RUNG LOCK: SKILLS ONLY GROW ON THE RIGHT SKILLS

Anonymous authors
Paper under double-blind review

ABSTRACT

During training, a capability that is about to emerge can look exactly like one that will never emerge. Both remain at chance, with no visible sign of what will happen next. We find that some capabilities are locked in this state until several pieces are in place at the same time. They earn no partial credit: if any one piece is missing, the training signal for the capability is exactly zero. The pieces therefore cannot be learned independently and assembled later. Each must help bring the others into place. We tested this by removing from the training data a simpler skill built from the same pieces, while changing nothing else. The target capability did not merely emerge more slowly. It never started. Across nine runs (thirteen with a preregistered extension), its score remained at chance; in two runs, it stayed there for $45\times$ the training normally required for emergence. We then tried to replace the missing skill with an unrelated skill of the same difficulty, training it until it reached perfect accuracy. The target capability remained exactly where it had been without the helper. When we could control the underlying pieces directly, the result became even simpler: a helper's value depended only on how many pieces it shared with the target, not on which pieces they were. A missing rung is not a delay. It is a lock—and the wrong key does nothing at all.

1 INTRODUCTION

The most consequential property of a training run is which capabilities it produces. Scaling laws (Kaplan et al., 2020) forecast the loss to within a few percent, and at fixed corpus and architecture models reaching the same pre-training loss reach the same downstream performance whatever their size (Du et al., 2024)—so the loss looks like a sufficient statistic for what a model can do. It is not one for what a model can still *learn*, and Section 5 shows the sharpest version: the same architecture, size and data acquiring a skill under one batch size and not another. We claim only that the scalar loss does not predict which run will form the skill, since ruling out *every* function of it would need a matched-loss construction we do not have.

The phenomena that make this urgent share one signature: a capability contributes nothing for a long time and then arrives quickly. Capabilities emerge at scale (Wei et al., 2022), and whether that emergence is real or a metric artefact is argued over (Schaeffer et al., 2023); networks grok long after they fit (Power et al., 2022; Nanda et al., 2023); induction heads arrive in a sharp step (Olsson et al., 2022). Existing theory explains single instances with single mechanisms—escape times governed by an information exponent (Ben Arous et al., 2021), sparse functions learned along staircases (Abbe et al., 2022; 2023), progress on parities hidden behind a flat loss curve (Barak et al., 2022). What is lacking is a common object that says which capabilities a dataset contains, prices when each forms, and explains how sharp per-capability events aggregate into the smooth curves we plot.

The natural explanation—that a capability appears once the model is large enough to represent it—sits badly with the observations, since capabilities appear at fixed size when only time increases (Power et al., 2022) and at fixed data when only size increases (Wei et al., 2022). What binds is not expressivity but a race, and what decides the race is not how much a model can express but how cold its run is.

We propose the object that makes the race precise. It is the *interaction spectrum* of the data distribution. Under a product base measure, any target functional decomposes uniquely into orthogonal interaction terms indexed by variable subsets A (Hoeffding, 1948; Efron & Stein, 1981; O'Donnell, 2014). We

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.

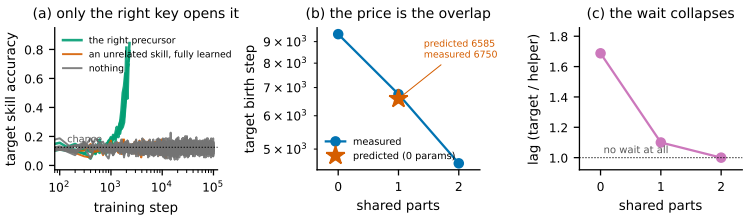

Figure 1: A missing rung is not a delay, and only the right helper pays for it. *(a)* One host, one schedule, three arms: with the target's true precursor the skill ignites at 2000 steps (green); with an unrelated skill of matched difficulty trained to accuracy 1.000 (orange), and with nothing supplied (grey), it stays at chance to 10^5 steps. *(b)* A helper's worth against how many of the target's parts it shares, amplitude fixed; the one-shared-part cell, predicted before it ran (geometric mean 6585, no free parameter), measured 6750. *(c)* The helper-to-target wait, falling to exactly 1 once one part is left.

identify a *capability* with one such term. Its *degree* $K = |A|$ is the number of parts that must be jointly engaged, and its *value* is the term's spectral mass. Its *supply* is the fraction of samples in which all of A is informatively observed. Learning by local updates then *condenses* the spectrum order by order, and

$$\textit{every training curve} = \textit{spectrum} \times \textit{clocks} \times \textit{capacity},$$

where the aggregation layer is, for squared loss, an identity rather than a metaphor, and the per-term clocks are conditional dynamical laws, exact under stated assumptions (Theorem 6, Lemma 1). On this basis threshold collapse, the delay in grokking, the induction-head transition, and the mirage debate stop being separate surprises: all are computable projections of one decomposition.

The mechanics: a capability of degree K earns no partial credit—while any part is unengaged the label signal is exactly zero (Theorem 1), so the parts can only lift one another, each pulled by the product of the others (Theorem 2). The one route past the resulting race is a rung already formed, and Figure 1 is what that route costs on a real model: a helper pays only for the parts it shares, and one that shares none of them pays nothing at all, however well learned.

Contributions.

1. **No partial credit, and the product that replaces it** (Section 2). While any part of a degree-K term is unengaged the label-carrying gradient of a blind parameter channel is *exactly* zero; once the parts are engaged the pull on each is *exactly* the product of the others, for every C^K activation. Both are identities; everything dynamical is separated from them and carries its assumptions explicitly—including why the network escapes the single-unit self-fit obstruction (network dilution, Proposition 2).
2. **The noise is blind to degree; the pull is not** (Proposition 1). Written in the interaction basis the sampling noise is isotropic—every channel receives the same power L/B whatever its degree—while the pull is exponential in it; in parameter space, where prior analyses measure it, it is anisotropic. That single asymmetry is the engine, and it collapses the kinetic layer to one dimensionless group, the *condensation number* Co.
3. **Noise here does not slow a composite skill, it deletes it** (Section 3.2). The target direction is downhill throughout, so fluctuations destroy the parts' coherence rather than surmount a barrier: under plain SGD smaller batches make high-degree terms slower, never faster, and past a computable point stop them. Because L cannot fall below the data's irreducible conditional variance, a run has a coldest state it never leaves—a bound on the *step*, not the summit. The lock has a lower bound—missing-part count linear in the waiting exponent (Theorem 5)—whose linearity makes the spectrum identifiable from arrival times, never from the loss (Proposition 4). The two regimes assemble into a phase diagram whose boundary is computable and whose structure axis only data can move (Section 5.2).
4. **Rung lock, tested causally, and priced** (Sections 5, 5.1). Removing a skill's degree-2 precursor from the data leaves the degree-3 skill at chance in $13/13$ runs across batch

sizes 2–64, two carried to $45\times$ the igniting budget with no partial progress at any point. Substituting an unrelated degree-2 skill of matched difficulty, trained on the same host to accuracy 1.000, leaves it at exactly the peak it reaches with nothing supplied: what is required is not a formed capability but one that shares the target's parts. And where the parts are ours to set, a helper's price depends only on how many of them it shares, not which: equal-overlap routes cost the same within 3%, and the one-shared-part cell, predicted with no free parameter, lands within 2.5%.

5. **Preregistered adjudication throughout** (Sections 4, 5). Every claim above was frozen with its kill condition before the run that tests it, including three that we lost and report as lost.

What is proved, and what is assumed. The layers here do not have the same standing and we mark each claim with its own. Theorems 1 and 2, the loss identity of Theorem 6, and the one-step residual identity are exact; the reduction to one scalar channel is conditional and labelled so; the diffusion rests on the (D), (T), (P) of Section 3, (D) the substantive one; the transfer to a transformer is supported by measurement, not derivation. The chain is *exact identity* $\rightarrow$ *conditional dynamics* $\rightarrow$ *stochastic approximation* $\rightarrow$ *empirical test*, and the causal results of Section 5 stand on the last link independently of how much of the middle survives.

What we do not claim to be first at. That a temperature of order $\eta L/B$ governs SGD (Jastrzębski et al., 2018; Smith et al., 2018; Sclocchi & Wyart, 2024), that a run cools itself (Feng & Tu, 2021), that low-information-exponent targets are hard (Kearns, 1998; Ben Arous et al., 2021; Barak et al., 2022), that learning proceeds along staircases (Abbe et al., 2022; 2023) and that skills come in discrete quanta (Michaud et al., 2023; Arora & Goyal, 2023) are established; we claim none of them. We add the basis in which the noise is isotropic while the pull is exponential in degree, a formation *clock* rather than a learnability threshold, and the causal tests of Section 5 (Appendix A in full).

2 Capabilities as interaction terms: two exact identities

2.1 Setup and definitions

Let x be uniform on $\{-1,1\}^n$ (a product base measure; the real-text extension is discussed in Section 6). For $A \subseteq [n]$ write $\chi_A(x) = \prod_{i\in A} x_i$. Every $f : \{-1,1\}^n \to \mathbb{R}$ has a unique orthogonal expansion $f = \sum_A \hat{f}(A)\,\chi_A$—the Fourier–Walsh representation, which on the Boolean cube under the uniform product measure coincides with the corresponding Hoeffding/ANOVA decomposition (O'Donnell, 2014). We call $\{\hat{f}(A)^2\}_A$ the *interaction spectrum* and $|A|$ the *degree* of a term.

Definition 1 (Capability, relative to a reference measure)**.** *Given a product reference measure μ, a capability of a data distribution is an interaction term $\hat{f}(A)\chi_A$ of its target functional decomposed with respect to μ. Its degree $K = |A|$ is the number of parts that must be jointly engaged; its value is $\hat{f}(A)^2$ (its contribution to the loss); its supply is the fraction of training samples in which all coordinates of A are informatively observed (the* joint-evidence share*).*

In the laboratory μ is the uniform product measure (canonical there); for real text we fix it to the unigram product measure, all real-machine claims relative to that convention (Section 6). Given μ, K is computable from data—not a modeling choice. Throughout, the running target is a pure term $y = \chi_S$ with $|S| = K$; mixed spectra are the subject of Theorem 3.

2.2 Exact gradient blindness

The first fact is elementary, and in substance known: it is the population, exact-zero form of the hardness that statistical-query and information-exponent analyses quantify from samples (Kearns, 1998; Ben Arous et al., 2021; Barak et al., 2022). We state it per parameter and per sample because that is what the rest of the paper needs: it says which *samples* can move which *weights*, and it makes the responsibility identity of Theorem 2 the *only* channel through which a capability can grow.

Theorem 1 (Exact gradient blindness)**.** *Let the loss gradient be affine in the label, $\partial\ell/\partial u = u - y$ plus label-independent terms; squared loss qualifies directly, and for cross-entropy the label-carrying*

component is affine in the target encoding, with $y = \chi_S$ read as the ± 1 contrast between the two logits rather than as a softmax one-hot label. Let θ be any parameter such that $g_\theta(x) = \partial f_{\text{model}}(x)/\partial\theta$ does not depend on some coordinate $i \in S$. Then the label-carrying part of the population gradient vanishes identically: $\mathbb{E}[\, y\, g_\theta(x)\,] = \mathbb{E}[x_i] \cdot \mathbb{E}[\chi_{S\setminus\{i\}}\, g_\theta] = 0$.

Proof. x_i is independent of the other coordinates and mean-zero; g_θ does not depend on x_i, so the expectation factorizes. □

Corollary 1. *(a) The statement is about a parameter* channel*, not a network: any parameter whose Jacobian is blind to a required coordinate receives exactly zero label-carrying population gradient. It does not say the network cannot reach χ_S by another route; what makes it bite is Theorem 2—the channels that do carry label gradient are exactly those whose pull is the product of the remaining amplitudes, so the blindness of the rest leaves that product as the only route. Claims below that a skill "receives no signal" carry that premise. (b) Parameters still move under label-independent noise—the reservoir clock (iii) draws on: blindness concerns signal, not motion. (c) Classifying samples by the subset $B \subseteq S$ they informatively expose and applying the theorem per class, samples with $B \neq S$ carry zero signal for χ_S; the joint-evidence share is then the currency the manufacturing clock reads (Theorem 7).*

2.3 Mutual gating is an exact identity

Blindness says a capability cannot grow through any channel that misses a part. We now prove what the one surviving channel looks like, and the answer is what makes the whole process bootstrap.

Theorem 2 (Mutual gating; exact responsibility identities)**.** *Let $f_S(w) = \operatorname{coef}_S[\sigma(w \cdot x + b)]$ for a unit with $\sigma \in C^K$, and let a be its readout weight, so the signal term of Theorem 1 routes through $V = a\, f_S(w)$. Then, exactly:*

(a) (Coefficient identity.) *$f_S(w) = \mathbb{E}_{x_{S^c}}\big[\Delta_S\, \sigma\big]/2^K = c_K(w) \prod_{i\in S} w_i$, where Δ_S is the K-fold alternating difference over sign flips of x_S and $c_K(w) = \mathbb{E}[\sigma^{(K)}(\xi_x)]$ is an explicit average of the K-th derivative.*

(b) (Responsibility identity.) *For every $j \in S$,*

$$\frac{\partial V}{\partial w_j} = a\, \kappa_j(w) \prod_{i\in S,\, i\neq j} w_i, \qquad \frac{\partial V}{\partial a} = c_K(w) \prod_{i\in S} w_i,$$

with $\kappa_j(w) = \mathbb{E}\big[\sigma^{(K)}(\xi'_{j,x})\big]$ another explicit average of $\sigma^{(K)}$. The prefactors are bounded by $\sup|\sigma^{(K)}|$ and share a common limit: $\kappa_j(w)$, $c_K(w) \to \sigma^{(K)}(b)$ as $\|w\| \to 0$ (the full vector—the averages retain $w_{S^c} \cdot x_{S^c}$, so $w_S \to 0$ alone does not suffice), uniformly in j.

Each part's gradient responsibility equals the product of all other parts' amplitudes, times an explicit bounded prefactor—an exact identity for every $\sigma \in C^K$, not a Taylor truncation. The consequence is the one the title turns on: each part's pull vanishes when any other is at zero, so the parts cannot be learned independently and assembled afterwards. They can only lift one another, and a part supplied from elsewhere is the only outside start. ((b) is not *obtained by differentiating (a) with c_K held fixed—invalid, as c_K depends on w—but by differentiating first, then applying (a) to σ' on $S\setminus\{j\}$; Appendix H.)*

3 Formation kinetics: three clocks and the condensation number

The identities of Section 2 say a capability grows only through a product of its own parts. How fast that product grows, what opposes it, and which wins is a stochastic question, and this section is where the paper stops being exact: the formal statements, proofs and the reduction's ten hypotheses are in Appendix B; what follows is their content and standing.

The clock. A pure degree-K term routes through $K+1$ multiplicatively coupled amplitudes, and under a symmetric initialization these collapse to one amplitude s whose label-attributed channel obeys $\dot{s} = \gamma_0 s^K(1 + O(s))$, $\gamma_0 = \eta|\sigma^{(K)}(b)|$ (Lemma 1, *conditional*). Three regimes follow

(Theorem 4): a degree-1 channel rises exponentially, $t^* = \gamma_0^{-1} \ln(\theta/c)$, as does the last part of a higher-degree term once the others are mature and it has an $O(1)$ drive; a drift-dominated term of degree $K \geq 2$ escapes by the power law $t^* = (c^{1-K} - \theta^{1-K})/\gamma_0(K-1)$, the slow saddle (cf. Saxe et al., 2014; Abbe et al., 2023); and when fluctuations dominate, the joint activation of K small amplitudes is a rare event, geometric with $\mathbb{E}[T] = p_0^{-K}$ under independence across factors and time steps—memoryless, CV $\to 1$, exponential in the number of missing parts. Which regime a run is in is empirical; the laboratory shows the high-K regime carrying the third's signatures.

A prediction of ours that failed. We froze the single-unit comparison as a kill condition—that a self-fit contraction should beat the signal at $K{=}3$ under $\tanh(\beta\,\cdot)$—and a preregistered β-sweep falsified it: the class stays near-deterministic down to $\beta{=}0.25$ while the clock slows as $\beta^{-2.5}$, confirming the prefactor $\gamma_0 \propto \sigma^{(K)}$ but not the competition. What survives decisively is the identity's own content—a degenerate bias kills the $K{=}2$ clock outright, $6/6$ censored against births at 500–750—and the failure is why the reduction is labelled *conditional* (Appendix B).

Why the network is not the unit. The contraction that wins in a single unit lives in the low-order channels, and in a network those are *fast* variables: if the low-order Gram block keeps $\lambda_{\min} \geq \lambda_0 > 0$ (H1), time scales separate (H2), and the low-order-to-target cross-block of Θ is small (H3—strictly weaker than (D)), the fast channels relax to a quasi-steady $O(s^K/\lambda_0)$ and the self-fit force falls from the unit's $\Theta(s^3)$ to $O(s^{K+1})$: one order *below* the signal (Proposition 2, outline standing). The unit is the $\lambda_0 = 0$ degeneracy, the narrow-width failures of Appendix B are (H1) failing where it must, and under SGD the fast channels hold the mean-zero OU plateau of Proposition 1(a2)—the slow drift unchanged, as the exact-enumeration probe measured.

The noise, and the three assumptions that carry it. In the Walsh basis one SGD step is $\delta\hat{f} = -\eta\Theta\hat{r}_{\text{emp}} + O(\eta^2)$, and because $\chi_A^2 \equiv 1$ the residual spectrum has $\text{Var}[\hat{r}_{\text{emp}}(A)] = (L - (\hat{f} - \hat{y})(A)^2)/B$. That is exact and it is all that is exact. The clean form L/B—the same for every channel whatever its degree, which is the isotropy the paper turns on—is a *zero-target* reading, and it becomes isotropy of the *update* only insofar as Θ acts diagonally in the Walsh basis: true of the static NTK (Jacot et al., 2018; Yang & Salman, 2019), not proved of the finite-width feature-learning Θ here—assumption **(D)**. Turning the recursion into a diffusion needs **(T)**—under i.i.d. sampling the increments are martingale differences by construction, so (T) is only the diffusion-limit freezing step (Ethier & Kurtz, 1986) the third clock also needs—and carrying the noise to a *parameter* needs the projection $x_j\sigma'(z)$ decorrelated from the residual, **(P)**, which pre-condensation discharges ($x_j^2 \equiv 1$ factors the joint moment; Appendix B); the multiplicative form is the parameter's own Jacobian, $\partial f/\partial w_j = a\sigma'(z)x_j$ with $a = s$ in the symmetric channel. Under (D), (T), (P),

$$ds = \eta\gamma s^K\,dt + \eta\,\sigma'(b)\,s\sqrt{L/B}\,dW, \qquad \text{Co} = \frac{\gamma c^{K-1}}{\sigma'(b)^2\,\mathcal{T}} = \frac{\gamma B c^{K-1}}{\eta L\,\sigma'(b)^2}, \qquad \gamma = |\sigma^{(K)}(b)| = \gamma_0/\eta,$$

and the scale function of *that* diffusion is exact within the diffusion approximation, whose measured domain is the width sweep of Appendix B. We call Co the *condensation number*. Birth is an extreme value—born once *any* of d units escapes—so the network transition sits at $dp \approx 1$: no single critical batch size, only a curve (derivations and floor exponents in Appendix B).

3.1 The same kinetics under Adam

Adam divides by $\sqrt{\hat{v}}$, the second moment of the gradient this coordinate receives—and that gradient carries its own Jacobian, $\partial f/\partial w_j = a\sigma'(z)x_j$ with $a = s$ in the symmetric channel: signal $O(s^K)$, fluctuation $s\sigma'(b)\sqrt{L/B}$. So $\hat{v}$ is fluctuation-dominated and $R = \sqrt{\hat{v}} \propto s\sqrt{L/B}$, not a constant. Under **(A)**, that this normalizer is the whole of Adam's effect, dividing drift and noise by R gives $ds = \eta\,g\,s^{K-1}\sqrt{B/L}\,dt + \eta\,dW$ with $g = (\sigma^{(K)}/\sigma')(b)$: the drift exponent drops by one and the noise becomes *additive*. Two earlier versions of this derivation—noise steeper; SGD exponent kept—were wrong, and the correction turns two measurements into predictions. Birth scales as $B^{-1/2}$ either way, so the measured slopes ($-0.450, -0.413, -0.448$ at $K = 3, 4, 5$; SGD: $0.023, 0.021$) fit both forms. What distinguishes them: additive noise does not die with s, so Adam rarely censors ($3/12$ cells against SGD's 12); and the saddle order drops by one, so $K{=}2$ and 3 arrive at the same step under Adam where SGD separates them. The additive form has its own boundary, further

out: on a denser ladder Adam is uncensored through $K{=}7$ ($8/8$ born, CV 0.22) and turns lottery at $K{=}8$—$2/8$ born at 3.7 and 6.6×10^4 steps, the rest unborn at 4×10^5 and still unborn when rerun to 10^6 (Appendix C): the lottery region is displaced by about two rungs, not deleted. **The deletion results below are established under SGD, and the corrected form says why: deletion is a multiplicative-noise phenomenon, and Adam's normalizer divides that factor out.**

3.2 What noise can and cannot do

In the drift-dominated branch noise can only slow formation, for a structural reason: along the direction that grows a capability there is *no barrier*, since the drift $\eta\gamma s^K$ points at the target for every $s > 0$. What noise does instead is multiply s by a random factor of size $\eta s\sqrt{L/B}$, randomizing the K parts independently while the pull requires them aligned at once; the pull grows as the Kth power and the shaking only as the first, so noise dominates exactly where the process spends nearly all its life. The prediction is opposite in sign to the escape-from-saddle picture: smaller batches make high-degree terms slower, never faster, and past a computable point stop them. Barak et al. (2022) showed parity learning is not a Langevin-like search; the condensation number says where the search is killed. Abbe & Sandon (2020)'s noise-as-resource result is constructive—an architecture built to decode the fluctuation—where ours concerns generic initializations with no decoder. In the fluctuation branch only a fluctuation ever completes the joint event, but it delivers a memoryless rare event, not an acceleration: neither branch is noise helping in the sense the SGD literature means.

Three preregistered measurements agree, all under plain SGD: shrinking B at $K{=}4$ over $64, 32, 16, 8, 4$ moves censoring from $0/4$ to $4/4$, with no run forming within $300{,}000$ steps at $B \le 8$; label noise independent of the target, which by Theorem 1 leaves the drift untouched, censors $4/4$ a term that otherwise forms in 3875 steps, every run given $77\times$ that budget, with quadrupled variance leaving $3/4$ censored and a single escape at $165{,}250$ steps; and lowering the target amplitude, which raises Co without touching the noise, rescues a censored configuration.

Capabilities couple across disjoint variables, but not through the bath. Since the per-channel variance is the same L/B everywhere, an unlearned companion should heat its neighbours; we ran the decisive test, holding the target's supply fixed and varying only a companion's residual. A companion on a *disjoint* variable set accelerates a degree-3 target $3.9\times$ (9375 against 36250 steps, 12 seeds), but the sign is wrong for a bath: the faster arm is the *hotter* one ($L = 1.361$ against 1.000 through the wait), and a dose–response over amplitudes 0.2–0.9 moves birth $44125 \to 7875$ while the pre-birth residual rises monotonically. The cause is Theorem 2's shared readout factor (1.51 against 0.06 at step 10^4). Earlier drafts had the mechanism and the sign wrong.

4 From single terms to the loss curve, and the staircase

For squared loss the total is a sum over interaction channels (Theorem 6, Appendix B), so a metric resolved by capability must look discontinuous while a token-averaged one looks continuous—the two sides of the emergence-versus-mirage debate are two layers of one identity. Only samples exposing every coordinate of S carry signal for χ_S, so the manufacturing clock scales as $1/f$ in the joint-evidence share and no "parts-first" curriculum contributes any signal to a pure degree-K term (Theorem 7).

Staircase catalysis (Theorem 3, Appendix B). For $y = \alpha\chi_B + \beta\chi_S$ with $B \subset S$, once the low-order term has condensed—which we assume rather than derive, since a small enough α^2 reverses the ordering—its parts are mature, and by Theorem 2 the missing-part product for χ_S runs over $S \setminus B$ only: the effective degree drops from $|S|$ to $|S \setminus B|$. Within a mixed spectrum the reuse is automatic, the same unit's product benefiting; catalysis *across* separate tasks additionally needs a representation-reuse assumption.

This is the theoretical content of rung lock: a formed sub-term is not a hint or a warm start—it removes a factor from the product that drives every remaining part, and the clock's exponent falls with it. Everything Section 5 measures (the price set by shared parts, the zero of a partless helper, the vanishing wait at one part left) is this theorem read at three values of $|S \setminus B|$.

The laboratory. Everything above is checked where it is computable exactly: $n{=}16$ Boolean inputs, enumerated, so the instrument adds no noise of its own. One capability followed act by act matches

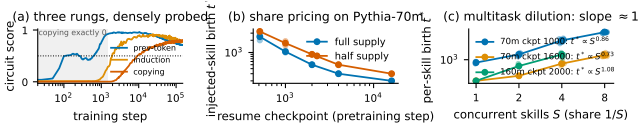

Figure 2: Real-machine correspondence on Pythia. (a) The three-rung ladder—previous-token, induction, behavioural copying—in that order at every size and in three training families; in the densely probed run the successor holds at its untrained value while its precursor matures. (b) Injected-skill birth vs. checkpoint at full/half share. (c) S distinct skills on one budget: per-skill birth $\propto S^{0.86}/S^{0.73}/S^{1.08}$, in Theorem 8's band.

the identities term by term, and birth medians across $K = 2, 3, 4, 5$ place $\ln t^*$ on *two* segments—flat where the drift clock rules, exponential where the joint event rules—whose exponential segment the real machine reproduces. Three preregistered kill shots hold: the clock reads spectral degree, not surface variables ($300\times$); birth times reorder by spectral mass alone ($8/8$); a mixed spectrum generates its own staircase. Nothing in Section 5 rests on it (Appendix C).

5 Real-machine correspondence

The laboratory laws should surface wherever their projection is measurable; all statements here are empirical laws under the unigram convention of Definition 1, tested prospectively.

Empirical Finding 1 (Precursor order and the dead window; Figure 2a)**.** *Three circuits—previous-token attention, induction attention, and behavioural copying—arrive in that order, without exception, on Pythia (Biderman et al., 2023) at* 70*m,* 160*m,* 410*m,* 2.8*b and* 6.9*b, on OLMo-2-1B (Groeneveld et al., 2024), and on a LLaMA-style* 70*m model we train ourselves. Three separations need no interpolation, one stored checkpoint having a circuit past threshold and the next at its untrained value: Pythia-2.8b step* 512 *(*0.549 *vs* 0.030*), Pythia-410m step* 1000 *(*0.91, 0.90 *vs* 0.48*), and, in another family on other data, OLMo-2-1B step* 300 *(*0.605 *vs* 0.013*). In the self-trained run, probed every* 25 *steps, previous-token crosses* 0.5 *at* $\approx$ 100*, induction at* $\approx$ 1900 *and behavioural copying at* $\approx$ 9200*; the induction score holds within noise of its untrained value through step* $\approx$ 600*, and behavioural copying is* exactly 0.0000 *until step* 1400*, while the circuit beneath it has already reached* 0.92*. The waiting is not slow growth.*

The zero is the point, and it is what makes the two cases indistinguishable: a capability whose precursor is still forming and one whose precursor is absent from the data both sit at the untrained value until the first of them moves—the signature Theorem 1 predicts for a term with an unengaged part, and what separates rung lock from a slow clock. The induction circuit is a degree-2 composition whose part is the previous-token head, so the theory expects exactly this shape, and the ordering extends past induction to the behavioural rung above it, where circuit-specific accounts (Olsson et al., 2022; Musat et al., 2025; Aoyama et al., 2025) do not reach.

The ratio between arrivals is not a constant. An earlier version reported $t^*_{\text{ind}} = 1.241\, t^*_{\text{prev}}$; on public checkpoints both crossings fall inside Pythia's single 512–1000 storage gap, so that number was interpolation output—the one resolved run gives 18. What the theory predicts, and what we report, is the order and the zero, never a universal constant (in full: Appendix C).

Empirical Finding 2 (Rung lock, tested causally)**.** *On Pythia-70m resumed at checkpoint* 1000 *we inject a degree-3 skill and vary one thing: whether its degree-2 precursor is present in the data. With the precursor, the degree-3 skill ignites at* 2100 *steps at batch* 16 *and at* 5650 *at batch* 4*. With the precursor removed and nothing else changed, it stays at chance across batch sizes* $2, 4, 8, 16$ *in* $9/9$ *runs; two runs were carried to* 10^5 *steps,* $45\times$ *the budget in which the same skill ignites when the precursor is present. Peak probe accuracy over all nine runs is* 0.18–0.23*, against a chance level of* 0.19 *and a with-precursor level of* 0.59–0.79*. A separately preregistered extension added two runs at each of batches* 32 *and* 64*,* 12,000 *steps each—*8–11$\times$ *the ignition expected at those batches under the* $B^{-1/2}$ *clock of Section 3.1—and all four locked: over the two grids together,* $13/13$ *runs across a* $32\times$ *batch range (per-cell budgets in Appendix C)—the axis asymmetry of Section 5.2, read causally.*

Removal does not slow the skill, it stops it: the no-precursor runs receive more budget, not less, register no partial progress, and do not creep—Theorem 1's signature for a term none of whose parts are engaged. As an intervention on the training data rather than an ablation of a trained network, it establishes that the precursor is required for the higher term to *form*, not merely correlated with it afterwards: the state Definition 2 names.

Empirical Finding 3 (The currency is specific; matched control)**.** *Run in one block on the same host, batch and schedule as Finding 2. In place of the true precursor we supply a degree-2 skill drawn from a* disjoint *symbol alphabet: same arity, same answer space, same episode rate, sharing nothing with the target but the shape of the task. The unrelated skill itself reaches accuracy* 1.000*—a fully formed capability, not a failed injection. The degree-3 skill nevertheless peaks at* 0.180 *in* $3/3$ *runs carried to* $12{,}000$ *steps, six times the budget in which the true precursor ignites it—and* 0.180 *is also the peak it reaches when nothing is supplied at all. The matched positive control in the same block ignites at* 2000 *steps and reaches* 0.695 *and* 0.789*.*

This is the control the causal claim needs: warm-up, general competence and more useful optimization all survive removal of the precursor but not this arm—the unrelated skill supplies all three in full, and they buy nothing. What the degree-3 skill waits for is not a formed capability; it is a formed capability that shares its parts.

What a helper is worth, priced (preregistered). In the laboratory the parts are ours to set, so overlap can vary at fixed helper amplitude. Against a degree-3 target we plant a degree-2 term sharing 0, 1 or 2 of its variables; median births over 12 seeds each are 9375, 6750 and 4625 steps. Two *different* two-variable routes cost the same to within 3% (4625 against 4750), so the price is set by how many parts are shared and not by which. The one-shared-part cell was predicted from the other two before it was run—their geometric mean, 6585, with no free parameter—and measured 6750, within 2.5%. The midpoint is the memoryless-branch prediction: waits p^{-3}, p^{-2}, p^{-1} for $0, 1, 2$ shared parts put the middle cell at exactly the geometric mean of the outer two; the adjacent ratios 1.39, 1.46 estimate one p^{-1}. The helper-to-target lag falls with overlap, 1.69, 1.10, 1.00: with two parts supplied the last has an $O(1)$ drive and arrives in the same probe as its helper—clock (i), not a second wait.

Why an unrelated skill is worth $3.9\times$ **there and nothing here.** It buys $3.9\times$ in the laboratory through the shared readout factor rather than overlap—a channel open only while the substrate is small. Pythia at checkpoint 1000 has already paid that factor, leaving the specific parts as the only remaining variable.

Three further cells (Appendix C): five runs with different injected calibrators spread 12% (1338–1500 steps)—the natural clock's honest precision, Finding 3 from the other side; the degree-2 batch sweep slope -0.689 misses the frozen band by 0.04, recorded undecided; and the task-attributed gradient jumps $\times 5.7$–6.6 at birth against $\times 1.5$ for a contaminated channel.

5.1 An irreducible-noise floor on raw capability degree

Because L cannot fall below the data's irreducible conditional variance σ^2_{irr}, the temperature obeys $\mathcal{T} \geq \eta\sigma^2_{\text{irr}}/B$ *for the entire run*, and a degree-K term is confined to the fluctuation branch for the rest of it once $\gamma_K c^{K-1} \lesssim \sigma'(b)^2\, \eta\sigma^2_{\text{irr}}/B$. The inequality, not a closed form, is the claim: it collapses to a threshold in K alone only if γ_K is K-independent (Appendix B). Past that point a term changes branch rather than slowing, and budget becomes the wrong axis.

A condensed term is itself a part—if a degree-m term on $S' \subset S$ has formed then $\chi_S = \chi_{S'}\chi_{S\setminus S'}$ needs only $K - m + 1$ things at once (Theorem 3)—so the floor caps the step, not the summit: complexity grows one rung at a time.

Definition 2 (Rung lock)**.** *A term* χ_S *is* rung-locked *at temperature* $\mathcal{T}$ *if its condensation number lies in the fluctuation-dominated branch at every sub-term that has actually formed. It is a property of the pair (data, run), not the architecture: relieved by lowering* $\mathcal{T}$*, raising supply, or forming a sub-term that reduces the parts that must coincide—not by more steps at the same temperature, because* $\mathcal{T}$ *has stopped falling. "Locked" is operational: the branch's rate is a rare-event rate, small but not zero (clause (iii)); we establish that it sits below every budget we could reach (*$45\times$*, no partial progress), not that it is exactly zero.*

The bound behind that exponent: modelling the K unengaged parts as the stationary zero-target channels of Proposition 1(a2), the first time all K sit above threshold *simultaneously* obeys $\mathbb{E}[T] \geq$

$\tau \exp\left(K[\theta^2/2\sigma_{\rm st}^2 - c_1\theta/\sigma_{\rm st} - \ln 2]\right)$ (Theorem 5, by Borell–TIS)—the measured ladder—while a single channel's escape probability is only polynomially small: the wait belongs to the joint event.

Rung lock is what makes the sequence strict rather than merely typical: a rung that has not formed does not make its neighbour slow, it leaves the neighbour on the branch where waiting is the wrong instrument (Finding 2 exhibits the state). We stop short of claiming arbitrary degree attainable—rungs occupy hidden units (Proposition 5), and whether spent rungs can be reclaimed we do not settle. Three consequences, in increasing practical bite: compute did not buy a skipped rung at any budget we reached; what binds is usually not the model's capacity but whether the data contains the intermediate rungs; and interventions that insert rungs—curricula, intermediate supervision, reasoning traces—should pay off out of all proportion to the loss they remove.

The floor's mechanism has been run, not only stated: label noise independent of the input *is* an irreducible conditional variance, so Section 3.2(ii) is the floor experiment itself—the degree-4 term absent in 7 of 8 runs at $77\times$ budget, the single exception at $43\times$. What we have *not* shown is that the floor binds where it would matter—whether natural text sits high enough to exclude a degree anyone cares about is open: the laboratory shows the mechanism, not the magnitude.

5.2 The two axes, assembled: a formation phase diagram

A capability's formation is positioned by two coordinates—a *structure* axis, the effective degree $K_{\rm eff}$ (how many parts the data still owes it), and a *temperature* axis $\mathcal{T} = \eta L/B$—and the plane has two regions separated by the computable line $\mathrm{Co} \approx 1$: clocked arrivals on the cold side (CV 0.15–0.3, no censoring), lotteries on the hot side (CV$\to 1$, censoring, no progress variable—no rising precursor in any of 2^{20} exhaustively catalogued subset channels, twelve locked waits). Three measured facts shape the plane (Appendix C): time does not distinguish the regions—t^* grows exponentially along K on both sides, its jump at the same K across a $16\times$ batch range. Determinism does—the first censored degree moves $4 \to 6 \to$none over $B = 16/64/256$, likewise in η, and the censored cells collapse to $\mathrm{Co} \in [0.11, 0.42]$. And the axes are not symmetric: optimizer, batch and learning rate all move the temperature coordinate—Adam pushes the boundary from $K^*{\approx}4$–6 to $K^*{\approx}8$—but none moves the structure coordinate: the same optimizer leaves a skill with a missing rung at chance in $13/13$ runs on the real model. Compute chooses how a capability is waited for; only data chooses what is waited for. Rung lock is the structure axis, and Findings 2–3 are its causal reading.

6 Scope, limitations, outlook

Every exact identity is proved on $\{-1, 1\}^{16}$ with one hidden layer and a linear readout; nothing is proved about a transformer, and what transfers is the *form* of the laws and the *direction* of each knob. Degree on real text is convention-dependent—no canonical product reference is known—and the aggregation layer needs a spectrum we cannot measure at scale. Of the named assumptions, (T) is a standard freezing step and (P) discharges pre-condensation (Appendix B); (D), (A) and the dilution outline's (H1)–(H4) remain assumed; two constants are fitted; minimality of the hidden cost is proved only at $K{=}2$. Deletion under Adam is predicted absent but untested, and the floor is stated, not established, each with its kill measurement; a third—bath heating—was tested and did not survive (Section 3.2).

7 Conclusion

Skills only grow on skills, and only on the right ones—the second half earned by experiment: a fully formed capability of matched difficulty sharing none of the target's parts buys exactly what nothing buys. The reason is one identity: no partial credit, so the parts can only lift one another, each pulled by the product of the rest—a race set by an exponent. Temperature alone deletes a degree-4 skill at $77\times$ the budget; three circuits of a real model form in strict order, the successor at exact zero until its precursor is there. And the barrier's linearity in K makes the spectrum measurable: the loss resolves $O(\log(1/\epsilon))$ parameters; arrival times, their rates separated by that barrier, determine it uniquely (Proposition 4)—what the loss can never tell you, arrival times can.

Reproducibility Statement

All laboratory experiments run in minutes-to-hours on CPU; real-machine runs use one consumer GPU on public checkpoints. Every experiment was preregistered in an append-only ledger with frozen predictions and kill conditions, included with the code and raw results.

AI Use Statement

In accordance with the ICLR AI policy, we disclose the following uses of large language models: (i) implementing experiment drivers and analysis scripts to the authors' specifications—all LLM-assisted code was reviewed, tested, and verified for correctness by the authors; (ii) processing and summarizing experimental results—every reported number was checked against the raw experiment logs included in the supplement. LLMs were not used to fabricate or select results: every preregistered criterion was frozen by the authors before the corresponding data existed, and censored or failed cells are reported as such. The authors take full responsibility for the entire content of the paper.

References

Emmanuel Abbe and Colin Sandon. Poly-time universality and limitations of deep learning. In *Advances in Neural Information Processing Systems*, 2020. Also Comm. Pure Appl. Math. 2023.

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory (COLT)*, 2022.

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory (COLT)*, 2023.

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer, 2007.

Tatsuya Aoyama et al. Predicting the emergence of induction heads in language model pretraining. *arXiv preprint arXiv:2511.16893*, 2025.

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Online learning and information exponents: on the importance of batch size, and time/complexity tradeoffs. In *International Conference on Machine Learning*, 2024.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

Boaz Barak, Benjamin L. 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.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: a suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: a memory viewpoint. In *Advances in Neural Information Processing Systems*, 2023.

Blake Bordelon and Cengiz Pehlevan. Learning curves for SGD on structured features. In *International Conference on Learning Representations*, 2022.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25, 2024.

Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. Discovering and explaining the representation bottleneck of DNNs. In *International Conference on Learning Representations*, 2022.

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.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *Annals of Statistics*, 9(3): 586–596, 1981.

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

Charles L. Epstein and John Schotland. The bad truth about Laplace's transform. *SIAM Review*, 50 (3):504–520, 2008.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

Yu Feng and Yuhai Tu. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: accelerating the science of language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Xinran Gu et al. Data mixing can induce phase transitions in knowledge acquisition. *arXiv preprint arXiv:2505.18091*, 2025.

Tom Henighan, Shan Carter, Tristan Hume, Nelson Elhage, Robert Lasenby, Stanislav Fort, Nicholas Schiefer, and Christopher Olah. Superposition, memorization, and double descent. *Transformer Circuits Thread*, 2023.

Tom Heskes. Bias-variance decompositions: The exclusive privilege of bregman divergences. *arXiv preprint arXiv:2501.18581*, 2025.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. The developmental landscape of in-context learning. *arXiv preprint arXiv:2402.02364*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. In *International Conference on Artificial Neural Networks*, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, 2023.

Clare Lyle et al. What can grokking teach us about learning under nonstationarity? *arXiv preprint arXiv:2507.20057*, 2025.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, 2023.

Tiberiu Musat, Tiago Pimentel, Lorenzo Noci, Alessandro Stolfo, Mrinmaya Sachan, and Thomas Hofmann. On the emergence of induction heads for in-context learning. *arXiv preprint arXiv:2511.01033*, 2025.

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.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Hannah Pinson. It's not a lottery, it's a race: Understanding how gradient descent adapts the network's capacity to the task. *arXiv preprint arXiv:2602.04832*, 2026.

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.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 2019.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024.

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.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

Antonio Sclocchi and Matthieu Wyart. On the different regimes of stochastic gradient descent. *Proceedings of the National Academy of Sciences*, 121(9), 2024.

Lee Sharkey et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Jamie Simon et al. There will be a scientific theory of deep learning. *arXiv preprint arXiv:2604.21691*, 2026.

Aaditya K. Singh, Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, 2023.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018.

Henry Teicher. Identifiability of finite mixtures. *Annals of Mathematical Statistics*, 34(4):1265–1269, 1963.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

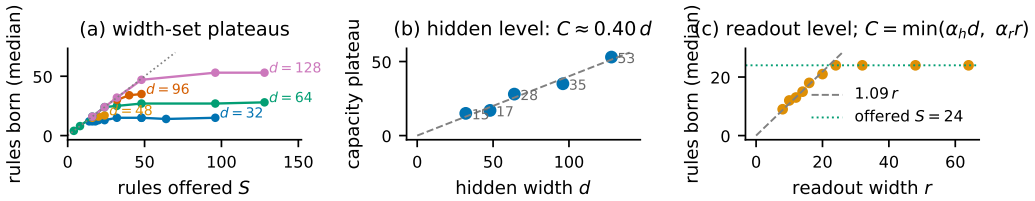

Figure 3: The two-level capacity law. (a) Rules born vs. rules offered: width-set plateaus. (b) Plateau vs. hidden width, at $K{=}2$ with $\tanh$: $C \approx 0.40\,d$, to be read as $d/m_\sigma(K)$ with $m_{\tanh}(2) \approx 2.5$ (Proposition 5). (c) Readout constraint at $d{=}128$: $C \approx 1.09\,r$ until the offered ceiling; jointly $C = \min(\alpha_h d, \alpha_r r)$.

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.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

Junpeng Zhang, Qing Li, Liang Lin, and Quanshi Zhang. Two-phase dynamics of interactions explains the starting point of a DNN learning over-fitted features. In *International Conference on Machine Learning*, 2024.

Hanxu Zhou, Zhou Qixuan, Tao Luo, Yaoyu Zhang, and Zhi-Qin John Xu. Towards understanding the condensation of neural networks at initial training. In *Advances in Neural Information Processing Systems*, 2022.

A Related work in full

Single-structure learning theorems. Escape times via the information exponent (Ben Arous et al., 2021), staircases (Abbe et al., 2022; 2023) and parity hardness (Barak et al., 2022; Kearns, 1998) each characterize one mechanism on one target; we assemble them into a spectrum-level dynamics.

Discrete-skill accounts of scaling and emergence. Quantization (Michaud et al., 2023) posits quanta learned in order of use; skill graphs (Arora & Goyal, 2023) posit compositional skills. Definition 1 supplies the microscopic object and the clocks the *when*; frequency-ordering is the share law at equal degree. The emergence debate (Wei et al., 2022; Schaeffer et al., 2023) is a change of layer (Corollary 2); data-mixing transitions (Gu et al., 2025) are share-law thresholds.

Induction heads and grokking. Accounts of induction-head formation (Olsson et al., 2022; Bietti et al., 2023; Reddy, 2024; Musat et al., 2025; Aoyama et al., 2025) and grokking (Power et al., 2022; Nanda et al., 2023; Lyle et al., 2025) describe two members of one family: a degree-2 composition on the staircase clock, and a full condensation cycle on the gate clock.

Interaction order as a representational statistic. A parallel line decomposes a *trained* network's function into Harsanyi interactions, finding a U-shaped order bottleneck and a two-phase dynamics of increasing order (Deng et al., 2022; Zhang et al., 2024), explained there by mid-order gradient variance; we decompose the *target* and supply kinetics, so the two lines meet on the noise side from opposite directions.

Batch size and target hardness. That batch interacts with target order is not new: a giant step needs $n = O(d^\ell)$ samples for leap index ℓ (Dandi et al., 2024), the optimal batch tracks the information exponent (Arnaboldi et al., 2024), SGD's global phases are organized by η/B with a critical batch

(Sclocchi & Wyart, 2024), and kernel-eigenbasis learning curves already carry $1/B$ noise (Bordelon & Pehlevan, 2022). Ours is the fixed-dimension, many-step, per-term version: one dimensionless group, a birth *curve* rather than a sample-complexity scaling, and the extreme-value step over units. On the Boolean cube the static NTK is already Walsh-diagonal with degree-graded eigenvalues (Yang & Salman, 2019); our Θ is its finite-width, feature-learning counterpart, which must itself grow—whence the plateau the kernel regime does not show. Note too that noise never assists escape here: the target direction is downhill throughout, so fluctuations destroy the product's coherence rather than surmount a barrier—why smaller batches make high-degree terms slower.

What the loss does and does not determine. Du et al. (2024) show downstream performance is a function of pre-training loss at fixed corpus, tokenizer and architecture, and Liu et al. (2023) that at equal loss the remaining variation tracks flatness; both concern what a model can do at a given loss, whereas our claim is about the rate at which the next capability forms. Heskes (2025) characterizes the losses admitting a clean bias–variance decomposition; Proposition 3 is the analogous statement along a different axis, additivity over an orthogonal function basis rather than over predictions. Pinson (2026) argues that capacity allocation is a race between units rather than a lottery over initializations; our measurements place both regimes in one system and supply the criterion separating them, near-deterministic birth at low degree and memoryless waiting at high, with the boundary at $\mathrm{Co} \approx 1$. The transience of in-context learning (Singh et al., 2023) is the death side of the same ledger, a single-part linear clock rather than the joint event that produced it. In the lazy or kernel regime (Chizat et al., 2019) the feature map does not move and no plateau of this kind arises, which is why Θ must itself grow here.

Noise as a resource. Abbe & Sandon (2020) prove that small-batch SGD is poly-time universal while large-batch gradient descent cannot learn functions of low cross-predictability, so in their setting sampling noise is a computational resource rather than a nuisance. The boundary is worth stating precisely. Their result is constructive: it fixes an initialization and architecture built to decode the fluctuation, and establishes what is *possible*. Ours concerns generic initialization and standard architectures, where no such decoder exists and the fluctuation only destroys the coherence of a product. Their lower bound—that averaging over a large batch removes the signal—is the same fact as Theorem 1.

Nearby names, different objects. Spectral bias (Rahaman et al., 2019; Xu et al., 2020) concerns continuous-input frequency and weight condensation (Zhou et al., 2022) parameter-space clustering, whereas ours lives on the *data's interaction structure*; singular learning theory and developmental interpretability (Watanabe, 2009; Hoogland et al., 2024) give equilibrium pictures where we give rate laws; superposition capacity (Elhage et al., 2022; Henighan et al., 2023) is the representational face of our capacity law. Surveys (Simon et al., 2026; Sharkey et al., 2025) bound what this framework reaches (Appendix J).

B THE KINETIC AND AGGREGATION LAYERS IN FULL

Theorem 3 (Staircase catalysis)**.** *Let $y = \alpha\chi_B + \beta\chi_S$ with $B \subset S$. Suppose the low-order term condenses first. We assume this rather than derive it: the lower degree makes it the likelier ordering, but a sufficiently small α^2 relative to β^2 reverses it, so $|B| < |S|$ alone does not settle which arrives first. Its parts' amplitudes then mature, and by Theorem 2 the missing-part product for χ_S then runs over $S \setminus B$ only: the effective degree drops from $|S|$ to $|S \setminus B|$. Quantitatively this requires the matured amplitudes to be $\Theta(1)$ and slowly varying on the reduced clock's time scale ($|\dot{w}_i| \ll |\dot{w}_j|$ for $i \in B$, $j \in S \setminus B$); under this explicit time-scale separation (given by the spectral mass ratio), the two-stage clock is a two-timescale composition of proven clocks. Two further conditions are part of the statement: the reduced clock's prefactor—κ averaged at the* matured *configuration, no longer controlled by $\sigma^{(K)}(b)$—must be nonzero, which for symmetric activations at special matured states can fail; and reading a formed* term *(rather than a raw coordinate) as one factor of the product is a reuse step of the same standing as the within-spectrum clause below. Staircases (Abbe et al., 2022) are the spectrum's inclusion relations made dynamical. Scope: within a mixed single-target spectrum, reuse of the shared parts is automatic—the same unit's missing-part product is what benefits (Theorem 2); catalysis* across *separate tasks additionally requires a representation-reuse assumption, which we state where used.*

THE KINETIC LAYER IN FULL (SECTION 3)

The identities of Section 2 say a capability grows only through a product of its own parts. We ask how fast that product grows, what opposes it, and which wins.

THREE CLOCKS

A pure degree-K term routes through $q = K+1$ multiplicatively coupled amplitudes: K part weights and the readout.

Lemma 1 (Symmetric signal-channel reduction; informal)**.** *Take one unit with a non-degenerate bias and a symmetric initialization $w_i(0) = a(0) = c > 0$, under population gradient flow on squared loss. Then the K part weights and the readout stay equal to one another and collapse to a single amplitude s, whose* label-attributed *channel obeys*

$$\dot{s}\big|_{\text{signal}} \;=\; \gamma_0\, s^{\,q-1}\,\big(1 + O(s)\big), \qquad \gamma_0 = \eta\,\big|\sigma^{(K)}(b)\big| > 0, \quad q = K+1.$$

The label-independent model-fit terms are not *uniformly higher order: after centering they contribute a self-interaction contraction $-\eta\,\Theta\big(\sigma'(b)^2\big)\,s^3$ at leading order. The ten hypotheses are listed and the three parts proved in Appendix H.*

A prediction of ours that failed. We froze the single-unit comparison in Lemma 1 as a kill condition—that under $\sigma = \tanh(\beta\,\cdot)$ the self-fit contraction should beat the signal at $K{=}3$—and a preregistered β-sweep falsified it: the class stays near-deterministic down to $\beta{=}0.25$ while the clock merely slows as $\beta^{-2.5}$. In function space the deterministic flow is one PSD descent (Lemma 2), so self-fit terms never oppose the signal there, and an exact-enumeration probe shows the low-order channels holding a noise-supported plateau rather than contracting. What survives is the identity's falsifiable content—a degenerate bias makes γ_0 vanish and kills the $K{=}2$ clock outright, $6/6$ censored against births at 500–750—and the reason Lemma 1 is labelled *conditional*. Sweep and probe: Appendix B.

Status. Clocks (i) and (ii) below are exact *within the conditional single-unit reduction of Lemma 1*; their standing in a multi-unit network is empirical, and the reduction's one quantitative failure is reported with the lemma. Clock (iii) is a conditional statement whose independence premise is assumption (T), tested rather than derived.

Theorem 4 (Three clocks)**.** *Let s be the symmetric signal-channel amplitude of Lemma 1, evolving with drift $\gamma_0 s^m(1 + O(s))$, $m = q - 1 = K$, from $s(0) = c$ to threshold θ; all clocks below are signal-channel clocks.*

- *(i)* $m = 1$ *(a degree-1 channel; $q = 2$): exponential mode, $t^* = \gamma_0^{-1}\ln(\theta/c)\,(1 + o(1))$; randomness enters only through c. Separately, once all but one factor of a higher-degree term are mature, the remaining factor has an $O(1)$ drive and hence linear arrival—a repair channel, distinct from a degree-1 target.* Exact given the signal-channel ODE of Lemma 1.
- *(ii)* $m \geq 2$*, drift-dominated: power-law escape $t^* = \frac{c^{1-m} - \theta^{1-m}}{\gamma_0(m-1)}\,(1 + o(1))$—the slow saddle (cf. Saxe et al., 2014; Abbe et al., 2023).* Exact given Lemma 1*; the $o(1)$ is controlled by the $O(s)$ term.*
- *(iii)* (Conditional.) *When drift is dominated by fluctuations, assume rare-event activation of the m small amplitudes that is independent across factors* and *across time steps, with stationary probabilities p_i. Then the joint activation time is geometric with $p = \prod_i p_i$: $\mathbb{E}[T] = \prod_i p_i^{-1}$, memoryless,* CV $\to 1$*; in the homogeneous case $p_i = p_0$, $\mathbb{E}[T] = p_0^{-m}$—exponential in the number of missing parts. Proposition 1 supplies part of this premise—mini-batch noise is white* across *spectral channels, with only $O(\mathbb{E}[r^2\chi_{C\triangle C'}]/B)$ off-diagonal coupling—but independence* across time steps *remains an assumption about SGD, not a consequence.*

Which clock a regime runs on is empirical, not part of the statement: in the laboratory the high-K regime carries the signatures clause (iii) requires, memoryless waiting with CV $\to 1$ and a ladder exponential in the number of missing parts (Appendix C).

Theorem 5 (Combinatorial barrier: the joint event is exponentially locked in K)**.** *Let $s_1, \ldots, s_K$ be independent stationary OU processes (relaxation rates ρ_i, stationary widths σ_i; the subscripted*

σ_i are widths, not the activation)—the zero-target channels of Proposition 1(a2). The independence premise is the fluctuation-branch idealization: on that branch the product drift of Theorem 2, of order s^K (the readout contributes its factor $a = s$), is negligible against the noise, so the K part channels decouple to their zero-target laws; this is exactly the regime the theorem prices, and outside it the drift clock of Theorem 4(ii) takes over. Let T be the first time all K satisfy $|s_i| \geq \theta_i$ simultaneously*, with $\theta_i \gg \sigma_i$. Then, with $\tau = c_0 \max_i \rho_i^{-1}$ and absolute constants c_0, c_1 (neither tracked),*

$$\mathbb{E}[T] \;\geq\; \tau \, \exp\Big(\textstyle\sum_i \big[\frac{\theta_i^2}{2\sigma_i^2} - c_1 \frac{\theta_i}{\sigma_i} - \ln 2 \big] \Big) \;\xrightarrow{\theta_i = \theta,\ \sigma_i = \sigma}\; \tau \, \exp\Big(K \big[\frac{\theta^2}{2\sigma^2} - c_1 \frac{\theta}{\sigma} - \ln 2 \big] \Big) :$$

the number of missing parts enters the exponent linearly.

Proof. Cut time into blocks of length τ. If the joint event occurs inside a block then every coordinate's block supremum exceeds its threshold, so by independence across coordinates the per-block hit probability is at most $\prod_i P(\sup_{[0,\tau]} |s_i| \geq \theta_i)$. Borell–TIS (Adler & Taylor, 2007) bounds each factor by $2\exp(-(\theta_i - \mathbb{E}\sup)^2 / 2\sigma_i^2)$, and a stationary OU process on a window of $O(1)$ correlation lengths has $\mathbb{E}\sup = O(\sigma_i)$, which produces the linear correction $c_1 \theta_i / \sigma_i$. Blocks are compared through the Markov property, so the number of blocks until a hit dominates a geometric variable with the stated rate. (Rice's upcrossing formula is unavailable—OU paths are not differentiable—which is why the bound runs through Borell–TIS.) □

Three remarks. *A single channel cannot lock*: in logarithmic coordinates the multiplicative diffusion of Proposition 1(b) has $2\mu/v = 1$, a critical coincidence under which the escape probability is polynomially, not exponentially, small—the exponential wait is a property of the *joint* event, which is why clock (iii) is a statement about K coordinates and is not derivable from any one-dimensional diffusion. *Division of labour with clock (iii)*: the theorem bounds the *rate* with no time-independence assumption; what clause (iii) still assumes—independence across steps—buys the full geometric law (memorylessness, $\mathrm{CV} \to 1$), not the exponential scale, which the bound already delivers. *Against the data*: the ladder's high-K segment has per-rung log-increment ≈ 1.5, and inverting the theorem gives $\theta/\sigma_{\mathrm{st}} \approx 1.7$—a consistency check, not a zero-parameter hit, since the per-part threshold is not measured independently.

Proposition 1 (Spectral noise, and the birth curve)**.** (a) The residual spectrum, exactly. *In the Walsh basis one SGD step is $\delta \hat{f} = -\eta \, \Theta \, \hat{r}_{\mathrm{emp}} + O(\eta^2)$, with $\mathbb{E}[\hat{r}_{\mathrm{emp}}] = \hat{f} - \hat{y}$ and, because $\chi_A^2 \equiv 1$, $\mathrm{Var}[\hat{r}_{\mathrm{emp}}(A)] = \big(L - (\hat{f} - \hat{y})(A)^2\big)/B$ where $L \equiv \mathbb{E}[r^2]$. That much is exact, and it is all that is exact.* (a1) *The clean form L/B is a zero-target reading: exact at all times on a channel carrying no target mass, and on a target channel only* once it has condensed*, its own residual then vanishing. Before condensation the subtraction is the target's own mass, negligible only while the bath dominates it. On that reading the residual is isotropic across the spectrum and $\mathcal{T} = \eta L/B$.* (a2) *An isotropic residual is not an isotropic update: $\Sigma_{\delta \hat{f}} = \eta^2 \Theta \Sigma_r \Theta^\top + O(\eta^3)$ is isotropic only insofar as Θ acts diagonally in the Walsh basis. The* static *NTK on the Boolean cube is Walsh-diagonal (Yang & Salman, 2019); the finite-width, feature-learning Θ this paper is about is not, and we do not prove that it is. We name that **(D)** and state rather than derive it. Under (D) each zero-target channel is an Ornstein–Uhlenbeck process and the loss* is *the temperature, so a run cools itself as it learns.* (b) One dimensionless group, under (D), (T) and (P). *Replacing the recursion by a diffusion needs **(T)**, increments uncorrelated across steps—an assumption about SGD, not a consequence of (a), and the same one clause (iii) names. One step more is needed because (a) concerns the residual spectrum while s is a* parameter*: the fluctuation reaching a part weight passes through its own Jacobian $\partial f / \partial w_j = a \, \sigma'(z) \, x_j$, and in the symmetric channel the readout equals the part weight, $a = s$, so the multiplicative form is that Jacobian factor rather than a postulate, giving $\mathrm{Var}[\Delta w_j \mid \mathcal{F}_t] = \eta^2 s^2 \sigma'(b)^2 L/B \, (1 + o(1))$. What it does not give free is that $x_j \sigma'(z)$ is decorrelated from the residual over the batch, **(P)**—though pre-condensation this discharges: $x_j^2 \equiv 1$ factors the joint moment exactly and $\sigma'(z) = \sigma'(b) + O(\|w\|)$, leaving only the correction the $(1 + o(1))$ already carries (see below). Under (D), (T), (P) a part weight follows $ds = \eta \gamma s^K dt + \eta \, \sigma'(b) \, s \sqrt{L/B} \, dW$, $\gamma = |\sigma^{(K)}(b)|$. The scale function of* that *diffusion is exact within the diffusion approximation—the error sits in the approximation itself, whose empirical domain the width sweep bounds—and equals $\exp(-\frac{2}{K-1} \mathrm{Co} \, (s/c)^{K-1})$ with*

$$\mathrm{Co} \;=\; \frac{\gamma \, c^{K-1}}{\sigma'(b)^2 \, \mathcal{T}} \;=\; \frac{\gamma \, B \, c^{K-1}}{\eta \, L \, \sigma'(b)^2},$$

the responsibility at the initial amplitude over the temperature as the parameter actually receives it (the Jacobian contributes $\sigma'(b)$ to the noise, hence $\sigma'(b)^2$ here): a Péclet-type group we call the condensation number. *For that one-dimensional first passage, at fixed threshold ratio and for a single unit, the escape statistics depend on $(\gamma, B, c, K, \eta, L)$ through* Co *and the explicit K alone; threshold ratio and unit count enter separately, the latter through (c). Drift- and fluctuation-dominated are* $\mathrm{Co} \gg 1$ *and* $\mathrm{Co} \ll 1$. (b′) One temperature, shared by everything. *The per-channel variance in (a) is the same L/B for every channel whatever its degree, so the temperature a capability must survive is set by the* total *residual—early in training, by everything not yet learned. Whether that bath coupling is what actually links capabilities is a separate question, and Section 3.2 reports that it is not.* (c) Birth is an extreme value, so the boundary is a curve. *A capability is born once* any *of the d units escapes, so with per-unit escape probability $p = p(\mathrm{Co}; K)$, $\Pr[\textit{birth}] = 1 - (1-p)^d$: the network-level transition sits at $d\,p \approx 1$, lower than the single-unit condition by a factor logarithmic in d, and its width comes from the same expression. There is no single critical batch size, only this curve.*

The standing of the three assumptions. They are not on the same footing, and two of the three can be pushed further than the main text claims for them.

(T) is a freezing step, not an independence axiom. Under i.i.d. sampling, successive mini-batches are drawn independently, so conditional on the parameter trajectory the noise increments are martingale differences *by construction*. What (T) actually supplies is the standard slow-drift freezing of the diffusion limit (Ethier & Kurtz, 1986): the multiplicative amplitude $s\sqrt{L/B}$ must vary slowly on the correlation time of the increments, which holds as $\eta \to 0$ and is the regime of every experiment here. The one caveat is real but bounded: L and s are themselves trajectory-dependent, which is what confines (T) to a local-in-time statement.

(P) discharges pre-condensation. The batch-level variance of $\partial\mathcal{L}/\partial w_j$ is $\frac{1}{B}\big(\mathbb{E}[r^2(x_j\sigma'(z))^2] - \mathbb{E}[r\,x_j\sigma'(z)]^2\big)$. The subtracted mean is the drift of Theorem 2, of order s^{K-1}, so its square is negligible. For the leading term, on ±1 inputs $x_j^2 \equiv 1$, so the joint moment factorizes *exactly* in x_j; and before condensation $\sigma'(z) = \sigma'(b) + O(\|w\|)$, so $\mathbb{E}[r^2(x_j\sigma'(z))^2] = \sigma'(b)^2 L\,\big(1 + O(\|w\|)\big)$. What remains of (P) is that $O(\|w\|)$ correction, which the $(1 + o(1))$ of part (b) already carries: pre-condensation, (P) is a consequence, not a postulate.

(D) is the wall, and its domain is measured rather than proved. Proving that the finite-width, feature-learning Θ acts diagonally on the relevant Walsh subspace would amount to a quantitative theory of feature-learning dynamics, which does not exist; we do not attempt it. What we have instead is the empirical boundary of its consequences: at widths $d = 64$ and 128 the parameter-free floor prediction that (D) licenses lands at 0.83–1.26 of the measurement across a $64\times$ batch range, while narrowing the layer breaks it in the predicted direction, 1.34 at d=16 and 2.09 at d=4. Those two failures are not noise; they are the edge of (D)'s domain, located experimentally. Every use of Co in this paper sits on the wide side of that edge.

Every part is tested, preregistered, in a squared-loss variant of the laboratory where the proposition is exact rather than local, with all spectral quantities obtained by enumerating the 2^{16} inputs. The zero-target floor scales as $B^{-0.545}\eta^{+0.531}$ against a predicted $B^{-1/2}\eta^{+1/2}$, and the parameter-free prediction of its magnitude lands at 0.97 of the measurement across a $64\times$ batch and $4\times$ learning-rate range; every factor of Co moves censoring in the predicted direction; and inverting the extreme-value form at four widths returns one per-unit escape probability over a $16\times$ range. Cells, sweeps, and the one factor that resists quantitative prediction: Appendix B and Figure 5.

ADAM, IN FULL

Real pretraining uses Adam, so the temperature must be re-examined rather than assumed. Adam updates w_j by $\eta\,\bar{g}/(\sqrt{\hat{v}} + \epsilon)$, where $\hat{v}$ is an exponential moving average of $\bar{g}^2$ and $\bar{g}$ is the mini-batch gradient of that coordinate. On the pre-condensation plateau the trajectory is slow, so the average converges to the current second moment, and that moment can be decomposed. The gradient carries the Jacobian $\partial f/\partial w_j = a\sigma'(z)x_j$ with $a = s$ in the symmetric channel: its mean (the signal) is $O(s^K)$ by Theorem 2, its batch fluctuation has standard deviation $s\sigma'(b)\sqrt{L/B}\,(1 + o(1))$ by Proposition 1(b). The second moment is therefore fluctuation-dominated, $\hat{v} = s^2\sigma'(b)^2 L/B\,(1 +$

$o(1)$): the signal being buried in the noise and the signal being invisible in $\hat{v}$ are the same statement, which is what makes assumption **(A)**—the per-coordinate normalizer is the whole of Adam's effect, moment schedules and ϵ neglected—self-consistent rather than circular. Dividing the one-step drift and noise by $R = \sqrt{\hat{v}}$ gives

$$ds = \eta\, g\, s^{K-1} \sqrt{B/L}\, dt + \eta\, dW, \qquad g = \frac{\sigma^{(K)}(b)}{\sigma'(b)},$$

additive noise and a drift one power flatter than SGD's. The scale density of that diffusion is $\exp\!\big(-\frac{2}{K}\,\mathrm{Co}_{\mathrm{Adam}}\,(s/c)^K\big)$ with

$$\mathrm{Co}_{\mathrm{Adam}} \;=\; \frac{g\,c^K}{\eta}\sqrt{\frac{B}{L}},$$

carrying the $1/\eta$ and the full power c^K that the multiplicative SGD group does not, and scaling as $\sqrt{B/L}$.

Two earlier versions of this derivation were wrong. The first assumed $\hat{v}$ was dominated by the target channel's own fluctuation without the Jacobian factor, giving a degree penalty one power steeper. The second—the version this text replaces—dropped the Jacobian's s from $\hat{v}$ altogether, treated R as a constant set by the bath, and concluded that the drift exponent was unchanged with multiplicative noise. Both fail the same audit: $\hat{v}$ is the second moment of *this coordinate's* gradient, and every term of that gradient carries the factor $a = s$. With only the optimizer changed on one grid, birth is batch-independent under SGD in the drift branch (log-log slopes 0.023, 0.021 at $K = 3, 4$) and scales as $B^{-1/2}$ under Adam (-0.450, -0.413, -0.448 at $K = 3, 4, 5$). The $B^{-1/2}$ is common to the corrected and the replaced form—the normalizer contributes $\sqrt{B/L}$ to the drift either way—so those slopes are consistent with both and adjudicate neither. Two other measurements do, and both were in hand before the correction: additive noise does not vanish with s, so small amplitudes are not trapped, and Adam censors 3 of 12 cells (all at K=6, smallest batch) where SGD censors 12; and the effective saddle order is $K-1$, so low degrees nearly collapse onto one clock, K=2 and K=3 arriving at the same step under Adam where SGD separates them by an order of magnitude. The corrected form also locates the deletion results: temperature-driven deletion is a multiplicative-noise phenomenon, and Adam's normalizer divides that factor out, which is why deletion is established under SGD and expected absent, not merely untested, under Adam in this regime. Aoyama et al. (2025) measure $B^{-0.37}$ for induction-head emergence and Section 5 measures $B^{-0.689}$ for an injected skill; the two bracket the $-1/2$ that $\mathrm{Co}_{\mathrm{Adam}}$ predicts without either landing on it, and we report that cell as undecided. The denser Adam ladder ($K \le 8$), its three zero-parameter clock ratios, and the boundary where the additive form itself turns lottery are in the phase-diagram subsection (Appendix C).

Noise, in full

Noise plays two roles here and they must be kept apart. In the drift-dominated branch it can only slow formation, for a structural reason: along the direction that grows a capability there is *no barrier*—the drift $\eta\gamma s^K$ points at the target for every $s > 0$—so the deterministic flow already goes the right way and only its speed is at issue. What noise does instead is multiply s by a random factor of size $\eta s\sqrt{L/B}$, randomizing the K parts independently while the pull requires them aligned at once. The pull grows as the Kth power of the amplitude and the shaking only as the first, so noise dominates exactly where the process spends nearly all its life, the long stretch at small s. The prediction is opposite in sign to the escape-from-saddle picture: smaller batches make high-degree terms slower, never faster, and past a computable point stop them. Barak et al. (2022) showed parity learning is not a Langevin-like search; the condensation number says where the search is killed. In the fluctuation-dominated branch the ordering reverses in one sense—there the pull is too weak to assemble the parts at all and only a fluctuation ever completes the joint event—but what it delivers is a memoryless rare event, not an acceleration; lowering the temperature still shortens the expected wait. Neither branch is a case of noise helping in the sense the SGD literature means.

Three preregistered measurements agree, all under plain SGD. Shrinking B at K=4 over $64, 32, 16, 8, 4$ moves censoring from $0/4$ to $4/4$, with no run forming within 300,000 steps at $B \le 8$. Raising the temperature directly—label noise independent of the target, which by Theorem 1

leaves the drift untouched—censors $4/4$ a term that otherwise forms in 3875 steps, every run given $77\times$ that budget; quadrupling the injected variance leaves $3/4$ censored with a single escape at 165,250 steps, the rare-event signature the fluctuation branch predicts. Lowering the target amplitude, which raises Co without touching the noise, rescues a censored configuration. We have found no regime in this laboratory in which added noise accelerates formation.

Capabilities are coupled when their variables are disjoint, but not through the bath. Part (b′) says an unlearned companion should heat its neighbours and a learned one cool them; the decisive test holds the target's supply fixed and varies only the companion's residual, and we ran it. A companion on a *disjoint* variable set accelerates a degree-3 target $3.9\times$ (9375 against 36250 steps, 12 seeds each)—but the sign is wrong for a bath, because carrying the companion *raises* the residual for the whole wait ($L = 1.361$ against 1.000 at matched steps), so the faster arm is the hotter one. A dose–response settles it: sweeping the companion's amplitude over $0.2, 0.4, 0.6, 0.9$ moves the target's birth to $44125, 18000, 9375, 7875$ while the pre-birth residual rises monotonically over $1.040, 1.160, 1.360, 1.447$. What explains it is already in Theorem 2: the pull carries the readout weight as a factor, and that factor is shared by every term, so learning anything raises it—at step 10,000 the readout norm is 1.51 in the companion arm against 0.06 alone. Capabilities do couple across disjoint variable sets, but through a shared multiplicative substrate and not through the temperature. We record that the version of this claim in earlier drafts had the mechanism and the sign wrong.

Theorem 6 (Spectral loss identity)**.** *For squared loss, exactly:* $\mathbb{E}[(f_{\text{model}} - y)^2] = \sum_A \big(\hat{f}_{\text{model}}(A) - \hat{y}(A)\big)^2$. *Hence the training curve is* $L(t) = \sum_A \hat{y}(A)^2\, S_A(t)$ *+ (model excess), where* $S_A(t)$ *is the uncondensed fraction of term* A*. The decomposition is the exact aggregation layer; the clock theory (Theorem 4) is our dynamical law for each* $S_A(t)$*, and capacity bounds which terms can hold. "Spectrum* $\times$ *clocks* $\times$ *capacity" is an identity at the aggregation layer and a model at the per-term layer.*

Theorem 7 (Joint-evidence share law; no curriculum shortcut)**.** *Classify samples by the informatively-observed subset* $B \subseteq S$*. By Corollary 1(c), only* $B = S$ *samples carry signal; their gradient weight is their share* f*, so the target-attributed drift is linear in* f *and, at fixed representation and threshold (the other dynamics held fixed), the signal-channel manufacturing clock scales as* $t^* \propto 1/f$*. Non-joint samples still occupy gradient norm and capacity: they dilute, and can push replaced coordinates toward label-independent solutions. Any "parts-first" curriculum contributes zero signal to a pure degree-*K *term.*

Two long-form records referred to from Section 3

A prediction of ours that failed, and what it cost. *What we predicted.* For a single unit the comparison in Lemma 1 is sharp: the signal beats the self-fit contraction at $K{=}2$ and, under $\sigma = \tanh(\beta\,\cdot)$, should lose at $K{=}3$ once $2\beta^3 < \beta^2$. We froze that as a kill condition for the multi-unit laboratory. *What happened.* A preregistered β-sweep falsified it: at $K{=}3$ the class stays near-deterministic down to $\beta{=}0.25$ (CV$\approx$0.2, no censoring) while the clock slows $\propto \beta^{-2.5}$—the $\gamma_0 \propto \sigma^{(K)}$ prefactor. *Why.* Lemma 2 gives part of the reason: in function space the *deterministic* part of the flow is one PSD descent, so self-fit terms never oppose the signal there; what remains is the noise of Proposition 1, which breaks that monotonicity under SGD. An exact-enumeration Fourier probe shows what the low-order channels then do: the low-order channels do *not* relax to the quasi-static level a population-flow argument predicts but hold a noise-supported plateau, so the self-fit contribution fluctuates at 0.2–2.2$\times$ the signal ($K{=}2$–4) instead of contracting systematically (Appendix H). *What we keep.* The class switch stays with the drift–fluctuation boundary of (iii), not with the self-fit term, and the identity's own falsifiable content survives decisively: a degenerate bias, $b{=}0$, makes γ_0 vanish and kills the $K{=}2$ clock outright (6/6 censored in both laboratories against births at 500–750). We report the failure at this length because it is the only place where a single-unit reduction misled us about the network, and it is the reason Lemma 1 is labelled conditional rather than exact.

Proposition 2 (Network dilution; outline standing)**.** *Let* $f = \sum_{h=1}^{d} a_h \sigma(w_h{\cdot}x + b_h)$ *under population squared loss with target* χ_S*,* $|S| = K$*, initialization amplitude* c*, and write* Θ_{low} *for the Gram block of the degree-*$< K$ *Fourier coordinates. Assume* ***(H1)*** $\lambda_{\min}(\Theta_{\text{low}}) \geq \lambda_0 > 0$ *throughout the pre-formation window—to be read at the window's amplitude scale,* $\lambda_0 = \lambda_0(c)$*, since the block's entries themselves carry* $\sigma'(b)^2$ *and powers of the amplitudes;* ***(H2)*** $\gamma c^{K-1}/\lambda_0 \leq \varepsilon \ll 1$*;* ***(H3)*** $\|\Theta_{\text{low},S}\| \leq$

δ. Then on the slow manifold the symmetric target channel obeys $\dot{s} = \gamma_0 s^K\,(1+O(s)+O(\varepsilon)+O(\delta))$: the low-order residual is a fast variable relaxing at rate $\eta\lambda_0$ to a quasi-steady level $O(s^K/\lambda_0)$, so the self-fit force on the slow channel falls from the single unit's $\Theta(\sigma'(b)^2 s^3)$ to $O(\sigma'(b)\, s^{K+1}/\lambda_0)$—one order below *the signal. Under SGD the fast variables relax instead to the OU plateau of Proposition 1(a2), whose mean is zero: with **(H4)**, decorrelation of the fast fluctuation from the slow channel on the averaging time scale (Khasminskii averaging under the mixing that the same spectral gap supplies), the net slow drift is unchanged, and the plateau's fluctuation joins the existing multiplicative noise—an $O(s)$-relative addition to the $\sigma'(b)^2 L/B$ width, inside the $(1+o(1))$ that Proposition 1(b) already carries.*

Standing and content. The fast–slow step is Tikhonov reduction (population) and Khasminskii averaging (SGD); every order has been checked, and we state the result at outline standing—the ε-δ constant tracking is not written out, and we label it so rather than calling it a theorem. Three points carry the content. *The unit is the degeneracy*: at $d = 1$ the low-order output direction is pinned to (w, a), the fast system cannot relax it without moving s, $\lambda_0 = 0$, and the $\Theta(s^3)$ contraction stays in the slow equation—the unit's failure and the network's escape are one quantity, λ_0. *The measurement is the mechanism*: the exact-enumeration probe's "noise-supported plateau, self-fit fluctuating 0.2–$2.2\times$ with no systematic sign" is the SGD clause verbatim. *(D) is needed only as (H3)*: the argument uses the smallness of one cross-block, not full Walsh-diagonality, and the narrow-width failures (1.34 at d=16, 2.09 at d=4) are (H1) degrading—which upgrades the width sweep from an empirical boundary of (D) to a quantitative test of (H1).

An earlier version of this derivation was wrong, and the experiment that corrected it. We previously assumed $\hat{v}$ was dominated by the target channel's own fluctuation, which gives additive noise and a degree penalty one power steeper. That premise does not hold for the reason just given. The correction is not cosmetic: it reverses the prediction about what the batch does. We tested it with the same driver and the same grid, changing only the optimizer. Birth time under SGD is *independent* of batch size in the drift-dominated branch (log-log slopes 0.023 and 0.021 at $K = 3, 4$), while under Adam it scales as $B^{-1/2}$ (slopes -0.450, -0.413, -0.448 at $K = 3, 4, 5$), matching the $\sqrt{B/L}$ above. Over the same grid SGD censored 12 runs and Adam censored 3, all three at K=6 and the smallest batch.

Two consequences follow, and we state them because they narrow the paper. *(i)* Under SGD the batch is a pure thermostat: within the drift-dominated branch it does not move the clock at all, and its only effect is whether the term forms. *(ii)* Under Adam the batch is mainly a clock, and censoring is rare in the reachable range. **The deletion results of Section 3.2 are therefore established under SGD, and their Adam-side scope is open.** The learning rate, by contrast, is a pure clock under Adam in the drift-dominated branch, which we confirm over a $30\times$ range at $K = 2, 3, 4$: birth at $3250/5750/20375$ steps at $\eta = 10^{-4}$ falls to $250/250/875$ at 3×10^{-3}, a $23\times$ speed-up at K=4 against $30\times$ predicted, with no censored runs at any setting.

The same $-1/2$ exponent has been measured on real pretraining without a theory attached. Aoyama et al. (2025) fit the number of updates to induction-head emergence across three orders of magnitude in model size and obtain $B^{-0.37}$, with the exponent estimated by regression. Section 5 measures $B^{-0.689}$ for an injected degree-2 skill on Pythia-70m. The two bracket the $-1/2$ that $\mathrm{Co}_{\mathrm{Adam}}$ predicts without either landing on it, and we report the real-machine cell as undecided against its preregistered band.

A capability is sharp, and a loss curve is smooth. Sections 2 and 3 price one term at a time; this section is about what happens when thousands of them are priced at once.

B.1 The loss curve is a spectral identity

Two boundaries of this identity are proved in Appendix H: among Bregman losses, spectral additivity for all targets holds *only* for squared loss, and for cross-entropy the identity survives as a second-order local statement whose weights are the softmax Fisher metric in the character basis—diagonal, hence additive, when the optimum is invariant under the cube's sign-flip group, and otherwise only up to the off-diagonal terms.

Corollary 2 (Emergence vs. mirage: both layers are real)**.** *Per-term progress $S_A(t)$ is sharp (a clock; Theorem 4); the loss is a sum of many terms and is therefore smooth. Capability-resolved metrics*

must look discontinuous while token-averaged metrics look continuous (Wei et al., 2022; Schaeffer et al., 2023): the two positions describe different layers of the same identity, and Sections D–E exhibit both layers with measurements in one system. This does not settle which benchmark metrics are discontinuous (a property of each metric's aggregation); it removes the appearance of contradiction.

B.2 Sample pricing and the share–threshold decomposition

When S capabilities share a fixed budget, the log-slope of birth time with respect to S splits exactly into a share-dilution term (equal to 1 when the share is $1/S$) and an occupancy-threshold term (Appendix H); measured, the two are 1.06 and 0.12, so the excess over 1 is a capacity-proximity meter readable without triggering collapse. Real pretraining fixes the token budget, making the mean-normalized mode the physical one: "more capabilities each form slower" is mostly share dilution.

B.3 Staircases are spectral inclusion; capacity

Under a random-superposition model (i.i.d. directions, Gaussian interference, $\arg\max$ over V candidates) the readout level admits the parameter-free asymptotic threshold $C \sim d/(2\ln V)$ (Appendix H; 15.4 at d=128, V=64, against a measured band 16–32—just below it, so we claim the order and the d-dependence, not the constant).

Two-level min. The full trained system behaves as a two-level series with each level linear: $C = \min\big(d/m_\sigma(K),\ \alpha_r r\big)$ (Appendix F). The hidden level is derived as an achievable rate—K ReLU units suffice, so $\alpha_h^{\mathrm{ReLU}} \geq 1/K$ with no fitted constant, and minimality (the ceiling reading) is proved only at K=2, where $1/K = 0.500$ (Proposition 5); how closely a run reaches that level is measured and setting-dependent (0.52 under one loss convention, 0.33 under another, Appendix F)—while $\alpha_r = 1.1 \pm 0.1$ is measured throughout.

All proofs are given in Appendix H; Table 1 maps each theorem to its experimental photographs.

Two of the quantities above are measured rather than derived—α_r at the readout level, and the per-unit escape probability p of Proposition 1(c). We emphasise that the contribution we claim is the functional form and the variable it depends on, not those numbers: a reader who disagrees with $\alpha_r = 1.1$ should still find the $\min(\cdot,\cdot)$ structure and the d/K level at issue, and every prediction we freeze below is a statement about a form or a direction, never about a fitted constant.

All three parts are tested, preregistered, in a *squared-loss variant* of the laboratory where the proposition is exact rather than local (scalar output; all spectral quantities obtained by enumerating the 2^{16} inputs, so the instrument adds no sampling noise). *(a)* We measure the floor on an *out-of-rule* coordinate. That is the only channel with no target coupling, and so the only one the proposition actually describes. Its width scales as $B^{-0.545}\eta^{+0.531}$, against a predicted $B^{-1/2}\eta^{+1/2}$. The proposition predicts a standard deviation while we measure a median absolute value, so the comparison carries the Gaussian factor 0.6745; with that conversion and nothing else, the parameter-free formula lands at 0.97 of the measurement, ranging over 0.83–1.26 across a $64\times$ batch range and a $4\times$ learning-rate range (Figure 5, six cells, no fitted quantity). Narrowing the layer breaks the agreement in the predicted direction: the Gaussian-interference premise fails at d=16 and d=4, where the ratios rise to 1.34 and 2.09. *(b)* Every factor of Co moves the outcome as predicted: raising K; shrinking B (at K=4, censoring over $B = 4, 8, 16, 32, 64$ runs $4/4, 4/4, 2/4, 0/4, 0/4$); raising η, which shifts that transition upward; flattening the activation ($4\times$ less γ costs $44\times$ in time); and, most directly, raising L with label noise *independent of the target*, which leaves the drift untouched yet stops a degree-4 term from forming at all ($4/4$ censored where it otherwise forms in 3875 steps). Lowering the target amplitude rescues a censored configuration, as preregistered. *(c)* Sweeping width at fixed B gives censoring $4/4, 3/4, 2/4, 1/4, 1/4$ for $d = 8, 16, 32, 64, 128$, and inverting $(1-p)^d$ at each width separately returns the *same* per-unit escape probability $(0.018, 0.021, 0.021, 0.011)$—one parameter over a $16\times$ range. This is why the single-unit condition overestimates the network threshold by an order of magnitude, and why the boundary is a curve, not a number. One factor resists quantitative prediction: bias enters γ but also sets $\sigma'(b)$ and saturation, so while $\sigma^{(K)}(b) = 0$ abolishes formation ($6/6$ censored at b=0 against 0–$1/6$ at every nonzero bias), the ordering across nonzero biases does not follow $1/\gamma$; we report the identity of γ, not a bias-to-threshold formula.

C THE LABORATORY IN FULL

THE FLOOR SUBSECTION IN FULL

AN IRREDUCIBLE-NOISE FLOOR (FULL VERSION)

The three factors of the temperature $\mathcal{T} = \eta L/B$ are logged by every training run: the learning rate, the loss, and the batch. (For cross-entropy the same expression is the second-order Fisher form, Proposition 3.) That makes one consequence immediate, and it is the sharpest practical statement the theory produces.

L has a floor. If the target is not a deterministic function of the input—as natural text is not—then even a perfect model leaves the irreducible conditional variance σ^2_{irr}, so $L \geq \sigma^2_{\text{irr}}$ and $\mathcal{T} \geq \eta\sigma^2_{\text{irr}}/B$ *for the entire run*. The bath has a temperature it never falls below. The condensation number of a degree-K term is $\text{Co}_K = \gamma_K c^{K-1}/(\sigma'(b)^2\mathcal{T})$, so a term is confined to the fluctuation branch for the rest of the run once

$$\gamma_K \, c^{K-1} \;\lesssim\; \frac{\sigma'(b)^2 \, \eta \, \sigma^2_{\text{irr}}}{B},$$

and this—not a closed form—is the general statement. We write γ_K and not γ deliberately: the responsibility prefactor is $|\sigma^{(K)}(b)|$, which depends on K through the order of the derivative, and for the activations used here it does so strongly; a flatter activation costs $44\times$ in time by that route alone (Section 3). Only under the additional assumption that γ_K is K-independent, or bounded below by a constant, does the condition collapse to a threshold in the degree alone,

$$K^* \;=\; 1 + \frac{\ln\big(\gamma B/(\sigma'(b)^2\eta\sigma^2_{\text{irr}})\big)}{\ln(1/c)},$$

and only if $\gamma_K c^{K-1} \to 0$ is the largest assemblable degree finite at all. We use the closed form as an illustration of the mechanism and the inequality above as the claim. Beyond K^*—to be read as the largest degree assemblable *from raw parts*—the term does not merely become slow. It leaves the drift-dominated branch entirely and enters the fluctuation-dominated one, where formation is a rare memoryless event rather than a scheduled one—and it stays there, because the quantity that would have to fall has already stopped falling. Waiting is therefore the wrong instrument: the laboratory shows the term absent in 7 of 8 runs at $77\times$ the unperturbed formation budget, the single exception arriving at $43\times$ the unperturbed time, which is the memoryless signature rather than a delayed schedule. We do not claim impossibility; we claim that budget is the wrong axis. The two axes that work are visible in the same formula: lower the temperature (B up, η down, or cleaner targets), or raise the effective c for that term—which is exactly what a formed lower-order neighbour does, and is why staircase catalysis (Theorem 3) is not a curiosity but the only route past the floor.

The floor bounds the step, not the summit. A term that has already condensed is itself a part. If a degree-m term on $S' \subset S$ has formed, then $\chi_S = \chi_{S'}\chi_{S\setminus S'}$ needs only $K - m + 1$ things to align at once, so the effective degree of its neighbour drops from K to $K - m + 1$ (Theorem 3; reading the formed term as one factor is the reuse step that theorem now states). The entropy floor therefore does not cap how high a model can reach. It caps how far a model can reach *in one step*. Complexity grows sequentially, one rung at a time rather than by skipping levels.

THE FORMATION PHASE DIAGRAM, IN FULL

Everything measured in this paper places one capability at one point of a plane: structure coordinate K_{eff} (how many parts the data still owes it), temperature coordinate $\mathcal{T} = \eta L/B$. The preregistered grid behind Section 5.2: $K \in \{2, \ldots, 6\} \times B \in \{16, 64, 256\}$ at $\eta = 0.05$, plus $\eta \in \{0.0125, 0.2\}$ at B=64, 8 seeds per cell, and an Adam column $K \in \{2, \ldots, 8\}$ at B=64, $\eta = 3 \times 10^{-4}$; the laboratory of Appendix D with the initialization amplitude read from the run ($c = 0.077$, median over units).

The boundary, measured on both axes. The first censored degree is K=4 at B=16, K=6 at B=64, and none through K=6 at B=256; the learning rate marches it the same way (K=4 at η=0.2, K=6 at η=0.0125). The four censored cells give $\text{Co} \in [0.11, 0.42]$ (with the $\sigma'(b)^2$ factor, γ_K averaged over the initialization's bias range): raw cell parameters spreading over two orders of magnitude

collapse onto one number within $4\times$, against the $\mathrm{Co} \approx 1$ prediction—order-correct, low by 2–$10\times$, in the direction the d-unit extreme-value step predicts.

The clock, with no parameters. In the clocked region the ratio of successive births, $t^*_{K+1}/t^*_K = \frac{K-1}{K}\, c^{-1}\gamma_K/\gamma_{K+1}$, carries no fitted quantity—c and the γ_K are measured independently. Predictions 0.85 for $K3/K2$ and 8.69 for $K4/K3$; measured $1.06/1.07/1.12/1.11$ and $7.00/7.36/7.39/9.43$ across the four temperature conditions: eight of eight inside the frozen $\pm 40\%$ band, median absolute deviation 25%, largest 32%. The sawtooth is the point—the flat steps are γ-ratio cancellations, not anomalies. Under Adam the same expression with the corrected drift order predicts $1.33/4.76/1.59$ for $K5/K4$ through $K7/K6$, against measured $1.13/3.58/1.09$: eleven zero-parameter hits in all, and the clock runs out exactly where the boundary arrives.

Adam's boundary. $K{=}7$: $8/8$ born, median 24200, CV 0.224. $K{=}8$: $2/8$ born (36650 and 66050), six unborn at 4×10^5 steps; four of the six rerun with a 10^6 budget remain unborn. Additive noise displaces the lottery region by about two rungs and no further.

No progress variable in the lottery region. In a 2^{20}-input variant with exact enumeration, twelve locked waits ($K{=}4$ targets, 4×10^5 steps each) were catalogued over *every* subset channel—$\approx 10^6$ channels per probe—with zero channels showing a rising precursor (frozen criterion: trend $z > 3$ with last-quarter mean above twice the first-quarter mean). The dead window of Section 5 is the real-machine face of this exhaustive zero.

Jump position is not a boundary fingerprint. The first large step of $\ln t^*$ sits at $K{=}3 \to 4$ at all three batches—it does not move with temperature, while censoring does. Time reads the structure axis; determinism reads the boundary.

The real-machine columns. The no-precursor skill stays at chance in $13/13$ runs over $B \in \{2, \ldots, 64\}$ under AdamW (Finding 2 and its extension)—the structure axis unmoved by any temperature operation available to us—while the five-run calibrator clock of Appendix C spreads 12%: the two regions' signatures, side by side on one host. Per cell (batch: runs, steps): 2: 2, 10^5; 4: 2, 4×10^4; 8: 2, 3×10^4; 16: 3, 2×10^4; 32: 2, 1.2×10^4; 64: 2, 1.2×10^4. Under the $B^{-1/2}$ clock the shortest cells still carry 8–$11\times$ their expected ignition budget; the longest carry $45\times$.

Cells moved from Section 5

The ratio between arrivals is not a constant, and we no longer report one. An earlier version of this paper reported $t^*_{\mathrm{ind}} = 1.241\, t^*_{\mathrm{prev}}$ as a law with a frozen constant. That number cannot be measured on public checkpoints. Pythia stores nothing between step 512 and step 1000, and at every size we examined both crossings fall inside that single gap, so each t^* is an output of the interpolation rather than an observation, and the ratio between them is fixed by the two endpoints. In the one run whose grid resolves the transitions the ratio is 18.0 under an absolute threshold and 18.8 under a plateau-relative one, not 1.24; the third rung sits a further 3.7–$4.8\times$ out. We therefore report the ordering and the dead window, which are threshold-insensitive and resolved, and treat any particular multiple as local to a training recipe. The theory predicts the order and the zero; it never predicted a universal constant.

Share pricing on a live model (Figure 2b). We resume Pythia-70m at checkpoints 512–16000, injecting a marked degree-2 skill at full or half evidence share during continued training. Halving supply delays birth $\times 1.4$–1.8 at every checkpoint—the share law's price on a real machine—and the degree-3 variant is born after the degree-2 at every checkpoint (the order ladder). Later checkpoints ignite faster: staircase catalysis by mature subcircuits, not raw optimization progress.

Multitask dilution, preregistered (Figure 2c). Resuming Pythia-70m with $S \in \{1, 2, 4, 8\}$ *distinct* injected skills on one fixed token budget under a mean-normalized loss (each sequence carries one skill's episode, so each share is exactly $1/S$; the sum-normalized convention isolates the other term of Theorem 8 and gives a slope near zero instead), per-skill birth scales as $t^* \propto S^{0.86}$ at checkpoint 1000 and $S^{0.73}$ at 16000, and replicates across scale—160m gives $S^{1.08}$—all inside the preregistered band 1 ± 0.3 of Theorem 8: on a real model, "more skills each form slower" is share dilution, priced by the theorem.

Batch size on a clean grid, and an exponent we did not predict well. The batch sweep reported in earlier work on this system co-injected the degree-2 and degree-3 skills, which confounds supply with

temperature. Injecting only the degree-2 skill and sweeping $B \in \{2, 4, 8, 16, 32\}$ at three seeds gives median birth at $2600/1400/800/500/400$ steps, a log-log slope of -0.689 ($R^2 = 0.975$, 14 births in 15 runs). We had frozen -0.50 ± 0.15 as the pass band for agreement with $\mathrm{Co}_{\mathrm{Adam}}$ and the band is missed, by 0.04. We record the cell as undecided rather than passed. What the measurement does establish is the elimination of both alternatives: the exponent is not 0, which is what a pure thermostat gives and what SGD gives in the same laboratory, and it is not 1, which is what a per-step sample count would give. Together with the -0.37 of Aoyama et al. (2025) on pretraining induction heads, two independent real-machine measurements bracket the predicted $-1/2$ without either landing on it.

The birth crossing, live (preregistered). On resumed Pythia-70m the task-attributed gradient (the signal channel of Lemma 1) sits at the instrument floor before an injected degree-2 skill is born and jumps $\times 5.7$–6.6 at birth—three probe powers, six seeds, lag $\leq 33\%$, batch-invariant criteria passed prospectively 9/9. Act 4's crossing into the gradient world, photographed live; a contaminated degree-3 channel moves only $\times 1.5$ (control).

D BIOGRAPHY OF ONE CAPABILITY

We choose the laboratory for one reason: it is the only setting in which every quantity the theory names can be computed rather than estimated. The base measure is exactly a product, so the degree of the target is known and not inferred; enumerating all 2^{16} inputs gives the spectrum, Θ and L exactly, so the instrument itself contributes no sampling noise. We instrument the smallest case end to end: $y = x_1 x_2$ (K=2) in a minimal laboratory (16 Boolean inputs, 64 $\tanh$ units, readout over 2 true classes plus 64 distractors, plain SGD at batch 64, lr 0.05, all-joint samples; median birth step 750). Figure 4 shows the acts.

Act 0, blindness (Thm 1). y is pure degree-2, so its projection on any single variable is exactly zero: an arm shown only single-part samples makes zero progress for 8×10^4 steps.

Act 1, the responsibility ledger (Thm 2). With $V = a\, c_2\, w_1 w_2$ each wire's responsibility is the product of the other two—$O(c^2)$ at initialization: small but, for K=2, not zero. The task-attributed SNR rises $-0.25 \to 0.03 \to 1.18 \to 6.7$ over steps 100–500, tracking that product.

Act 2, which clock is on duty (Thm 4). Three wires drive each other: $\dot{s} = \gamma s^2$, take-off at $T \approx 1/(\gamma c)$. Four frozen corollaries, all confirmed (Fig. 4b,c): $T \propto \eta^{-0.86}$; near-determinism in the drift regime (CV ≈ 0.13); batch-freeness *within* that regime (8→256 flat; below B^* the term never forms, Prop. 1); width acts only as lottery tickets ($\times 1.67$ over d=16 → 128).

Act 3, sample pricing (Thm 7). Shares $f = 1,\ 0.5,\ 0.25$ give births $875/1875/4250$ ($\times 2.1$–2.3 per halving); doubling single-part samples *slows* birth ($\times 1.53$); every parts-first warm-up is a pure delay.

Act 4, birth, and Act 5, projection. The grown feature must still beat 64 distractors at the readout; the first sustained win is the birth, after which the responsible weights jump into the gradient world (post-birth SNR 13.9–19.6) and are pinned by a restoring force—maintained, not frozen. One term then gives a flat-then-drop loss; many terms sum to the smooth curve (Thm 6).

The same laboratory, across K. Birth medians for $K = 2, 3, 4, 5$: $750/875/3875/17375$ (Figure 4c). The pattern is a change of *kinetic class*, not a monotone hardness slope ($2 \to 3$ costs $1.17\times$; $3 \to 4$ and $4 \to 5$ cost $4.4\times$ each): $K \leq 3$ ride the near-deterministic slow saddle; at $K \geq 4$ the relative drift rate ($\gamma_0 c^{K-1}$) loses to fluctuation and the memoryless gate takes over (pre-birth SNR 0.16, exponential waiting). Adam flattens the saddle (K=2, 3 both born at 250—the $K-1$ drift of the corrected Adam kinetics, a prediction rather than an anomaly) but does not touch the gate (K=4 still $3.5\times$): optimizers change the push, not the orthogonality.

E THREE PREREGISTERED KILL SHOTS

The theory's distinctive claim is that the *spectral degree*—computable before training—controls the clock. Three predictions were frozen with kill conditions before running (ledger in Appendix K; Figure 6).

P-I1: the clock reads the degree, not the surface. Targets whose surface variable count and true degree disagree—a degree-3 rule padded with 2 or 4 always-present but irrelevant variables,

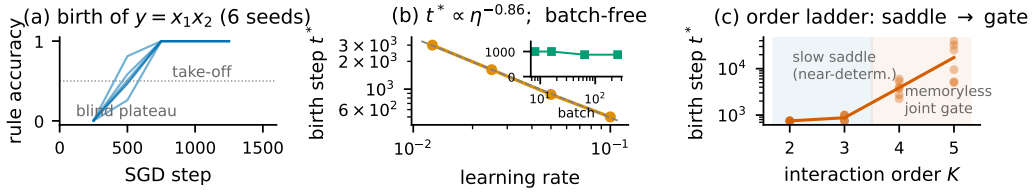

Figure 4: Biography of $y = x_1x_2$ and its K-generalization. (a) Per-seed rule accuracy: blind plateau, then a take-off resolved within ∼250 steps (probe resolution). (b) The saddle clock prices learning rate as $t^* \propto \eta^{-0.86}$ and is batch-free within the drift regime (inset). (c) The order ladder 750/875/3875/17375 for K=2..5 with the drift-to-gate switch at $K \approx 4$. Proposition 1 places the switch through Co; its location depends on the laboratory's residual scale, gain and width, and in the squared-loss variant it sits one step later.

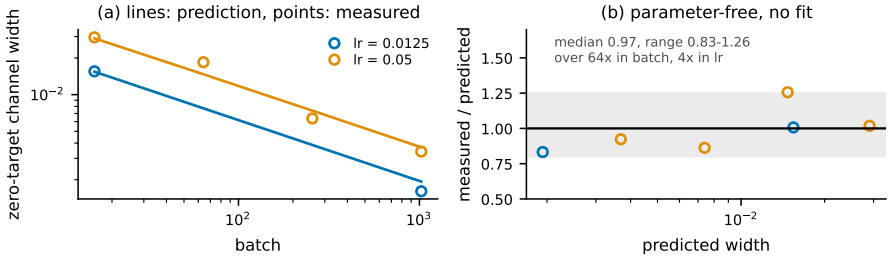

Figure 5: The fluctuation floor is predicted with no fitted quantity. *(a)* Lines are Proposition 1(a) evaluated on the measured Θ_{AA}, $(\Theta^2)_{AA}$ and L; circles are the measured width of an out-of-rule channel. *(b)* The ratio is flat at 0.97 across a 15× range of predicted widths. The comparison carries the Gaussian factor $\text{median}|X| = 0.6745\,\sigma$, since the proposition predicts a standard deviation and we measure a median absolute value; nothing else is adjusted.

versus true degree-5 and degree-7 parities—would slow like high-K arms if the clock read surfaces. Outcome: the padded arms are born at median step 1000, indistinguishable from the true degree-3 clock; true degree-5 at median 1.1×10^4 (16 seeds, wide spread, as the memoryless gate requires); true degree-7 unborn in all 8 seeds at 3×10^5—a $\geq 300\times$ separation attributable to degree alone.

P-I2: fix the marginals, permute the spectrum. Three degree-3 rules share one stream, distinguished only by spectral mass (0.6/0.3/0.1; all lower-order marginals identical), and birth times should follow that mass. Outcome: strict ordering in 8/8 seeds, medians 1560/3900/20600, share exponent 1.22 on no-stronger-neighbor points. The 0.1 rule also pays a *competition step* of ×1.5, the same whether its stronger neighbours hold 0.6, 0.45+0.45 or 0.3+0.3 (20.6k/20.3k/18.5k)—pricing the existence, not the size, of a stronger neighbour; an equal-share control shows no step.

P-I3: mixed spectra self-generate the staircase. For $\sigma(\alpha\chi_B + \beta\chi_S)$ with $B \subset S$, Theorem 3 predicts the low-order term first, then the high-order one on a reduced-degree clock, with catalysis growing in α. Outcome: solo high-order birth at 7375; with $\alpha = 0.4/0.8/1.2$ it falls to 3750/2500/2375 (catalysis 3×, monotone), while the low-order birth falls as its own share law prices it (2500/1438/1125).

F CAPACITY: THE TWO-LEVEL MIN LAW

How many terms can coexist? Scanning offered rules S against hidden width d and readout width r (Figure 3, Appendix G), the number born saturates linearly in d (15/17/28/35/53 at $d = 32/48/64/96/128$) with $\alpha_h = 0.40$, successive doublings giving $15 \to 28 \to 53$—while constraining the readout at d=128 gives capacity linear in r with $\alpha_r = 1.09$. The two levels compose

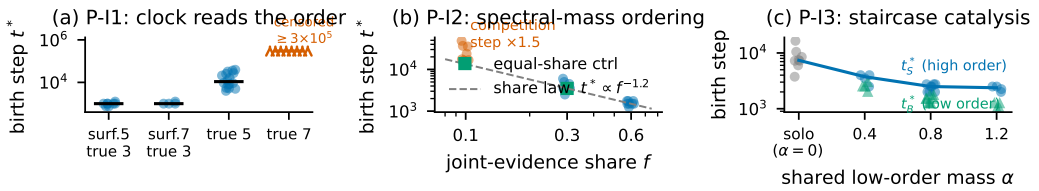

Figure 6: Three preregistered kill shots. (a) P-I1: padded surface arms follow the true degree-3 clock; true degree-7 is censored at 3×10^5 in all seeds ($\geq 300\times$). (b) P-I2: with marginals fixed, birth times follow spectral mass (strict order 8/8 seeds; share exponent 1.2); the weakest rule pays a gap-independent competition step. (c) P-I3: staircase catalysis of the high-order term, monotone in shared mass; the low-order term is priced by its own share.

as $C = \min(\alpha_h d, \alpha_r r)$: at $(64, 48)$ the hidden level binds (predicted ≈ 26, measured 23) despite the wide readout, and at $(1024, 32)$ neither binds.

The hidden constant is not free. On ± 1 inputs a degree-K term is exactly representable by K ReLU units and by no fewer at $K{=}2$ (Proposition 5), so a width-d layer should hold d/K terms on disjoint variables. At $K{=}2$ this gives $\alpha_h^{\mathrm{ReLU}} = 0.500$ with no fitted parameter, against 0.52 measured under the sum-normalized loss.

We state the condition on that agreement, because it is not incidental. Holding the input dimension, the degree, the batch and the learning rate fixed and changing only whether the multi-rule loss is summed or averaged moves the measured α_h from a median of 0.47 (range 0.44–0.53, nine cells) to 0.33 (range 0.18–0.44, seven cells). The second value sits *below* the representation bound, so the parameter-free agreement holds under the sum convention and not in general. This is the same instrument effect described in the next paragraph, seen from the other side: averaging divides each rule's effective learning rate by S, which slows formation and truncates the plateau before the representational ceiling is reached. What the construction bounds is the *representational* capacity d/K; the measured capacity approaches that bound when learning is easy and falls short of it when learning is hard. We therefore claim the form $C = \min(d/m_\sigma(K), \alpha_r r)$ with $m_{\mathrm{ReLU}}(K) = K$ as an upper bound with a learnability discount, not a fitted-free constant. All capacity measurements here are at $K{=}2$; $\alpha_h^{\mathrm{ReLU}}(K) \cdot K = 1$ remains frozen as a prediction at $K{=}3, 4$ and is not yet tested.

An instrument lesson. Averaging the loss over rules scales each rule's effective learning rate as $1/S$, so "birth slows with load" is share dilution, not a rising threshold; fixing supply separates them (1.06 vs. 0.12), and any multi-task curve read without fixing supply conflates the layers (Appendix K).

G Statements moved from the main text

Lemma 2 (Function-space descent)**.** *Let $\Theta_{AB} = \langle \partial_\theta \hat{f}(A), \partial_\theta \hat{f}(B) \rangle$. Under population gradient flow with squared loss, $\dot{\hat{f}} = -\eta\, \Theta\, (\hat{f} - \hat{y})$ with $\Theta \succeq 0$, so $\|\hat{f} - \hat{y}\|$ is non-increasing. The label-independent (self-fit) terms are thus part of one descent flow in function space, not a force opposing the signal—the apparent competition in the single-parameter calculation is an artifact of that projection.*

Proposition 3 (Exclusivity, and the cross-entropy version)**.** *Among Bregman losses, spectral additivity as in Theorem 6 holds for all targets iff the generating function is quadratic (squared loss). For cross-entropy, writing $\Delta_A = \hat{f}(A) - \hat{y}(A)$, near the optimum $L - L^* = \frac{1}{2} \sum_{A,B} \Delta_A\, H_{AB}\, \Delta_B + O(\|\Delta\|^3)$, where H is the softmax Fisher metric expressed in the character basis. The additive form $\frac{1}{2} \sum_A \kappa_A \Delta_A^2$ holds exactly when H is diagonal in that basis—for which invariance of the optimum under the cube's sign-flip (translation) group is a sufficient condition—and otherwise up to the off-diagonal Fisher terms; in practice the Hessian is measured, not assumed diagonal. The identity is thus second-order local for cross-entropy—and this limitation is itself a theorem, not an open item.*

Theorem 8 (Share–threshold decomposition; an exact bookkeeping identity)**.** *When S capabilities share a fixed gradient budget, the log-slope of birth time with respect to S decomposes identically as* slope $= 1$ (*share dilution*) $+$ *threshold term* (*occupancy*)*, where the first term equals 1 exactly*

when the share is $1/S$. Real pretraining fixes the token budget, so the mean-normalized mode is the physical one.

Both terms are separately measurable, and the excess over 1 is then a capacity-proximity meter readable without triggering collapse. Under the mean-normalized loss (share $\propto 1/S$) we measure slope 1.06; under a fixed-supply, sum-normalized loss, which isolates the threshold term, 0.12 on the capacity plateau. So on a fixed token budget "more capabilities each form slower" is mostly share dilution.

Proposition 4 (What the loss cannot identify, arrival times can)**.** *Let a spectrum hold n_K terms of degree $K \le M$, each arriving at rate λ_K in the fluctuation branch, and let the two observables be the loss trajectory $L(t) = L_\infty + \sum_K n_K v_K e^{-\lambda_K t}$ and the arrival-time distribution $T \sim \sum_K (n_K/N)\,\mathrm{Exp}(\lambda_K)$.* (Negative half.) *For every $\epsilon > 0$ there are distinct spectra $\{n_K\} \neq \{n'_K\}$ with $\sup_t |L(t) - L'(t)| < \epsilon$ once $M > C\log(1/\epsilon)$: a loss trajectory identifies $O(\log(1/\epsilon))$ effective parameters and no more—the exponential decay of the Laplace kernel's singular values (Epstein & Schotland, 2008). (A Müntz-type extremal function makes the construction explicit; its coefficients can grow with M, so we claim the parameter count, not a uniform lower bound on the spectra's separation.)* (Positive half.) *If $\lambda_K = \nu p^K$ with fixed $p < 1$—the geometric rate ladder the barrier law supplies—then $\{n_K\}$ is determined uniquely by the law of T for every M (Teicher, 1963), and stably at fixed M: the moment map is a Vandermonde system in the geometrically separated nodes p^{-K}, whose relative condition number at fixed M is controlled by the ratio $1/p$ (it grows with M, as all Vandermonde conditioning does).*

The pair interlocks with the experiments rather than sitting beside them: the measured barrier $\propto K$ is precisely the separation that makes the positive half true, and the paper's clock measurements are all of the second kind. Two boundaries. The negative half's $L(t)$ is the fluctuation-branch idealization; real losses add drift-branch and non-exponential segments, which only worsen the inversion, so the negative direction is if anything understated. And reading rung M from arrival times needs an observation window of order $\lambda_M^{-1} = p^{-M}/\nu$—finite budgets censor the deepest rungs, which is what the $45\times$ budgets of Section 5 are buying.

Proposition 5 (Capacity, hidden level: the cost of one term)**.** *On ± 1 inputs a degree-K interaction term is* exactly *representable by K hidden ReLU units together with the readout bias. Sufficiency is what that construction gives, and it bounds capacity from* below*: a width-d hidden layer hosting terms on disjoint variable sets admits at least $\lfloor d/K \rfloor$ of them. The opposite direction—that K units are also necessary, which is what would make d/K a ceiling—we prove only at $K{=}2$; for $K > 2$ we do not, and d/K should be read there as an achievable rate rather than a general upper bound. For* $\tanh$ *no exact finite representation exists, so the cost carries a surcharge and $\alpha_h^{\tanh} < \alpha_h^{\mathrm{ReLU}}$ at every K.*

Proposition 6 (Capacity, readout level)**.** *Under the random-superposition model—stored feature directions drawn i.i.d. uniformly on the sphere in $\mathbb{R}^d$, readout by* $\arg\max$ *against V candidates, interference treated as Gaussian with variance $1/d$ per stored pair—readout at fixed reliability $1 - \delta$ has the leading-order asymptotic threshold $C \sim d/(2\ln V)$, with δ- and $\ln\ln V$-corrections at lower order and no fitted constant.*

Measured against it: at $d{=}128$, $V{=}64$ the formula gives 15.4 and the observed band is 16–32—the prediction sits just below the band, so the claim is the order and the d-dependence, with the Gaussian-interference constant under-predicting by up to $2\times$.

H PROOFS

H.1 THEOREM 1 (EXACT GRADIENT BLINDNESS)

Write the population gradient of the loss at parameter θ: $\mathbb{E}[\partial\ell/\partial u \cdot g_\theta(x)]$ with $\partial\ell/\partial u$ affine in y; the label-carrying part is $-\mathbb{E}[y\, g_\theta(x)]$ (squared loss; for cross-entropy the same term appears with the label one-hot). With $y = \chi_S$ and g_θ independent of x_i, $i \in S$: since the base measure is a product measure, $\mathbb{E}[\chi_S g_\theta] = \mathbb{E}[x_i]\,\mathbb{E}[\chi_{S\setminus\{i\}} g_\theta] = 0$. $\square$

H.2 THEOREM 2 (EXACT RESPONSIBILITY IDENTITIES)

(a) For any f, the Walsh coefficient obeys $\mathrm{coef}_S[f] = \mathbb{E}\big[\chi_S(x)f(x)\big]$. Apply, for each $i \in S$, the identity $\mathbb{E}_{x_i}[x_i h(x_i)] = \frac{1}{2}(h(1) - h(-1))$: iterating over $i \in S$ yields $\mathrm{coef}_S[f] = \mathbb{E}_{x_{S^c}}[\Delta_S f]/2^K$ where Δ_S is the alternating sum over the 2^K sign patterns of x_S (the K-fold finite difference). For $f = \sigma(w \cdot x + b)$, the mean-value theorem for iterated divided differences ($\sigma \in C^K$) gives $\Delta_S \sigma = \sigma^{(K)}(\xi_x) \prod_{i \in S}(2w_i)$ with ξ_x in the convex hull of the evaluated arguments, whence $f_S(w) = c_K(w) \prod_{i \in S} w_i$ with $c_K(w) = \mathbb{E}_{x_{S^c}}[\sigma^{(K)}(\xi_x)]$.

(b) Differentiate *first*, under the expectation:

$$\frac{\partial f_S}{\partial w_j} = \mathbb{E}\big[\chi_S(x)\, x_j\, \sigma'(w \cdot x + b)\big] = \mathbb{E}\big[\chi_{S\setminus\{j\}}(x)\, \sigma'(w \cdot x + b)\big],$$

using $x_j^2 = 1$ so that $\chi_S x_j = \chi_{S\setminus\{j\}}$. The right-hand side is the Walsh coefficient of σ' on the set $S\setminus\{j\}$ of size $K-1$; applying identity (a) to $\sigma' \in C^{K-1}$ on that set gives

$$\frac{\partial f_S}{\partial w_j} = \mathbb{E}_{x_{(S\setminus\{j\})^c}}\big[(\sigma')^{(K-1)}(\xi'_{j,x})\big] \prod_{i \in S,\, i \neq j} w_i = \kappa_j(w) \prod_{i \in S,\, i \neq j} w_i,$$

with $\kappa_j(w) = \mathbb{E}[\sigma^{(K)}(\xi'_{j,x})]$. The readout responsibility $\partial V/\partial a = f_S(w)$ is (a) itself; no differentiation of any prefactor occurs anywhere in the argument. Boundedness is immediate; as $\|w_S\| \to 0$ the intermediate points converge to b uniformly, so $\kappa_j, c_K \to \sigma^{(K)}(b)$. (Differentiating the *formula* (a) in w_j while holding c_K fixed would be invalid—c_K depends on w; the theorem does not use that route.) □

Remark (degenerate biases). If $\sigma^{(K)}(b) = 0$ (e.g. odd σ at $b = 0$ with K even), the leading prefactor vanishes and the next order takes over; with generically initialized biases this is a measure-zero configuration, and in a wide layer it simply reweights the initialization lottery of Appendix D.

H.3 LEMMA 1 (REDUCTION) AND THEOREM 4 (CLOCKS)

Formal statement of Lemma 1. Assume (i) one unit $\sigma(w \cdot x + b)$ with (ii) $\sigma \in C^{K+1}$, (iii) a linear readout a, (iv) the constant Fourier component projected out (a centered readout; in practice the trained output bias), (v) a generic bias, $\sigma^{(K)}(b) \neq 0$, (vi) the readout oriented so the target-aligned mode grows, (vii) target $y = \chi_S$ with $|S| = K$, (viii) squared loss, (ix) population gradient flow without regularization, (x) symmetric initialization $w_i(0) = a(0) = c > 0$, and (xi) $w_{S^c}(0) = 0$, which part (c) uses; the flow does not preserve (xi) exactly—w_{S^c} drifts under the model-fit gradient, and that drift is not bounded by R_w—so (xi) is one more reason the reduction is conditional. Then: *(a)* the configuration $w_i \equiv w$ for $i \in S$ is exactly invariant, by permutation symmetry, and κ_j is j-independent there; *(b)* the pair (w, a) obeys $\dot{w} = \eta\, a\, \kappa(w)\, w^{K-1} + R_w$ and $\dot{a} = \eta\, c_K(w)\, w^K + R_a$, with $a/w \to 1$ maintained to $O(\theta)$ along small-amplitude trajectories; *(c)* the signal channel of the common amplitude obeys the displayed ODE with $q = K + 1$ and $\gamma_0 = \eta|\sigma^{(K)}(b)|$. *Lemma.* Under squared loss, the population gradient on any parameter θ splits as $\mathbb{E}[(f_{\mathrm{model}} - y)g_\theta] = -\mathbb{E}[y\, g_\theta] + \mathbb{E}[f_{\mathrm{model}}\, g_\theta]$: a label-attributed (signal) part and a label-independent model-fit part.

(a) Invariance. For the signal part, Theorem 2(b) gives, at a configuration with $w_i \equiv w$ for all $i \in S$: $-\partial L/\partial w_j\big|_{\mathrm{signal}} = a\, \kappa(w)\, w^{K-1}$, and κ_j is independent of j there because permuting two coordinates of S maps the configuration to itself and permutes the κ_j; the model-fit part is equally symmetric. Hence equality of the w_i is preserved exactly.

(b) The pair (w, a). The signal equations are $\dot{w}|_{\mathrm{signal}} = \eta\, a\, \kappa(w)\, w^{K-1}$, $\dot{a}|_{\mathrm{signal}} = \eta\, c_K(w)\, w^K$, with $\kappa, c_K = \sigma^{(K)}(b) + O(w)$ (Theorem 2(b), uniform limit). The ratio drifts slowly: $d\ln(a/w)/dt = \eta w^{K-1}\big(c_K w/a - \kappa a/w\big) = O(w^K)$ near $a \approx w$. Bounding the accumulated drift by the growth law is a continuation argument, not a circle: on the maximal interval where $|\ln(a/w)| \leq \delta$ (any fixed $\delta < 1$) the signal equations hold with constants uniform in δ, so $\int w^K dt = O(\theta/\gamma_0)$ there and the accumulated $\ln(a/w)$ is $O(\theta)$; taking θ small enough that this is $< \delta$ prevents the interval from ending before threshold. The symmetric description thus holds as $a = s(1 + O(\theta))$, $w = s$. Substituting into the signal equations gives $\dot{s}|_{\mathrm{signal}} = \gamma_0 s^K(1 + O(s) + O(\theta))$ with $\gamma_0 = \eta|\sigma^{(K)}(b)|$ after orientation; taking the threshold small, the $O(\theta)$ is absorbed into the stated error term.

(c) The model-fit term, computed rather than assumed away. Write $u = \sum_{i \in S} x_i$ and center the unit: the trained output bias removes the constant coefficient $\hat{f}_{\text{model}}(\emptyset)$, so $f_{\text{model}} = a(\sigma(b + su) - \sigma(b))$ on the symmetric configuration. Then, expanding in s,

$$\mathbb{E}[f_{\text{model}}\, g_{w_j}] = a^2\, \mathbb{E}[(\sigma - \sigma(b))\, \sigma'\, x_j] = s^3\, \sigma'(b)^2\, (1 + O(s)), \qquad \mathbb{E}[f_{\text{model}}\, g_a] = s^3\, K\, \sigma'(b)^2\, (1 + O(s)),$$

using $\mathbb{E}[x_j u] = 1$ and $\mathbb{E}[u^2] = K$. Both enter descent with a minus sign: a self-interaction *contraction* of order $\Theta(\sigma'(b)^2 s^3)$, uniformly in K—not $O(s^{2K-1})$. This is why the lemma is stated for the signal channel, with the contraction reported alongside. The comparison against the signal holds for a single unit; the main-text Remark reports the preregistered network-level test, where the summed low-order modes are themselves fitted to zero and the contraction self-cancels. □

Theorem. (i) is the solution of $\dot{s} = \gamma_0 s(1 + O(s))$; the $O(s)$ contributes $o(1)$ to the log clock. (ii) Integrating $\dot{s} = \gamma_0 s^m(1 + O(s))$ between c and θ: $T = (c^{1-m} - \theta^{1-m})/(\gamma_0(m-1)) \cdot (1 + o(1))$, where the correction is bounded by $\sup_{[c,\theta]} O(s)$ because the integrand is multiplied by $(1 + O(s))^{-1}$. (iii) is elementary probability under the stated independence assumption: per-step joint success is $\prod_i p_i$, so T is geometric, memoryless, with $\text{CV} = \sqrt{1-p} \to 1$; the assumption itself is empirical for SGD and is tested, not derived (measured memorylessness and CV, Appendix D). A heuristic for why the high-K regime lands in this class: each small amplitude aggregates many independent mini-batch contributions, so their increments decorrelate by central-limit mixing—a heuristic, stated as such, not a proof. □

H.4 Lemma 2 (low-order relaxation)

(a) With squared loss, Parseval gives $L = \frac{1}{2}\|\hat{f} - \hat{y}\|^2$, so $\dot{\theta} = -\eta \sum_A (\hat{f}(A) - \hat{y}(A))\, \partial_\theta \hat{f}(A)$ and, by the chain rule, $\dot{\hat{f}}(B) = \sum_\theta \partial_\theta \hat{f}(B)\, \dot{\theta} = -\eta \sum_A \Theta_{BA}(\hat{f}(A) - \hat{y}(A))$, i.e. $\dot{\hat{f}} = -\eta\Theta(\hat{f} - \hat{y})$ with Θ a Gram matrix, hence PSD (the Walsh-basis form of kernel-flow dynamics, cf. Jacot et al., 2018). Therefore $\frac{d}{dt}\frac{1}{2}\|\hat{f} - \hat{y}\|^2 = -\eta(\hat{f} - \hat{y})^\top \Theta(\hat{f} - \hat{y}) \le 0$.

□

A conjecture we preregistered and killed. The same equation suggests more: with amplitudes $\asymp s$, $\Theta_{rr} \asymp s^{2r}$ and $\Theta_{r,S} = O(s^{r+K})$, so the degree-$r$ block should relax at rate ηs^{2r} to a quasi-static $O(s^{K-r})$, making every label-independent channel contribute $O(s^K)$—and the regimes tile exactly ($2r < K-1$, where the un-relaxed contribution would dominate, is precisely where relaxation is adiabatic). We froze this and tested it with the exact-enumeration probe. It fails under SGD: the degree-1 coefficient holds a plateau instead of decaying (last-third/first-third ratio 1.1–1.6; jitter $\pm 2\times$; K=2–4), and the self-fit/signal ratio still scales as the un-relaxed s^{3-K} (fitted -1.2 to -1.6 at K=4 vs. the conjecture's 0). Mini-batch noise re-excites the low-order channels as fast as descent drains them, so they equilibrate at a noise-supported level. We note, because a reader may take this balance for temporal correlation in the noise, that it is the opposite: a drained-and-re-excited plateau is exactly what an Ornstein–Uhlenbeck channel with *independent* increments equilibrates to—the stationary width of Proposition 1(a2)—so this observation neither supports nor threatens assumption (T). Proposition 2 states, at outline standing, why the drained-and-re-excited plateau leaves the slow drift unchanged; beyond that outline the network-level behaviour is reported as measured, not derived.

H.5 Theorem 6 and Proposition 3

The identity is Parseval for the orthonormal characters. Exclusivity: for a Bregman loss D_ϕ, take $f = u_0 + t\chi_A + r\chi_B$ ($A \neq B$ nonempty, $y = 0$); on the cube (χ_A, χ_B) is uniform on $\{\pm 1\}^2$, so the population loss at base u_0 is $L(t, r) = \frac{1}{4}\sum_{\pm,\pm} \phi(u_0 \pm t \pm r) - \phi(u_0)$ up to terms linear in the mean-zero characters. Additivity for all (t, r) forces $\partial^2 L / \partial t\, \partial r = 0$, i.e. $\phi''(u_0 + t + r) + \phi''(u_0 - t - r) = \phi''(u_0 + t - r) + \phi''(u_0 - t + r)$: the midpoint equation at every base u_0, so ϕ'' is affine. Convexity of ϕ on $\mathbb{R}$ makes an affine, everywhere-nonnegative ϕ'' constant; hence ϕ is quadratic up to affine terms, which do not affect D_ϕ. The cross-entropy local version is the second-order Taylor expansion of the log-partition function around the optimum: the Hessian is the softmax Fisher information, and expressing the perturbation in the character basis gives the quadratic form $\frac{1}{2}\Delta^\top H \Delta$. H is not diagonal in that basis in general; it is diagonal whenever the optimum is invariant under the cube's sign-flip (translation) group, since the characters are exactly the irreducible representations of that

group and a commuting operator is diagonal in them. (Permutation exchangeability alone only block-diagonalizes by degree.) In the diagonal case the eigenvalues are the weights κ_A. We use the additive form only in that regime and the quadratic form otherwise. □

H.6 THEOREMS 7 AND 8

One hypothesis is used silently in the per-class application of Corollary 1(c): within each class of samples (indexed by the informatively observed subset B), the conditional law of the unobserved coordinates must remain the independent mean-zero product measure. For the injected episodes used in our experiments this holds by construction; for natural data it is part of the reference-measure convention of Definition 1. Classwise application of Theorem 1 shows classes $B \neq S$ contribute zero label-signal; the signal term of the population gradient is thus the full-joint class conditional expectation weighted by its share f. Dilution: non-joint classes contribute to the label-independent part (norm, occupancy) only. For Theorem 8: with S rules sharing a budget, per-rule signal is (share)×(per-sample signal), and the birth threshold rises with occupancy; taking logarithms splits the slope additively into the share term (exactly 1 when share $= 1/S$) and the threshold term. The threshold term's value is measured, not derived. □

H.7 THEOREM 3

By Theorem 4 the B-clock completes at $t^*(B)$; for $t \gg t^*(B)$ the amplitudes of parts in B are $O(1)$, so in Theorem 2's product only parts in $S \setminus B$ remain small: the effective missing-part product has degree $|S \setminus B|$. Time-scale separation ($t^*(B) \ll$ the reduced clock, controlled by mass ratio α/β; matured amplitudes $\Theta(1)$ and slowly varying on the reduced scale) makes the composition a two-timescale sequential composition of the two proven clocks. □

H.8 PROPOSITION 5

Let $u = \sum_{i \in S} x_i$. On $\{\pm 1\}^K$, u takes the $K{+}1$ values $K - 2j$ ($j = 0, \ldots, K$) and $\chi_S = (-1)^j$, so $\chi_S = g(u)$ for the function g with $g(K - 2j) = (-1)^j$. Any continuous piecewise-linear interpolant of $K{+}1$ alternating values has K linear pieces with alternating slopes, hence is realized as $c + \sum_{m=1}^{K} b_m \operatorname{relu}(u - t_m)$ with one threshold in each of the K gaps; the resulting system for (b_m) is triangular and has a unique solution. Since $u = w \cdot x$ with $w = \mathbf{1}_S$, each summand is one hidden ReLU unit and c is the readout bias. Explicitly, $K{=}2$: $t = (-1, 1)$, $b = (-2, 6)$, $c = 1$; $K{=}3$: $t = (-2, 0, 2)$, $b = (2, -6, 10)$, $c = -1$.

Minimality at $K{=}2$. One unit gives $c + b \operatorname{relu}(w_1 x_1 + w_2 x_2 + \beta)$, which must take equal values on $\{(1,1), (-1,-1)\}$ and on $\{(1,-1), (-1,1)\}$. With $a = w_1 {+} w_2$ and $a' = w_1 {-} w_2$, the first equality forces $a = 0$ or $\operatorname{relu}(\pm a + \beta) = 0$, and the second forces $a' = 0$ or $\operatorname{relu}(\pm a' + \beta) = 0$. Every surviving branch makes all four arguments non-positive or both weights zero, so the unit is constant and cannot separate the two pairs. Hence two units are necessary and sufficient at $K{=}2$. For $K > 2$ the construction is an upper bound on cost; we do not prove minimality. □

H.9 PROPOSITION 6

Under the stated model, each stored unit-norm direction receives interference from the other $C - 1$ stored directions; i.i.d. uniform directions have pairwise overlaps that are $O(1/\sqrt{d})$ with sub-Gaussian tails, so the total interference on a readout score is approximately Gaussian with variance $(C-1)/d$. An $\arg\max$ over V candidates fails when interference exceeds the signal margin; Gaussian extreme values put the largest of V competitors at $\sqrt{2 \ln V} \cdot \sqrt{(C-1)/d}\,(1 + o(1))$, so reliability requires $(C-1)/d \cdot 2 \ln V \leq 1 + o(1)$, i.e. $C \leq (1 + o(1))\, d/(2 \ln V)$. The model assumptions (i.i.d. directions, Gaussian interference) are stated in the proposition; the measured band (16–32 born at $d{=}128$, $V{=}64$) sits just above the prediction 15.4, so the asymptotic order and d-dependence are supported while the constant under-predicts by up to $2\times$. □

I THEOREM–EVIDENCE AUDIT

Statement	Rigor status	Experimental photographs
Thm 1 gradient blindness	exact	parts-only arm: zero progress for 8×10^4 steps; P-I2 fixed marginals carry no signal
Thm 2 mutual gating	exact (responsibility identity, any $\sigma \in C^K$)	pre-birth SNR tracks the amplitude product ($-0.25 \to 6.7$ over steps 100–500)
Lem 1 reduction	conditional (assumptions stated); signal-channel ODE + self-fit contraction $\Theta(s^3)$	clock corollaries; β-sweep prefactor $\propto \beta^{-2.5}$; b=0 kills K=2 (6/6); real-machine SNR jump $\times$6 at birth (9/9)
Lem 2 function-space descent	exact	exact-Fourier probe: self-fit is 0.2–2.2$\times$ the signal and fluctuating, not a systematic contraction
Thm 4 three clocks	(i),(ii) exact *within the single-unit reduction* of Lem 1 (multi-unit standing empirical); (iii) conditional (independence, tested)	lr slope -0.86; birth CV 0.13; under SGD batch-free in the drift regime (slopes $0.023/0.021$), under Adam a clock ($-0.450/-0.413/-0.448$)
Prop 1 spectral noise, birth curve	(a) step identity exact; (b) scale function exact within the stated 1-D diffusion, whose domain the width sweep bounds (assumption D); (c) extreme value, one parameter	floor parameter-free at 0.97 of formula (range 0.83–1.26; exponents $-0.545/+0.531$ vs. $\mp 1/2$); all six factors of Co move as predicted (incl. label-noise kill $4/4$); $(1-p)^d$ fits 16$\times$ width range with one p
Thm 6 spectral identity	exact (squared); CE local 2nd order, exclusivity proven	single-rule flat-then-drop; multi-rule smooth sum; mirage adjudication
Thm 7 share law	signal-attributed drift exact; $1/f$ clock scaling conditional (fixed representation)	f-halving $\times$2.1–2.3; warm-ups pure delay; P-I2 ordering 8/8
Thm 8 share–threshold	decomposition exact; threshold value measured	slopes 1.06 vs. 0.12; real-machine $S^{0.86}/S^{1.08}$ (70m/160m, preregistered)
Thm 3 staircase	exact under stated time-scale separation, matured $\kappa \neq 0$, and the reuse clause	I3 catalysis 3$\times$ monotone in α; real-machine three-rung ordering, five sizes and three families, with the successor at exactly zero until its precursor forms
Thm 5 combinatorial barrier	lower bound on the rate (independent-OU premise stated; constants untracked)	high-K ladder log-increment $\approx$ 1.5/rung; real-machine barrier $\propto K$; Adam boundary at $K \approx 8$
Prop 2 network dilution	outline (H1–H4 stated; constants untracked)	plateau probe: self-fit 0.2–2.2$\times$, no systematic sign; width sweep $1.34/2.09$ at d=16/4
Prop 4 identifiability pair	exact (classical tools; premises stated)	the measured barrier $\propto K$ supplies the separation; all clock measurements are arrival-time reads
Prop 5 hidden capacity	exact construction; minimal at K=2; upper bound on cost for $K > 2$	α_h^{ReLU}=$1/K$ is a ceiling: 0.500 at K=2 vs. 0.52 measured under the sum convention, but 0.33 under the mean convention, i.e. the approach to the ceiling is setting-dependent; tanh below ReLU as predicted (0.40)
Prop 6 readout capacity	conditional (random-superposition model); two-level min = empirical	plateaus $15/17/28/35/53$; α_r=1.09; two crossover checks

Table 1: Each claim carries its rigor status and its data. Every conditional statement lists its assumptions explicitly, and every empirical constant is labeled as measured.

J THE FIELD'S BLACK-BOX LIST VS. THIS FRAMEWORK

We audited the framework against a merged list of 20 open training phenomena drawn from recent surveys (Simon et al., 2026; Sharkey et al., 2025) and topic-specific literature. Grade A = quantitative account with numbers checked in this paper; B = mechanism identified, constants not located here; C = outside current scope.

Phenomenon	Grade	What the framework gives
Emergence at scale	A	gating ⇒ blindness ⇒ joint arrival; clock prices the delay
Emergence-vs-mirage debate	A	adjudicated as two layers of Theorem 6
Grokking	B	full condensation cycle; plateau = waiting, jump = event
Training phase transitions	B	threshold collapse width $\propto$ capacity$^{-1/2}$ (predicted)
Induction-head transition	A	strict order with a dead window, three families; fixed multiples are recipe-local (Section 5)
Superposition capacity	A	$C = \min(d/m_\sigma(K), 1.1r)$; both levels derived: $m_{\mathrm{ReLU}}{=}K$, readout $d/(2\ln V)$
Circuit-formation dynamics	A	the main subject of this paper
ICL transience	B	birth (joint, exponential) vs. death (single-part, linear) clocks
Memorization vs. generalization	B	single-sample vs. multi-part terms competing for capacity
Catastrophic forgetting	B	death is single-part linear; constants not located here
Plasticity loss	B	occupancy raises thresholds; linear slowdown measured in lab
Feature learning vs. lazy	B	regime map from the clock boundary
Lottery tickets	B	initialization pre-charge $s(0)$; distribution not derived
Scaling-law exponents	C	needs the measured structure spectrum; inversion from loss ill-posed
Double descent	C	aggregate-layer; outside per-term clocks
Loss spikes; warmup	C	optimizer state $\times$ batch interactions not in the equations
Generalization theory	C	explicitly out of scope
Neural collapse; implicit bias	C	partial geometric remarks only

Table 2: Field black-box audit (20 items, condensed). Much of the emergence family is grade A with numbers in this paper; the aggregate and optimizer families are honestly C.

K Protocol, preregistration ledger, and instrument lessons

Protocol. Every experiment follows: freeze predictions and kill conditions → run → adjudicate against the frozen text → append to the ledger. The ledger (entries 5a–5gn, ∼380 sub-experiments in 23 batches) is append-only and included in the supplement. Surprises trigger, in order: instrument check, mapping to known mechanisms, only then theory revision.

Adjudications reported as misses. The attention laboratory (1-layer, 1-head attention with LayerNorm and residual) did not yield reliable births within 6×10^5 steps in 3/4 seeds at the preregistered budget, and learning-rate warm-up rescues failed in 4/4 configurations; the laboratory redesign is deferred and the negative result stands.

Instrument lessons (all preregistered reruns). (1) *Loss normalization*: mean-over-rules rescales per-rule learning rate by $1/S$, masquerading as a capacity effect; fixed-supply normalization separates share from threshold exactly as Theorem 8 predicts. (2) *Masked-mean normalization* in the multi-rule stream divided out the share being measured; corrected to full-batch normalization, the P-I2 ordering held. (3) *Degenerate rules*: offering $S > n/2$ rules on n inputs creates trivially born rules by index overflow; validity requires $S \le n/2$. (4) *Probe cadence vs. clock*: a fixed probe interval leaves only 4–5 samples before birth at high learning rate, which flattened a fitted floor exponent from 0.35 to 0.29; probe intervals must scale as $1/\eta$. The reported Proposition 1 exponents use the corrected cadence. (5) *bf16 probes*: large-vocabulary softmax scores collapse in bf16; all probes are computed in fp32.

L Hyperparameters and assets

Minimal laboratory (biography, shots). Inputs $n{=}16$ Boolean; hidden 64 $\tanh$ units (width scans 16–128); readout = 2 true classes + 64 distractors ($V{=}64$). Both layers carry trained biases at the framework-default initialization (uniform on $\pm 1/\sqrt{\text{fan-in}}$): hidden biases are therefore nonzero almost surely, meeting Lemma 1's genericity condition $\sigma^{(K)}(b) \neq 0$ (the degenerate $b{=}0$ case of $\tanh$ at even K is measure-zero), and the trained output bias absorbs the constant channel (the Lemma's centering assumption). Loss = binary cross-entropy over class scores (label-affine gradient, Theorem 1); plain SGD, batch 64 (scans 8–256), lr 0.05 (scans 0.0125–0.1); birth = first probe step

with rule accuracy ≥ 0.5 sustained for two consecutive probes (probe every 250 steps); 4–8 seeds per cell.

Spectral laboratory (Lemma 2, Proposition 1). The same inputs and hidden layer, but a *scalar* output $f(x) = \sum_h a_h \sigma(\beta(w_h \cdot x + b_h)) + c$ with *squared* loss $\frac{1}{2}\mathbb{E}[(f-y)^2]$—the setting in which Theorem 6 is an identity and Proposition 1 exact. Because n=16, every Fourier coefficient, Θ_{AA}, $(\Theta^2)_{AA}$ and $\mathbb{E}[r^2]$ is computed *exactly* by enumerating all 2^{16} inputs (no sampling noise), with $(\Theta^2)_{AA} = \mathbb{E}_x[G_A(x)^2]$ obtained from a directional derivative along $\partial_\theta \hat{f}(A)$. Birth = correlation of f with χ_S crossing 0.5 on two consecutive probes. Sweeps: $B \in \{4, \ldots, 1024\}$, $\eta \in \{0.0125, 0.05, 0.2\}$, $K \in \{2, \ldots, 5\}$, $d \in \{4, 16, 64\}$; 3–4 seeds per cell. Clock-ladder and capacity numbers elsewhere in the paper come from the cross-entropy laboratory above; the two differ in prefactors and, as Proposition 1 implies, in where the kinetic-class boundary falls.

Capacity laboratory. Same base; S simultaneous degree-2 rules on n=96–256 inputs (validity $S \leq n/2$); hidden $d \in \{32, 48, 64, 96, 128, 1024\}$; optional linear readout bottleneck $r \in \{8, \ldots, 64\}$; fixed-supply (sum) loss normalization; 10^5–2×10^5 step budget; 3 seeds per cell; capacity = median rules born at the plateau.

Real-machine runs. Pythia-70m resumed from public checkpoints (512–16000) with AdamW (10^{-4}, $\beta = (0.9, 0.95)$, wd 0.01, clip 1.0), batch 16, sequence length 256, fp32; injected skills are marked two-token mappings written into natural-text windows at controlled episode rates; probes are 96–128 fixed sequences, fp32, birth = accuracy ≥ 0.5 on two consecutive probes (probe every 100 steps). Precursor-law measurements use the standard previous-token and induction scores on Pythia 14m–1B over Pile and deduped-Pile checkpoint grids, thresholds crossed in log-interpolation; constants frozen on the fit set before held-out evaluation.

Assets. Pythia checkpoints (Biderman et al., 2023) and OLMo (Groeneveld et al., 2024) are public; all laboratory code is pure PyTorch/NumPy; total compute for the laboratory results is CPU-scale, and each real-machine run fits a single 24GB consumer GPU.