clearing observation uncertainty.

For allocation, $g_{\mathrm{FR}}(u, v) = \sum_i u_i v_i / p_i$. Any smooth spectral summary $\psi(p)$ satisfies

$$\dot{\psi} = g_{\mathrm{FR}}(\nabla_{\mathrm{FR}} \psi, \dot{p}), \qquad \frac{|\dot{\psi}|}{\left\| \nabla_{\mathrm{FR}} \psi \right\|_{\mathrm{FR}}} \le \left\| \dot{p} \right\|_{\mathrm{FR}} = v_{\mathrm{alloc}}. \tag{83}$$

Equality holds along the signed Fisher gradient. Thus Fisher speed is the direction-free score; entropy and participation ratio are matched directional summaries.

D.3 Causal path statistics

For checkpoint times t_j, define displacement, path length, and sampled speed by

$$D_{j|0} = d_g(K_{t_0}, K_{t_j}), \qquad L_{j|0} = \sum_{i=1}^j d_g(K_{t_{i-1}}, K_{t_i}), \qquad v_j = \frac{d_g(K_{t_{j-1}}, K_{t_j})}{t_j - t_{j-1}}. \tag{84}$$

The triangle inequality gives $D_{j|0} \le L_{j|0}$. Displacement records net change; path length also records reversals and is invariant under increasing reparameterization.

Map the current backward tangent Ξ_j and its causal history into one tangent space, and let $\mathcal{N}_j$ be a continuation fitted only to earlier transported tangents. Define

$$I_j = \inf_{W \in \mathcal{N}_j} \left\| \Xi_j - W \right\|_{g, K_{t_j}}. \tag{85}$$

For a linear history span this is the orthogonal residual norm. If $\Xi_j = X + bY$, with $X \in \mathcal{N}_j$, Y a unit normal, and $|b| \ll \|X\|_g$, then

$$I_j = |b|, \qquad \|\Xi_j\|_g - \|X\|_g = \frac{b^2}{2\,\|X\|_g} + O\left(\frac{b^4}{\|X\|_g^3}\right). \tag{86}$$

Innovation is first order in a new direction, whereas total-speed change is second order. Smooth established motion is instead summarized by displacement or path length. Finite differences incur Eq. (73). Table 1 in the main text summarizes how each proposed change selects its measurement.

D.4 Canonical watch, confirmation, rearming, and forecasting

At checkpoint t_j, the online monitor uses only fixed probes, the current model, and observations through t_j. A *watch* is the first threshold crossing, *confirmation* satisfies persistence, a *peak* is a later maximum within the episode, and *expression* is an offline behavioral event.

Given a causal no-change family, define

$$s_j = S_{\mathcal{N}_j}(x_j), \qquad T_{\text{watch}} = \min\{t_j : s_j > h_j\}. \tag{87}$$

If confirmation requires n_{pers} consecutive crossings,

$$T_{\text{alarm}} = \min\{t_j : s_{j-\ell} > h_{j-\ell} \text{ for } \ell = 0, \ldots, n_{\text{pers}} - 1\}. \tag{88}$$

With an additional response clock, a forecast may use only history through confirmation:

$$\widehat{T}_{\text{expr}} = \mathcal{F}(K_{t_0}, \ldots, K_{T_{\text{alarm}}}). \tag{89}$$

Smooth departures use displacement, speed, or path length; sharp path breaks use causal continuation distance. Rolling calibration begins only after sufficient history. Learned-system alarm times are confirmation times, with watches retained separately where available.

A sustained episode runs from watch until the release rule returns to the no-change state, after which the monitor rearms. A spike is localized to its sampled interval; sparse gaps are censored by Eq. (73). Post-watch peaks cannot be backdated as alarms, though a causal decline may define a second event. Episode duration and warning lead are distinct. Table 4 lists the score and confirmation rule used in each learned system.

Table 4: Causal event construction. Thresholds and persistence confirm the geometric score; behavior enters only through the expression record.

System	Causal score	Confirmation and rearming	Reported output
Deep linear	Residual-projector displacement after causal removal of the established mode	Threshold 0.1 for three checkpoints, with residual dominance at least 0.95; one event per run	Confirmed alarm before cumulative top-two subspace capture
Induction	Probe-consensus continuation departure s_j^{ind} in Eq. (100)	Threshold 4.5 for two checkpoints; 16-observation refractory period; at most three alarms	Confirmed path break and nearest pre-induction alarm
Grokking	Off-cone distance r_j, followed by entropy drawdown C_j	Entry confirms at $\sum r_j \geq 0.05$; concentration watches at $C_j \geq 0.005$ and confirms at $C_j \geq 0.01$ for four checkpoints	Post-interpolation entry, concentration watch and confirmation, and offline velocity crest
Pythia	Typed union s_j^{union} of shape and soft-support continuation departures	Threshold 3 for two checkpoints; local release and rearming; no alarm cap	Watch, confirmation, geometric core, and later peak

Event-local evaluation counts alarms before the admissible pre-event window as too early, alarms after expression as late, and absent alarms as missed. Forecast evaluation uses a fixed horizon and treats alarms outside it as false positives, preventing an always-on detector from receiving unbounded credit.

The later expression time is not the geometric episode itself. The forecast map $\mathcal{F}$ is analytic in exact models and calibrated on earlier runs in the learned systems.

D.5 How familiar measures arise from the geometry

Table 5 lists what familiar scalars retain and discard. These summaries need not be monotone; they respond only to retained components. Standard change-point methods apply to the score stream (Truong et al., 2020), and time-uniform confidence sequences can provide formal causal thresholds (Howard et al., 2021).

Table 5: Existing measurements as partial views of the observed learning path.

Measurement	Geometric object	Information retained and lost
Activation norm	$a = \operatorname{tr} K$	Growth only; normalized organization is removed.
Entropy, effective rank, participation ratio (Roy & Vetterli, 2007; Garrido et al., 2023)	a declared direction in the spectrum of ρ	Signed allocation change; other allocation directions, axes, and support motion are removed.
Projector speed or distance	$P = KK^{+}$	Equal-weight support motion; growth, allocation, and internal rotation are removed.
CKA and Gram similarities (Kornblith et al., 2019)	finite K or centered K	Endpoint similarity under a declared invariance; tangent type and timing require a path.
Gradient norm	$\|\mathrm{d}L\|_{A_\theta}$	Global force in Eq. (57); capacity and alignment are removed.
Continuation error or change point	distance from a causal continuation family	Temporal novelty of the supplied geometric path; the continuation model must be declared separately.
Local loss complexity	local loss geometry	Complementary structure near parameters or functions; not a coordinate of one representation.

E Grokking details and robustness

A single grokking path contains two geometric stages: entry into a post-interpolation reorganization near memorization, and a later spectral concentration before generalization. Neither occurs in no-weight-decay controls, although they memorize. This section defines both stages, reports their robustness, and records two offline landmarks and a local forecasting clock.

E.1 Overview

For multiplication modulo 97, the observer uses penultimate representations of 300 fixed uniformly sampled input pairs, without training membership, labels, the Fourier solution, or future accuracy. The first event, entry, accumulates the distance of each checkpoint step from the family in which growth is at least as large as normalized-shape motion. It appears in all 40 standard runs and all 12 runs stopped before 90 percent test accuracy, but in none of 12 no-weight-decay controls, even though those controls memorize. Entry therefore marks a post-interpolation reorganization rather than memorization alone.

Within that regime, a causal drawdown of spectral entropy identifies a later concentration stage. Its watch precedes the first sustained above-chance test lift in 36 of 40 standard runs and its stronger confirmation in 27 of 40; both precede 90 percent test accuracy in all 40. A symmetry analysis shows eventual concentration into the true modular solution sectors. Figure 6 summarizes both stages across runs, and Figure 7 gives the full trajectory and mechanism view. The claim is stage resolution, not a universal grokking forecast: the local forecasting gain is cohort dependent (Appendix E.6).

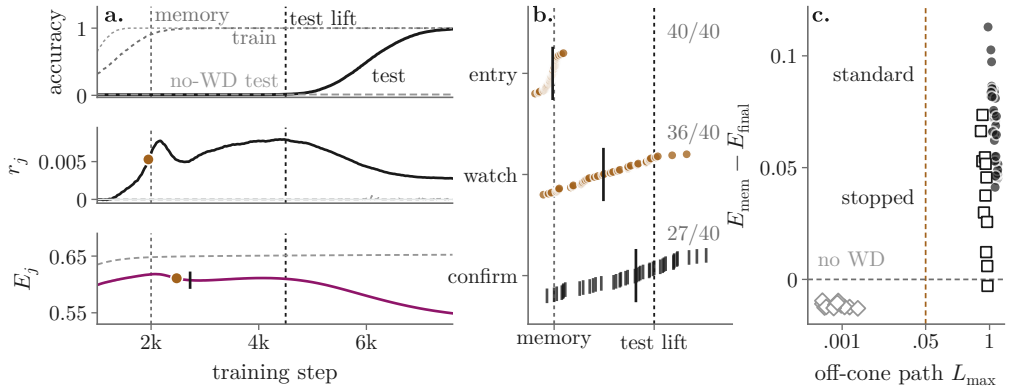

Figure 6: **One grokking path contains two geometric stages. a,** In a representative run, off-cone entry (gold circle) occurs near memorization, while entropy-drawdown watch and confirmation (gold dot and ink tick) occur later and before the first sustained test lift. The gray curves are a no-weight-decay control. **b,** Event times in all 40 standard runs, rescaled separately so memorization is zero and test lift is one; black bars are medians and right labels count events before the lift. **c,** Every standard, stopped, and no-weight-decay run plotted by its maximum off-cone path and entropy decrease. Standard and stopped runs show both changes; controls memorize but show neither.

E.2 Stage definitions

Let a_j and ρ_j be the scale and normalized shape of the fixed-probe Gram matrix at checkpoint j. If $Y_j Y_j^\top = \rho_j$, the aligned Bures chord between consecutive checkpoints is

$$\delta_j = \min_{O^\top O = I} \|Y_j O - Y_{j-1}\|_{\mathrm{F}}, \tag{90}$$

a finite-checkpoint form of the quotient geometry of Section 2 with the same infinitesimal metric. The growth and shape increments and their distance from the growth-dominated half-plane are

$$g_j = |\Delta \log a_j|, \qquad q_j = 2 d_{\mathrm{B}}^{\mathrm{chord}}(\rho_j, \rho_{j-1}), \qquad r_j = \frac{[q_j - g_j]_+}{\sqrt{2}}. \tag{91}$$

The complete finite increment is $(\Delta \log a_j)^2 + 4\delta_j^2$, and r_j is its Euclidean distance from the growth-dominated half-plane, without variance fitting or unstable division. Group $r_j > 0$ into maximal intervals; the first initializes the monitor. On later intervals, a positive episode beginning at s accumulates path length and confirms entry when

$$L_{s:j} = \sum_{i=s}^{j} r_i \geq \ell, \qquad \ell = 0.05, \tag{92}$$

resetting at $r_j = 0$. The rule is causal and independent of memorization.

After entry, the normalized spectral entropy and its causal drawdown are

$$E_j = -\frac{1}{\log N} \sum_{k=1}^{N} p_{k,j} \log p_{k,j}, \qquad C_j = \max_{i:\, T_{\text{entry}} \leq t_i \leq t_j} E_i - E_j. \tag{93}$$

Four consecutive values of C_j at 0.005 give a watch, and four at 0.01 give confirmation. For the symmetry-resolved analysis of Appendix E.5, with normalized sector masses $m_t(k,l)$,

$$E_t^{\mathrm{sym}} = -\frac{1}{\log M} \sum_{k,l} m_t(k,l) \log m_t(k,l). \tag{94}$$

The location of entry has a natural optimizer interpretation. Write one local update as

$$\Delta\rho_t \simeq J_\rho(\theta_t)\big(-\eta u_t - \eta\lambda\theta_t\big), \tag{95}$$

where u_t is the loss-driven update and λ is weight decay. Near interpolation, the fitting component of u_t contracts while decay remains, changing the observed direction before test accuracy rises. This local interpretation is not an exact optimizer solution.

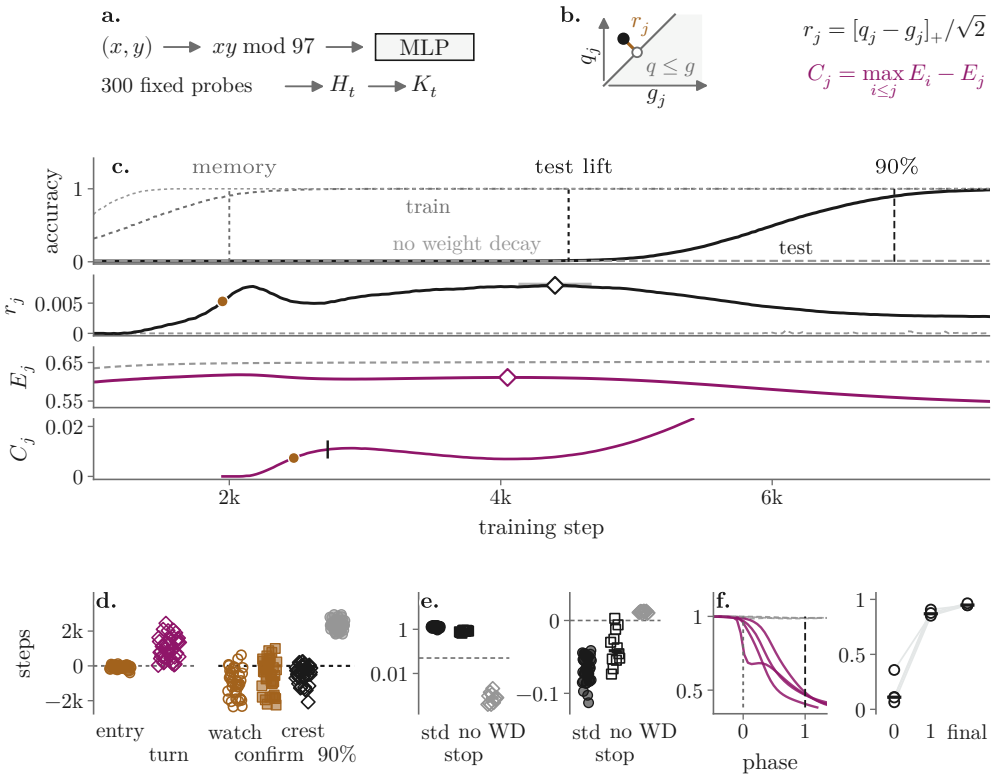

Figure 7: **Grokking has post-interpolation entry and later concentration. a,** Fixed unlabeled probes define the Gram path. **b,** r_j measures departure from growth dominance and C_j later concentration. **c,** Accuracy, off-cone motion, entropy, and drawdown; markers denote entry, offline crest, entropy turn, watch, and confirmation. **d,** Stage timings over 40 runs. **e,** Standard and stopped cohorts have long off-cone paths and falling entropy; controls do not. **f,** Cyclic-sector entropy contracts and mass concentrates into $k = \pm l$, unlike the random-order control.

E.3 Entry and concentration results

Entry thresholds 0.02–0.10 yield events in all 40 standard and 12 stopped-early runs and none of 12 no-weight-decay controls. Median confirmation ranges from 275 steps before to 250 after memorization; at 0.05 it is 25 steps before. The longest post-startup interval spans 216–271 checkpoints in standard runs, 138–148 in stopped runs, and 3–6 in no-weight-decay controls; the corresponding path lengths are 0.93–1.49, 0.67–0.91, and 0.0004–0.0021 (Fig. 6c).

All 40 standard and 12 stopped runs reach both the concentration watch and its confirmation; no no-weight-decay control is armed. From memorization to the final checkpoint, entropy falls in all 40 standard and 11 of 12 stopped runs, and rises in every control. The test lift is defined by chance: with 4,656 held-out examples and chance $1/97$, 72 correct exceeds the one-sided 99.9 percent binomial bound, and the lift is the first of three consecutive checkpoints at or above $72/4656$. The entropy watch precedes it in 36 of 40 runs and confirmation in 27; both precede sustained 90 percent accuracy in all 40. Confidence levels from 99 to 99.99 percent preserve the cohort ordering (Fig. 8d).

Drawdowns 0.005–0.020 detect all 40 complete standard runs and no no-weight-decay controls; larger thresholds occur later and exclude more stopped runs. Persistence of two to six checkpoints and retaining every second or fourth checkpoint shift the median by at most one original checkpoint when duration is matched (Fig. 8b,c).

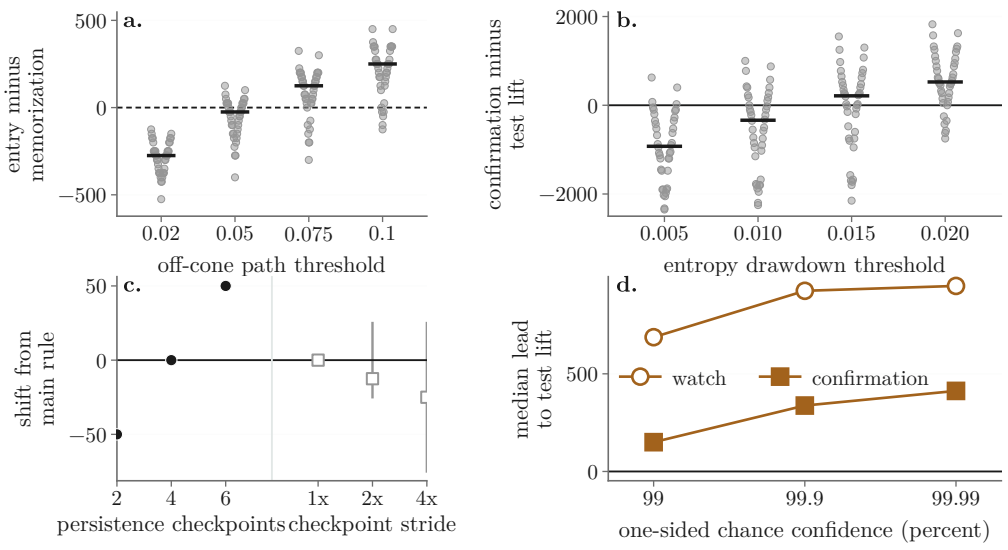

Figure 8: **Robustness of both grokking stages. a,** Entry moves smoothly with path threshold. **b,** Concentration moves later with drawdown threshold; all 40 standard runs confirm at every threshold, 12, 12, 11 and 9 of the 12 stopped runs do, and no control does. **c,** Persistence and subsampling shift median confirmation by at most one original checkpoint. **d,** Watch and confirmation retain positive median lead under three chance bounds.

E.4 Offline landmarks: velocity crest and entropy turn

A degree-three, 21-checkpoint centered Savitzky–Golay filter, restricted to 500-step supports after entry, places the offline velocity crest before the lift in 32 of 40 runs. Because it uses future observations, the crest is a path summary, not an alarm.

The last smoothed entropy maximum before terminal descent is a sharper offline landmark (Fig. 9). It lies between memorization and lift in all 40 runs, with median offsets $+987$ and -875 steps, and shifts by at most 100 steps across tested windows. A causal entropy-slope monitor alarms before lift in all 40; the turn-matched alarm confirms in 37 with median lead 788. The turn correlates 0.62 with post-memorization delay and reduces leave-one-run-out median error from 206 to 136 steps. In the representative run it occurs at step 4,050, 450 steps before lift. This is exploratory, not the main forecasting claim.

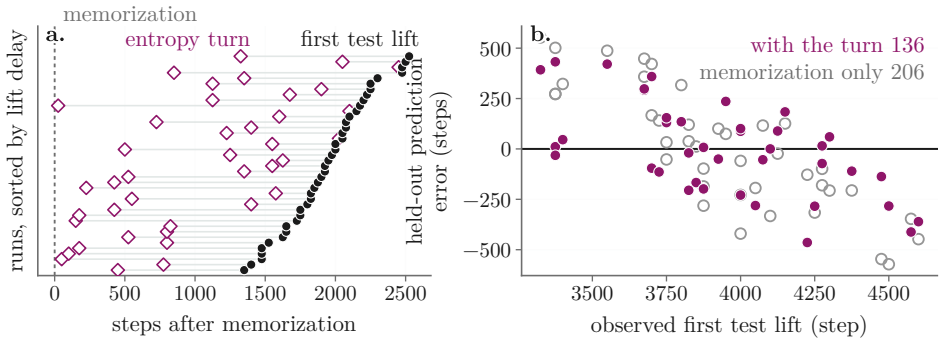

Figure 9: **Entropy-turn timing. a,** In all 40 standard runs, the turn lies between memorization and first test lift. **b,** Adding the turn to memorization reduces leave-one-run-out median error from 206 to 136 steps. This is an exploratory offline association; transfer was not tested.

E.5 SYMMETRY SECTORS

At 250-step checkpoints, double center each hidden unit on the complete nonzero table. Writing $x = g^a$, $y = g^b$ for a primitive generator, the two-dimensional DFT diagonalizes independent cyclic translations. Normalize the energy in sectors with both frequencies nonzero to $m_t(k, l)$. Changing the primitive generator only permutes these masses; sorted values differ by at most 1.1×10^{-8}. Effective sector count falls below 80 percent of its initial value for two checkpoints in all four runs, after memorization and before generalization. Independent row and column permutations produce no alarm and final entropy near one. The detector does not use $k = \pm l$; post hoc, its median mass rises from 0.095 at memorization to 0.858 at generalization and 0.948 finally (Fig. 7f).

E.6 A LOCAL FORECASTING CLOCK

For a separate local clock, six quantities from the preceding 12 observations describe shape, scale, and spectrum. Sixteen runs fit a shape boundary and a ridge map to remaining time. On 24 later runs, pooled median absolute error is 160.3 steps, versus 237.5 for a fixed boundary delay and 500 for a fixed schedule; 23 forecasts precede generalization and one follows it.

The gain is cohort dependent (Fig. 10b, in Appendix G): local-clock versus fixed-delay median errors are 246.3 versus 175 in growth replication and 149.5 versus 350 in complete confirmation. No no-weight-decay control is forecast. Four right-censored runs cross the boundary, but all predictions lie beyond their horizon; the clock is therefore not a stable general rule.

F DEEP-LINEAR DETAILS AND CONTROLS

This section gives the offline event, the online score, and the controls behind Section 5.1. The deep-linear task trains a five-layer factorization on a teacher with sequential singular modes, using fixed Gaussian probes. If the teacher response is $H_\star = U_\star \Sigma_\star V_\star^\top$, define $\Pi_2^\star = U_{\star,:2} U_{\star,:2}^\top$. For $H_t = U_t \Sigma_t V_t^\top$, let $P_2(t) = U_{t,:2} U_{t,:2}^\top$ and define the cumulative teacher-subspace mass

$$S_2(t) = \operatorname{tr}(\Pi_2^\star P_2(t)), \qquad 0 \le S_2(t) \le 2. \tag{96}$$

The offline event is the first of three consecutive checkpoints satisfying

$$2 - S_2(t) \le 0.05, \tag{97}$$

while the second singular value carries at least 0.95 of the combined second-plus-third squared mass. The teacher projector is used only offline. Indexed second-mode alignment is diagnostic only, since weak directions may reshuffle while S_2 remains a stable cumulative measure.

Online, the first mode is anchored after eight observations, dominance at least 0.95, and four consecutive projector speeds below five percent of their running maximum. Removing the current leading direction gives $P^{\mathrm{res}}(t) = u_2(t) u_2(t)^\top$ and score

$$D^{\mathrm{res}}(t_j) = 2^{-1/2} \left\| P^{\mathrm{res}}(t_j) - P^{\mathrm{res}}(t_0) \right\|_{\mathrm{F}}. \tag{98}$$

An alarm requires three consecutive scores at least 0.1 and second-versus-third dominance at least 0.95. The rule is fixed before the 12 evaluation runs are scored.

Under the rank-one structural null, all eight runs anchor and none alarms; the largest displacement is 0.0082. Four fixed-horizon runs that expressed a second mode are slow positives and are not counted as controls.

As a rank-one projector, P^{res} moves only by support transport; finite tangents are reconstructed to error below 2.5×10^{-15}. The warned direction aligns with the next teacher direction in every run. An independent weight-space projector alarms in all 12 runs, with median absolute lag 37.5 steps and maximum lag 300. In Fig. 2, the alarm leads the $S_2/2 = 0.975$ event by 525 steps. Across runs, the median angle contracts from 36° at alarm to 13° at acquisition; median lead is $1{,}750$ steps (range 50–11,300).

G INDUCTION DETAILS AND INTERVENTIONS

This section gives the blind score and forecast behind Section 5.2, then the repetition-conditioned observations and head ablations used to localize the mechanism.

G.1 BLIND CONTINUATION SCORE AND FORECAST

For sampled state x_j, let $\dot{x}_j$ be the shortest aligned finite difference and $\widehat{x}_{j,w,o}$ a constant or linear continuation fit to the preceding w observations. Define

$$\mathcal{I}_j[x] = \max_{w,o} \left\{ \mathcal{Z}_{<j}\left(\left\| \dot{x}_j - \widehat{x}_{j,w,o} \right\|_{g,x_j} \right), \mathcal{Z}_{<j}\big(\mathrm{dist}_{g,x_j}(\dot{x}_j, \mathrm{span}\{\dot{x}_{j-w}, \ldots, \dot{x}_{j-1}\})\big) \right\}. \quad (99)$$

For the full, even, and odd first-half probe banks $\mathcal{B}$, the blind induction score is

$$s_t^{\mathrm{ind}} = \mathrm{median}_{b\in\mathcal{B}}\, \mathcal{I}_t[\rho^{(b)}]. \quad (100)$$

The blind observer uses centered final-residual representations of fixed random prefixes, before any repeated suffix. For principal factors $Y_j Y_j^\top = \rho_j$, choose O_j minimizing $\|Y_j O - Y_{j-1}\|_{\mathrm{F}}$ and set

$$V_j = \frac{2(Y_j O_j - Y_{j-1})}{t_j - t_{j-1}}. \quad (101)$$

Its Frobenius norm is the sampled quotient norm. For $w \in \{4, 8, 16\}$, constant and linear fits give the first residual in Eq. (99); projection onto the recent span gives the second. Standardize each sequence from at least 12 earlier finite values:

$$\mathcal{Z}_{<j}(r_j) = \frac{r_j - \mathrm{median}_{i<j}\, r_i}{\max\{1.4826\,\mathrm{MAD}_{i<j}(r_i), 0.05|\,\mathrm{median}_{i<j}\, r_i|, 10^{-12}\}}, \quad (102)$$

using at most 32 values. The maximum over six fits and the span residual gives $\mathcal{I}_j$; the median over full, even, and odd banks gives s_j^{ind}. Detection begins after warmup, requires two consecutive values above 4.5, uses a 16-observation refractory period, and permits at most three alarms. It sees no heads, lags, repetition contrasts, or behavior. In Fig. 3, the blind alarm leads induction by 330 steps and the forecast by 10; the conditioned entropy alarm (Appendix G.2) leads by 275 steps.

For confirmed alarms A_i and induction events T_i, four development runs set

$$\widehat{\Delta} = \mathrm{median}_{i\in\mathcal{C}}(T_i - A_i) = 320, \qquad \widehat{T} = A + \widehat{\Delta}. \quad (103)$$

Applied unchanged to 24 later runs, median and mean absolute errors are 30 and 30.8 steps, versus median 50 for a fixed transition step (Fig. 10a). Every forecast precedes induction with median issue lead 320 steps; none of 12 no-repetition controls is forecast.

G.2 CONDITIONED OBSERVATIONS AND HEAD INTERVENTIONS

For the repetition-conditioned analysis, let A_t be the mean first-half response, B_t the second-half response to $[x, x]$, and B_t^π the second-half response to $[x, \pi(x)]$. With P_t the leading rank-eight support of A_t,

$$C_t^\pi = (I - P_t)(B_t - B_t^\pi), \qquad K_t^{\mathrm{rel}} = \mathbb{E}_\pi[C_t^\pi C_t^{\pi\top}]. \quad (104)$$

If $p_i(t) = \lambda_i(K_t^{\mathrm{rel}})/\operatorname{tr} K_t^{\mathrm{rel}}$ and r is the maximum resolved rank, the normalized relational entropy is

$$E_t = -\frac{1}{\log r} \sum_i p_i(t) \log p_i(t). \quad (105)$$

For $\rho_t^{(-h)}$ computed after ablating head h, its signed contribution along the observed allocation direction is

$$b_h = E(\rho_t^{\mathrm{rel}}) - E(\rho_t^{(-h)}) \simeq \left\langle \nabla_\rho E, \rho_t^{\mathrm{rel}} - \rho_t^{(-h)} \right\rangle. \quad (106)$$

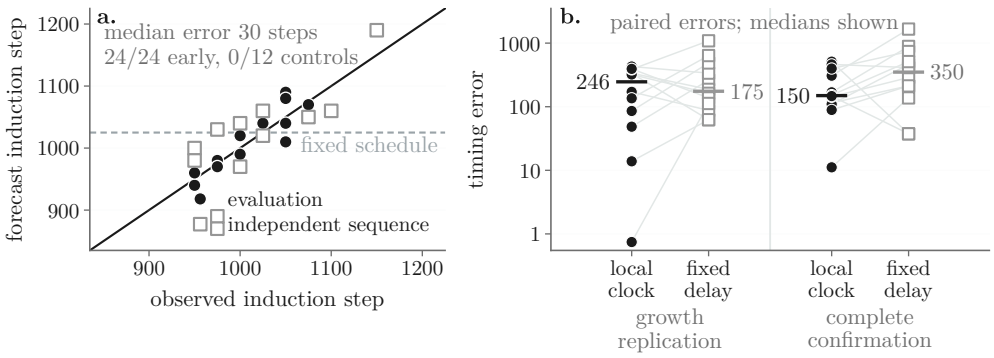

Figure 10: **Geometric clocks for later expression. a,** A 320-step induction latency calibrated on four runs gives median error 30 on 24 later runs, versus 50 for a fixed schedule. **b,** The grokking local clock improves pooled timing but not both cohorts; paired points compare it with a fixed delay from the same boundary.

For mechanism localization, M second-half permutations preserve token counts but destroy repeated-position relations. Equation (104) averages their residual Grams before normalization; the maximum resolved rank normalizes E_t to $[0, 1]$. The alarm covers both 12-run cohorts and leaves four slow and eight matched-rate no-repetition controls quiet.

At the first checkpoint after induction, each head is ablated and scored by b_h. The two largest values define the broadening pair and the two smallest the concentrating pair, without behavioral selection. The rule is fixed before the replication runs; in all 12, broadening-pair ablation causes the larger ICL drop.

H Pythia details, cross-scale audit, and behavior battery

This section specifies the Pythia observer of Section 5.3, reports motion fractions and the cross-scale audit, ranks events by intrinsic size without thresholds, and records the behavior battery.

The observer uses 96 fixed length-64 probes. Natural probes are fixed WikiText validation offsets; relational probes repeat a 32-token half, with token-matched permuted controls (Fig. 11a). Hidden states are sampled at eight second-half positions; attention features average each head over relative lag. Features are centered and split into three interleaved probe banks. For natural, repeated, and permuted inputs x, x^{rep}, and x^{perm}, define

$$H^{\mathrm{nat}}_{t,\ell} = h_{t,\ell}(x), \qquad H^{\mathrm{rel}}_{t,\ell} = h_{t,\ell}(x^{\mathrm{rep}}) - h_{t,\ell}(x^{\mathrm{perm}}), \qquad H^{\mathrm{att}}_{t,\ell} = A_{t,\ell}(x^{\mathrm{rep}}) - A_{t,\ell}(x^{\mathrm{perm}}). \tag{107}$$

Here $A_{t,\ell}$ is the flattened head-by-lag profile, and $K^b_{t,\ell} = d_b^{-1} \bar{H}^b_{t,\ell} \bar{H}^{b\top}_{t,\ell}$. Geometry uses every public Pythia-1B checkpoint through 143,000; behavior is excluded from event construction.

For each layer and split, Eq. (99) uses four- and eight-checkpoint constant and linear continuations, with a 12-observation baseline after six-checkpoint warmup. The states are normalized shape and soft support

$$P^\tau(K) = K(K + \tau I)^{-1}, \qquad \tau = 0.1 \, \frac{\operatorname{tr} K}{N}. \tag{108}$$

Aggregate each departure by the 75th layer percentile and split median, then standardize their maximum using 12 earlier view-level values after warmup of eight. With scores $c^{\mathrm{shape}}_{j,b,s}$ and $c^{\mathrm{supp}}_{j,b,s}$,

$$s_{j,b} = \mathcal{Z}_{<j}\left(\max\left\{\operatorname{median}_s c^{\mathrm{shape}}_{j,b,s}, \operatorname{median}_s c^{\mathrm{supp}}_{j,b,s}\right\}\right), \qquad s^{\mathrm{union}}_j = \max_b s_{j,b}. \tag{109}$$

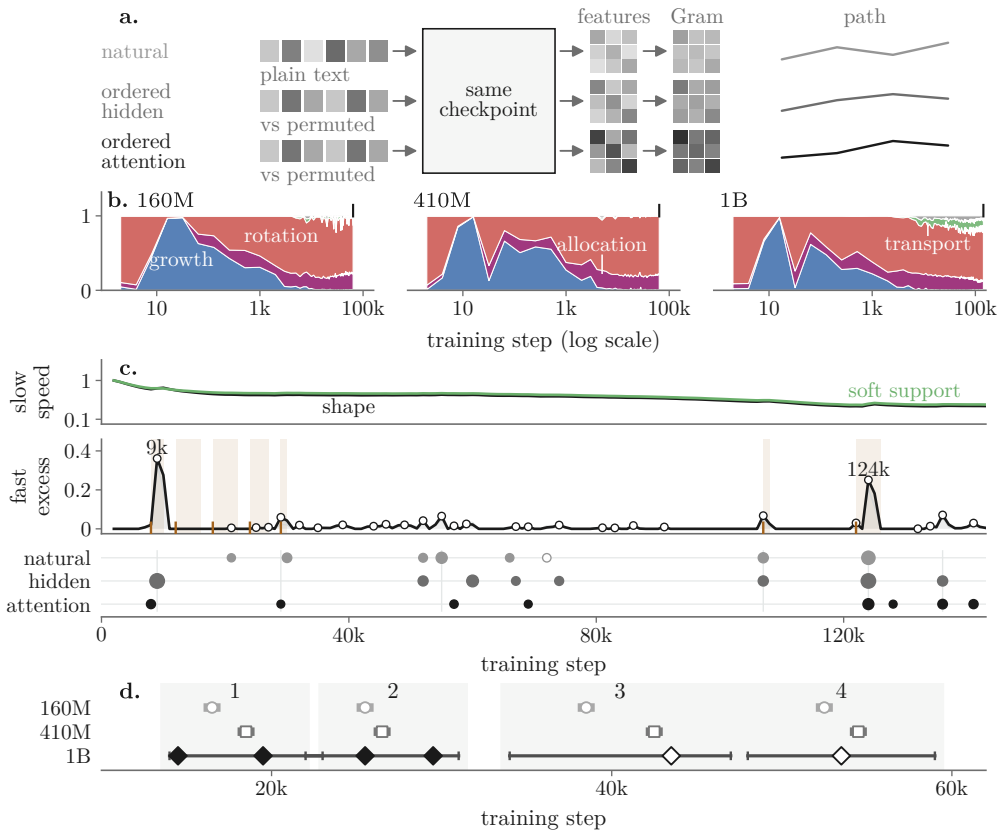

Figure 11: **Pythia geometric atlas.** Three fixed probe views show the growth-to-rotation progression, compare 1B causal surprise with intrinsic magnitude, and provide the descriptive cross-size audit.

A watch requires $s_j^{\text{union}} \geq 3$ and the next crossing confirms. Gaps over four times the median of the preceding three are censored. Returning below threshold rearms the uncapped monitor.

The local layer score and positive-layer fraction are

$$c_{j,b,\ell} = \max\left\{\text{median}_s\, \mathcal{I}_j[\rho^{b,\ell,s}], \text{median}_s\, \mathcal{I}_j[P^\tau(K^{b,\ell,s})]\right\}, \qquad B_{j,b} = \frac{1}{L_b}\sum_{\ell=1}^{L_b} \mathbf{1}\{c_{j,b,\ell} > 0\}. \tag{110}$$

Separately, applying Eq. (4) to each checkpoint chord gives

$$f^b_{m,j} = \frac{\sum_{\ell,s}(v^{b,\ell,s}_{m,j})^2}{\sum_{m'}\sum_{\ell,s}(v^{b,\ell,s}_{m',j})^2}, \qquad \sum_m f^b_{m,j} = 1. \tag{111}$$

These fractions sum to one when defined and are computed only after detection, separating event discovery from motion typing.

After step 2,000 (Fig. 11b), relational-attention median rotation fractions are 0.750, 0.796, and 0.666 for 160M, 410M, and 1B; allocation fractions are 0.195, 0.191, and 0.158, and transport fractions 0.023, 0.002, and 0.122. Positive residual-layer breadth rises from zero at step 128 to one at 512 in all sizes, then falls by 1,000. The seven 1B watches occur at 8k, 12k, 18k, 24k, 29k, 107k, and 122k; every active bank is rotation dominated.

Four windows derived from the 160M and 410M paths alone each contain a 1B branch. Because they cover 41,000 of the 45,000 steps from 14k to 59k and lack matched random windows, this is a descriptive audit only (Figs. 11d and 12c).

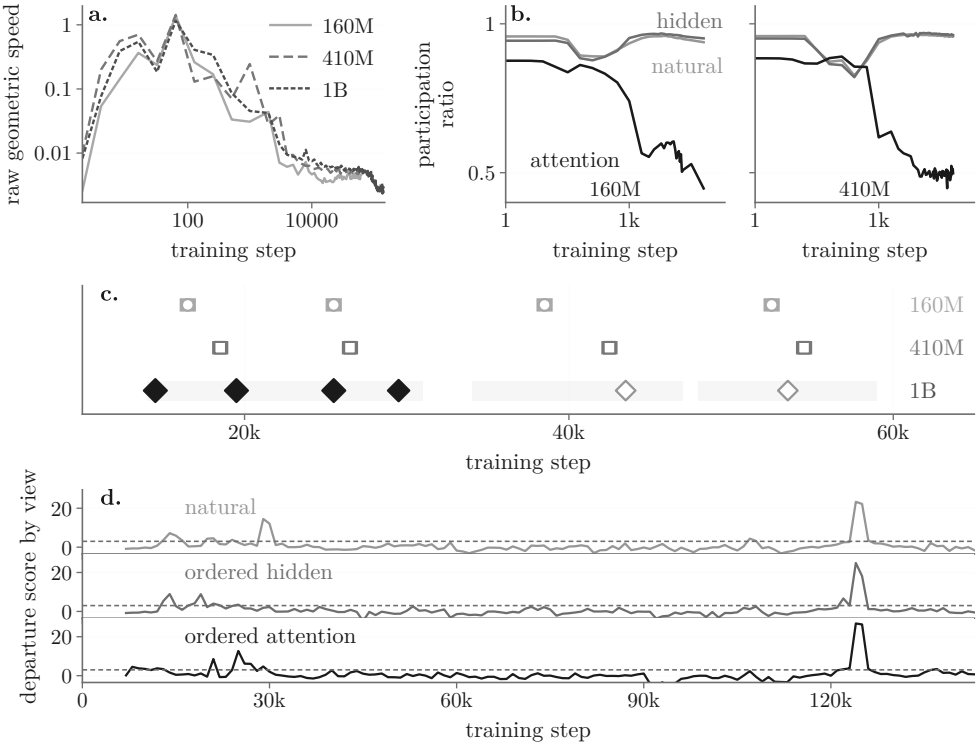

Figure 12: **Pythia audit. a,** Raw relational-attention speeds. **b,** Participation ratios by view. **c,** Smaller-model windows and 1B branches; this is descriptive because coverage is broad and no matched null is available. **d,** Per-view departure records with threshold 3.

H.1 An intrinsic, threshold-free event hierarchy

To measure intrinsic magnitude rather than local surprise, form for each view the Bures shape speed and dimension-normalized soft-support Frobenius speed. The support resolution is 0.1 of the mean eigenvalue; 0.05 and 0.20 preserve all named peak locations. For six positive paths q_a, let $\ell_a = \log q_a$ and define

$$\tau\, \dot{b}_a = \ell_a - b_a, \qquad q_a = e^{b_a}\, e^{\ell_a - b_a}, \qquad E_\tau = \Big(\tfrac{1}{6}\textstyle\sum_a [\ell_a - b_a]_+^2\Big)^{1/2}. \tag{112}$$

Here e^{b_a} is the slow training clock and E_τ the pooled event field (Fig. 11c). Local maxima are ranked by log prominence, invariant to constant score rescaling, without thresholds, window scans, polynomial order, layer quantiles, or a maximum over views. At $\tau = 4{,}000$, dominant 1B pulses are at 9k and 124k, with a second tier at 29k, 55k, 107k, and 136k; all appear in three probe splits. Catalog events at 8k, 29k, 107k, and 122k coincide with coherent pulses. Events from 12k–24k are locally surprising but intrinsically modest; 55k and 136k are intrinsic candidates missed by the surprise detector.

Near-zero recent MAD can inflate unplotted intermediate surprise scores. It does not change event times, and the plotted 1B maximum is 26.8, but surprise magnitude should not be interpreted as geometric size; use the intrinsic hierarchy for magnitude.

H.2 Behavior battery

After the geometric events are fixed, 48 checkpoints are evaluated on next-token loss, repetition gain, induction attention, eight few-shot tasks, parameter norm, and gradient norm. Because the 45 same-cadence intervals concentrate near events, we report global ranks among them and local ranks within contiguous blocks. The 24k event has the largest induction-attention change, with the opposite transferred sign; the 122k peak has the largest local few-shot and gradient changes. The 107k event remains unresolved. External head analyses are context only (Yin & Steinhardt, 2025). Figure 13 shows the raw 1B trajectories around the 24k and 122k events.

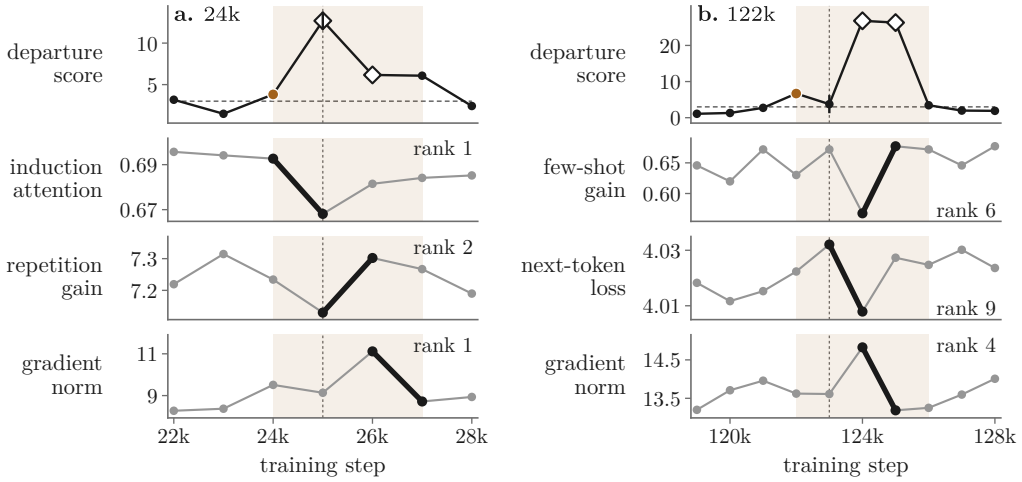

Figure 13: **Behavior around Pythia events.** Raw 1B trajectories around 24k and 122k show readouts changing at different checkpoints within one geometric interval. Labels rank changes among 45 same-cadence intervals.

I Validation, computational cost, and reproducibility

This section records numerical checks of the geometry and response machinery on controlled inputs, the cost of the observer, and the conventions used in the figures.

I.1 Controlled validation of the geometry and response machinery

Table 6 lists the checks. They verify the numerical construction on controlled inputs and are not claims about training.

I.2 Computational cost and implementation

The observer is inexpensive relative to training because, once representations have been extracted on a fixed probe bank, all geometric calculations take place in probe space rather than parameter space. Let n be the number of probes in one observation bank, d the extracted feature dimension, L the number of monitored layers, B the number of observation views, and w the longest continuation window. Table 7 gives the cost of each step for one layer and one bank.

Over all views and layers, the post-processing cost at one checkpoint is

$$O\big(BL\left[n^2d + n^3 + w^2n^2\right]\big)\,, \tag{113}$$

up to the small constant number of probe splits. The dense linear algebra depends on the number of probes, not on the number of model parameters: with n and w fixed, it grows only linearly with feature width and the number of monitored layers, and model size enters through the forward passes on the probes. For Pythia, 96 fixed probes in three interleaved banks give $n = 32$ and $n^3 = 32{,}768$, negligible next to a language-model forward pass; even the unsplit 96-probe Gram

Table 6: **Controlled numerical checks.**

Check	Result	What it establishes
Pure prescribed paths, 64 mixtures	Tangent reconstruction maximum relative error 1.18×10^{-15}; zero numerical cross terms	Each pure motion activates only its own term of Eq. (4)
Repeated eigenvalues	Maximum equivariance error 7.2×10^{-16}; within-block basis rotations contribute no internal rotation	The decomposition does not depend on the basis chosen
Controlled flows	Every block speed in Eq. (57) reconstructed with zero recorded relative error; randomized trace estimates have maximum relative error 0.0258	The response map is computed correctly
Planted transport	Transport singular vectors recover the planted source and destination; a perturbation along the predicted destination changes its coordinate by 0.20 versus 0.04 for a matched control	The controlled response map identifies where a response goes
Innovation and cadence	Innovation error fits Eq. (73) with $R^2 = 0.992$; optimal spacing exponent 0.452 versus the predicted $1/2$	The cadence law holds
Conditioning	The residual is recovered at 1.00–1.54 times the combined leakage-and-noise scale; maximum cross-curve standard deviation 0.300, above the 0.10 collapse criterion	Recovery is stable only up to constants

Table 7: **Cost of the observer per checkpoint, for one layer and one observation bank.**

Step	Cost
Gram matrix of the centered features	$O(n^2 d)$
Normalization $\rho = K/\operatorname{tr} K$	$O(n^2)$
One eigendecomposition, reused for the Bures factor, the soft support $P^\tau(K) = K(K+\tau I)^{-1}$, and the motion decomposition	$O(n^3)$
Procrustes alignment of consecutive principal factors	$O(n^3)$
Constant and linear continuation fits over w tangents, with residuals	$O(w^2 n^2)$
Memory for the recent w Gram states	$O(wn^2)$

matrix has $n^3 < 10^6$. The dominant observer-side step is therefore usually the Gram contraction itself. Representations need not be stored across checkpoints, so observer memory is $O(BLwn^2)$ rather than the $O(BLnd)$ needed to keep every activation tensor.

Evaluated once every c training steps, with C_{train} the cost of one step and C_{probe} the cost of the fixed-probe forward passes, the fractional overhead is approximately

$$\frac{C_{\mathrm{probe}} + O\big(BL[n^2 d + n^3 + w^2 n^2]\big)}{c\, C_{\mathrm{train}}}. \tag{114}$$

It falls with fewer probes, fewer monitored layers, or a lower cadence, without changing the construction. In the Pythia observer, natural, repeated, and permuted inputs need separate forward passes, but the hidden-state and attention views come from the same passes. The experiments need neither parameter Jacobians nor backpropagation through the observer. The response-operator analysis of Appendix B.4 is optional: a realized response needs one Jacobian–vector product, and stochastic estimates of response capacity need vector–Jacobian products, but no dense Jacobian is formed.

Algorithm 1 gives the generic online construction. The observation view and the no-change family change between experiments; the computational skeleton does not.

Algorithm 1 Causal geometric observer

Require: Checkpoints $\theta_{t_1}, \theta_{t_2}, \ldots$; fixed probes $\mathcal{X}$; monitored layers $\mathcal{L}$; views $\mathcal{B}$; probe splits $\mathcal{S}$; continuation windows $\mathcal{W}$; threshold h; persistence n_{pers}

Initialize an empty causal history for every (b, ℓ, s)
for checkpoint t_j **do**
  Evaluate the model on the fixed probe inputs with gradients disabled
  for view $b \in \mathcal{B}$ and layer $\ell \in \mathcal{L}$ **do**
    Construct the feature matrix $H_j^{b,\ell}$
    Center the features across the fixed probes
    for probe split $s \in \mathcal{S}$ **do**
      $K_j^{b,\ell,s} \leftarrow d_b^{-1} H_j^{b,\ell,s} H_j^{b,\ell,s\top}$
      $\rho_j^{b,\ell,s} \leftarrow K_j^{b,\ell,s} / \operatorname{tr} K_j^{b,\ell,s}$
      $\tau \leftarrow 0.1 \operatorname{tr} K_j^{b,\ell,s} / n$
      $P_j^{b,\ell,s} \leftarrow K_j^{b,\ell,s} (K_j^{b,\ell,s} + \tau I)^{-1}$
    end for
  end for
  for each stored path $x_j \in \{\rho_j, P_j\}$ **do**
    Align x_j with x_{j-1} in the quotient geometry
    Construct the backward finite-checkpoint tangent $\dot{x}_j$
    for $w \in \mathcal{W}$ **do**
      Fit constant and linear continuations using only $\dot{x}_{j-w}, \ldots, \dot{x}_{j-1}$
      Compute the quotient-norm continuation residual
    end for
    Compute distance from the span of recent transported tangents
    Standardize all residuals using only observations before t_j
    $I_j[x] \leftarrow$ maximum standardized continuation departure
  end for
  Aggregate splits, layers, and fixed views into causal score s_j
  if $s_j \geq h$ for n_{pers} consecutive checkpoints **then**
    Confirm an alarm at t_j
  end if
  if $s_j < h$ after an active episode **then**
    Rearm the monitor
  end if
  Retain only the history required by the largest continuation window
end for

I.3 Reproducibility

Appendices A.3 and D–H define every score, threshold, persistence rule, calibration split, and control. Representative traces use the lowest-numbered eligible run unless stated otherwise; aggregate panels include all eligible runs. Figures are computed from the recorded runs without retraining or rescoring.