Life of Research · Published 19 November 2026 · DOI 10.5281/zenodo.24355741

A Sharp Turn Is Not a Phase Transition: What Training-Time Transitions in Small Language Models Establish, Why Threshold Readouts, Fast Crossovers, Developmental Stages and Critical Behaviour Are Reported Under One Name, and the Controls That Would Tell Them Apart

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A 2025 study trains a 3.6-million-parameter character-level transformer on Tiny Shakespeare and reports a phase transition in the middle of training: around epochs 230 to 250, the dispersion indices of correct and incorrect word counts exchange which one sits near the Poisson value, divergence from a Poisson fit peaks, the unique misspellings rise and then collapse, and mean correct-word length jumps from about 1.5 to 2.5 characters, while the loss curves stay smooth. Its authors read the agreement of these four probes as evidence of a real reorganisation rather than a metric artefact. In this analysis the author reviews and synthesises published results to show that "phase transition" is used for four different claims in the training-dynamics literature, and that the anchor's evidence supports only the weakest. The four are: a threshold readout, where a nonlinear score turns sharply while the quantity beneath it moves smoothly; a fast crossover, where smooth dynamics move quickly between plateaus; a developmental stage, where a new mechanism forms and can be found and intervened on; and a phase transition proper, a singularity that sharpens as a size parameter grows. The paper derives that, when words are counted in fixed 21-word windows, the dispersion flip follows from arithmetic once the share of correct words passes one half, and the divergence pattern follows from fitting a Poisson curve to what is a bounded count. All four probes therefore depend on one underlying share, two of them by identity, and their agreement does not count as independent confirmation. The record elsewhere is stronger: induction heads, syntactic attention and loss-landscape stages are established developmental events, several of them tested causally, and solvable models and percolation theory give transitions in the strict sense, though along data amount, task count, mixture, model size or a hyperparameter, and in no located study along training time in a real language model. The paper consolidates 30 published results in a table and a figure, states a test for each sense, and gives a classification procedure that would say which sense a reported transition has earned.

Keywords: phase transitions; emergence; training dynamics; developmental interpretability; grokking; in-context learning; evaluation metrics; language models

1. Introduction

The vocabulary of statistical physics entered the study of language models through scale. Wei et al. (2022) called an ability emergent if it is absent in smaller models and present in larger ones, so that it could not be forecast by extrapolating the smaller models, and that paper itself described the step-like plots as phase transitions. Ganguli et al. (2022) framed the same observation as a policy problem: loss on the training distribution falls predictably with scale, while specific capabilities arrive unpredictably. Schaeffer et al. (2023) replied that much of the sharpness was put there by the evaluation. Exact string match and multiple-choice grade, which are nonlinear or discontinuous in the per-token error rate, accounted for more than 92 percent of the emergent abilities that annotators had flagged in BIG-Bench, and switching to a continuous score made several of them smooth.

A parallel literature uses the same words for something else: change over training time within one run. Grokking, the delayed jump from chance to perfect test accuracy on small algorithmic tasks long after the training set is fit (Power et al., 2022), was the first widely discussed case. Induction heads appear in the same short window of training as a sharp gain in in-context learning, which shows up as a bump in the loss curve (Olsson et al., 2022). Masked language models acquire syntactic attention structure in a brief window together with a steep loss drop (Chen et al., 2024). Loss-landscape degeneracy, measured by the local learning coefficient, divides training of a small transformer into stages that line up with changes in its internal computation (Hoogland et al., 2024). In each case the authors speak of a phase transition, a phase change, a breakthrough or a stage, and the terms are often used interchangeably; Chen et al. (2024) say so explicitly.

Hong and Hong (2025) bring the two literatures together. They train a small GPT-style character model on Tiny Shakespeare and track the words it produces over training. They report that four statistics of the generated text reorganise together in one narrow band of epochs, that the loss curves show nothing at that point,

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

and that the coincidence answers the metric-artefact critique of Schaeffer et al. (2023). They draw three conclusions: phase transitions are not unique to large models, they can be detected on a linear training axis without logarithmic rescaling, and they occur early in training as coherence emerges.

This paper does not dispute that something happens in that model around epoch 240. Its question is what kind of something the evidence shows, and its contribution is to separate four claims that the shared vocabulary runs together. A reported transition can be a property of the readout, a property of fast but smooth dynamics, a property of the model's developing mechanism, or a property of a limit in which a quantity becomes non-analytic. These carry very different consequences. A readout transition says nothing about the model; a crossover predicts that a finer probe will see continuous change; a developmental stage predicts that something specific can be located and manipulated; a phase transition proper predicts scaling laws for its width and location. Treating them as one has let weak evidence for the first stand in for the last.

In this analysis the author reviews and synthesises published results to show four things. First, the probes in the anchor study are all functions of one quantity, the share of generated words that are in the corpus vocabulary, and two of the four are tied to that share by arithmetic, so the synchronisation the anchor offers as converging evidence is the expected behaviour of a single smoothly changing share (Section 4). Second, the observation that a transition is invisible in the loss neither supports nor undermines its reality, because aggregate losses both hide genuine breakthroughs and average over offsetting trends that can manufacture apparent ones (Section 5). Third, fast crossovers with no singularity are common, are produced by exactly solvable models, and can look identical to transitions at the resolution of training curves (Section 6). Fourth, the strict sense has been earned only in solvable or idealised models and in transitions along data, task-count or hyperparameter axes; no located study has shown that a training-time transition in a real language model sharpens with model size (Sections 7 to 9). The paper consolidates the measured values in Table 1 and Figure 1, sets out what each sense predicts in Table 2, and states a classification procedure as Algorithm 1.

The paper proceeds from the anchor outward. After a description of the anchor's measurements (Section 2) and a definition of the four senses (Section 3), Section 4 examines how the anchor's probes are built. Sections 5 to 8 then take, in turn, aggregation, fast crossovers, developmental stages and the strict sense, and Section 9 relates all of this to emergence over scale. The consolidated values (Section 10) and the classification procedure (Section 11) follow, then objections to the argument (Section 12), open questions (Section 13) and a short conclusion (Section 14).

2. What the anchor result measured

The model in Hong and Hong (2025) has an embedding dimension of 192, 8 layers and 6 attention heads, about 3.6 million parameters, and a context of 128 characters. It is trained on Tiny Shakespeare, about 1.1 million characters drawn from 65 distinct symbols, with 5 independent seeds over 0 to 600 epochs. At each checkpoint the authors sample 30,000 tokens with a fixed decoding configuration, which the methods section gives as "temperature $T = 1.0$, greedy top-1 unless otherwise noted". The two parts of that description are not compatible: greedy decoding takes the most probable character and makes temperature irrelevant, while sampling at temperature 1.0 draws from the full distribution. The paper does not say which was used for which figure, and this matters below, because the two produce very different text from the same model.

The generated text is split into words at whitespace and punctuation. A word is labelled correct if it appears in the corpus vocabulary and incorrect otherwise; the authors acknowledge that this misclassifies rare valid words. From these labels the paper computes four families of statistics. The unique counts of correct and incorrect word types measure diversity. The mean length of correct words measures compositional complexity. The text is cut into non-overlapping windows of 21 words, and for each class the number of words per window is summarised by its index of dispersion, variance over mean, against the value of one that a Poisson count would have. The same count histograms are compared with a fitted Poisson distribution by Kullback-Leibler divergence.

The reported signature has four parts, all placed in one band, epochs 230 to 250, and averaged over seeds with standard-deviation bands. Correct-word dispersion starts near one and ends below one, with a temporary move back toward one inside the band; incorrect-word dispersion starts below one and ends near one. Divergence from the Poisson fit rises sharply for correct words and falls toward zero for incorrect words. The unique incorrect vocabulary plateaus, peaks near epoch 250 and then collapses, while the unique correct vocabulary accelerates between roughly 250 and 400. Mean correct-word length rises along an S-curve from about 1.5 to about 2.5 characters. A case study tracks the strings "y", "yo" and "you": the prefixes decline as the full word rises. Throughout, the authors state that the training and validation losses are smooth and show no corresponding feature.

The interpretive claims are as important as the measurements. The authors argue that the probes are "continuous, internal statistical measures", that they measure different properties of the output, and that their simultaneous change therefore "rules out metric artifacts". They read the temporary rise in errors during the band as barrier crossing in the sense of a first-order

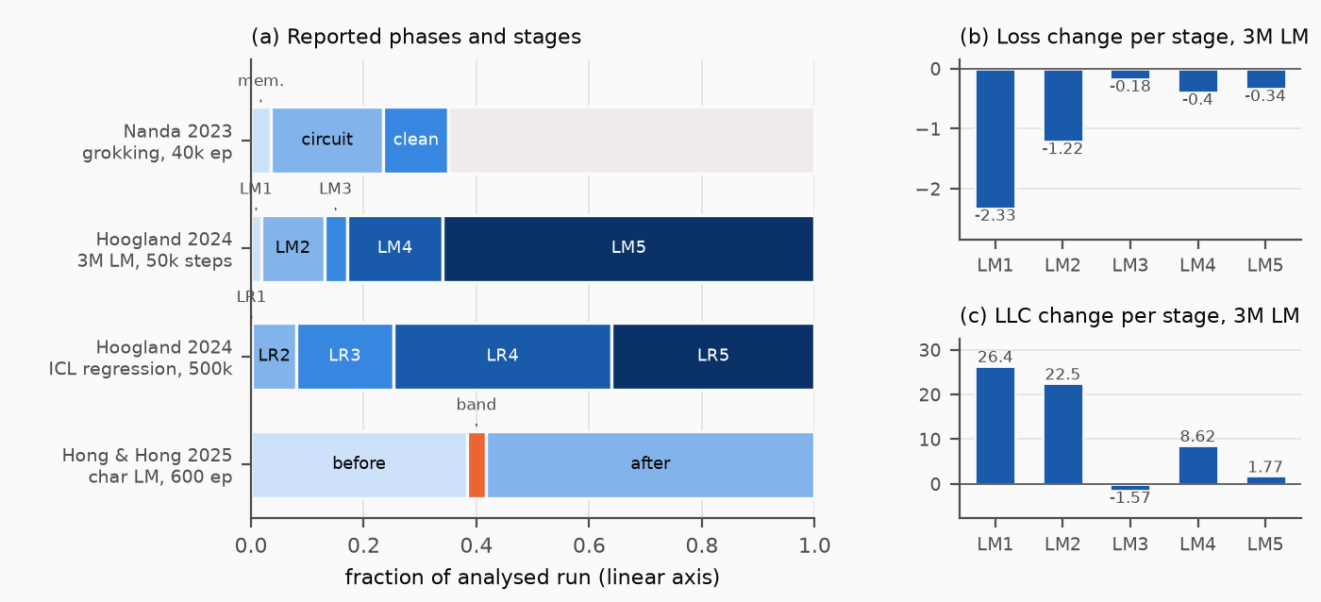


Figure 1. (a) Phase and stage boundaries reported on a linear axis, as fractions of each analysed run: grokking in a one-layer transformer on modular addition (Nanda et al., 2023; 40,000 epochs), local-learning-coefficient stages of a 3M-parameter language model (Hoogland et al., 2024; 50k steps) and of an in-context regression model (Hoogland et al., 2024; 500k steps), and the transition band of the anchor (Hong and Hong, 2025; 600 epochs); the grey segment is the final stable phase after cleanup. (b) Change in loss and (c) change in the estimated local learning coefficient over each of the five stages of the 3M-parameter language model, as tabulated in Hoogland et al. (2024), Figure 1.

transition, citing the mapping of grokking onto a first-order transition by Rubin et al. (2024). They name the dispersion as their "primary order parameter" and training epoch as the control variable. In their limitations they note a single architecture and dataset, character-level tokenisation, an untested dependence on decoding, and the absence of any analysis across model sizes; their future work proposes finite-size scaling to tell genuine critical behaviour from a threshold effect.

Two features of the record should be kept in view. The paper uses "epoch" and "step" interchangeably for the training axis, so the unit of the band at 230 to 250 is not fixed by the text. The location of the band is not fixed either: the summary places it at 230 to 250, but the snapshot section labels step 300 "approaching transition" and says of step 350 that "This checkpoint corresponds precisely to the epochs where dispersion and KL divergence exhibit synchronized cusps". And none of the four statistics is computed from the model's probabilities: all are computed from sampled or decoded text after a binary vocabulary test. Those two facts shape what the evidence can show.

3. Four things called a phase transition

The author distinguishes four claims that a reported transition can carry. Each is stated with what it would predict and what would refute it.

T1, threshold readout. A latent quantity, such as the per-token probability of the correct continuation, changes smoothly. The reported score is a nonlinear or discontinuous function of it, such as exact match or a count above a cutoff, and turns sharply as the latent quantity crosses the region where the function is steep. Under T1 the sharpness belongs to the instrument. The prediction is that a score linear in the latent quantity, or a direct measurement of it, shows smooth change, and that the location of the turn moves when the threshold moves. Schaeffer et al. (2023) state the mechanism for model scale; nothing in it depends on the axis being scale rather than training time.

T2, fast crossover. The latent quantity itself changes quickly, after a plateau, but along a smooth trajectory with no singularity at any size. Plateaus followed by rapid drops arise in exactly solvable deep linear networks (Saxe et al., 2014) and in sigmoidal skill acquisition in a solvable multilinear model (Nam et al., 2024). Under T2 a probe with enough time resolution sees a steep but differentiable curve, and the width of the change, measured in absolute units rather than as a fraction of the run, does not shrink toward zero as the system grows. The distinction matters: in deep linear networks the transition becomes "exceptionally brief relative to the total training time" as the initial mode strength goes to zero (Saxe et al., 2019), because the plateau lengthens while the transition itself does not shorten, so a T2 model can look like a step in rescaled time.

T3, developmental stage. A new internal mechanism forms during a bounded interval, and its formation changes behaviour. The paradigm is the induction head (Olsson et al., 2022). A T3 claim predicts that the mechanism can be located, that intervening on it changes the behaviour, and that the stage recurs across

seeds. It need not involve any singularity: a stage can be a T2 crossover in time that happens to install a qualitatively new circuit. T3 is a claim about the model's structure, T2 about the shape of a curve, and the two are independent.

T4, phase transition proper. In statistical physics a phase transition is a point of non-analyticity in a thermodynamic quantity, and it exists strictly only in a limit, usually of infinite system size. At finite size it is rounded, and the rounding shrinks with size in a way described by finite-size scaling. A T4 claim therefore needs a control parameter, an order parameter, and either an analytic limit, of system size or of a structural parameter, or a demonstration that the transition sharpens with system size, with the width measured in absolute units. First-order transitions add coexistence of two phases and hysteresis; continuous transitions add diverging fluctuations and characteristic exponents.

The senses are not exclusive. A developmental stage may be revealed by a threshold readout; a crossover in time may be the finite-size shadow of a true transition in data amount. Their value is that each assigns a different follow-up. T1 is settled by changing the metric. T2 is settled by resolving the curve. T3 is settled by finding and manipulating a mechanism. T4 is settled only by a size sweep or a derivation. Lavie et al. (2026) state the central difficulty plainly when contrasting their own analysis with timescale results for induction heads: such characterisation "cannot distinguish a true phase transition from a crossover, as both would have tight asymptotic timescales". Table 2 sets out the predictions side by side.

4. Four probes, one share

The anchor offers the agreement of its probes as its defence against the metric-artefact reading: four statistics that measure different things all change in the same band. That argument depends on the probes being independent. On the plain reading of the method in Section 2, they are not.

Start with dispersion, which the anchor treats as its primary order parameter. Each 21-word window contains some number C of correct words and I of incorrect words, and C + I = 21 in every window. It follows that the variance of C across windows equals the variance of I, whatever the dependence between words. The index of dispersion of each class is that common variance divided by the class's own mean. So the class with the larger mean always has the smaller index, and the two indices cross exactly when the mean share of correct words passes one half. If words were independently correct with probability p, the count of correct words in a window would be binomial, and the indices would be 1 - p for correct words and p for incorrect words. Both are below one for every p strictly between zero and one, so "sub-Poisson" is the default for a bounded count, not a sign of order; a class looks "near-Poisson" exactly when it is rare. Early in training, when correct words are rare, the binomial bookkeeping alone puts correct-word dispersion near one and incorrect-word dispersion well below it. Late in training the positions are reversed. That is the anchor's dispersion flip, in its direction and its form, from the arithmetic of a fixed-size window with no structure in the text at all.

The divergence probe follows the same share. A binomial count with W trials and success probability p, compared with a Poisson distribution of the same mean, diverges little when p is small and increasingly as p rises toward one, because the binomial becomes much narrower than a Poisson of that mean. The author checked the direction numerically for W = 21. Applied to the two classes, the correct-word divergence should start near zero and rise steeply as correct words become common, and the incorrect-word divergence should start high and fall toward zero as errors become rare. That is the anchor's "divergence-convergence duality". It is the same share again, seen through a misspecified null: the Poisson baseline is the wrong reference for a count bounded by the window size, and departures from it are guaranteed whenever a class is common.

The one feature of the dispersion record that binomial bookkeeping does not produce is the temporary move of correct-word dispersion back toward one inside the band. A possible reason is heterogeneity between windows. If the probability that a word is correct varies from window to window, because some stretches of generated text are coherent and others are fragments, the variance of the count gains a term proportional to the variance of that probability across windows, and the index rises above its binomial value. That variance is plausibly largest when the model is midway between producing mostly fragments and mostly words, which is when the mean share is moving fastest. This is an untested mechanism, but it needs no barrier and no metastable state: any change from one output mode to another, smooth or abrupt, mixes the two modes for a while.

The other two probes are not tied to the share by an identity, but both depend on it strongly. Mean correct-word length is a conditional mean under a selection that relaxes as the model improves. A generated string of length L enters the "correct" class only if every one of its characters matches some vocabulary word; in a stylised model where each character is right with probability q, the chance is roughly q to the power L, with a small allowance for chance hits. Early on, only one-letter words and short fragments pass the test, so the mean is pinned near the shortest lengths. As q rises, longer words pass at a rate that grows geometrically in L, and the mean climbs along an S-curve in q. This is the exact-match mechanism of Schaeffer et al. (2023), whose model gives accuracy on a target of L tokens as the per-token success rate raised to the power L, applied here to

Table 2. What each sense of "phase transition" predicts, and the observation that would discriminate it (author's classification).

Sense	What is sharp	Predicts	Key test
T1 threshold readout	the score	latent smooth	linear or direct metric
T2 fast crossover	the latent curve	finite width at all sizes	finer time resolution
T3 developmental stage	the mechanism	locatable, manipulable	ablation, clamping, seeds
T4 transition proper	a limit	width shrinks with size	size sweep or derivation

words instead of answers. The unique incorrect vocabulary has the same structure in reverse: when outputs are short and repetitive, there are few distinct misspellings; as the model attempts longer strings, the number of distinct wrong strings grows combinatorially; as per-character accuracy approaches one, wrong strings become rare. A rise and fall is the expected shape under steady improvement.

The point is not that the anchor's model underwent no reorganisation. The point is that all four statistics are functions of the decoded text after a binary vocabulary test, two are tied by arithmetic to one share, and the other two respond to that share through an exact-match nonlinearity. Their coincidence in one band is what a single smoothly rising quantity would produce, and it therefore cannot do the work of independent confirmation. Lubana et al. (2024) make the relevant distinction in their definition of emergence: simultaneous improvement on several tasks is suggestive because the tasks are different and share only an underlying structure. Four summaries of one set of samples are not several tasks.

The probes also do not answer the metric critique on its own terms. The anchor describes its measures as continuous, and they are continuous in the sense of taking real values. But Schaeffer et al. (2023) located the problem in the mapping from the model's per-token behaviour to the score, not in whether the score is real-valued; an average of exact-match outcomes is also real-valued. The test their account implies is to compute a score that is linear in per-token success, such as per-character log-likelihood on held-out text or the expected number of characters matching a vocabulary continuation, and see whether it turns sharply in the band. The anchor reports that its loss curves, which are per-character cross-entropies, are smooth. Under T1 that is not a puzzle to be explained by a hidden reorganisation; it is the expected result.

Decoding adds a further layer. If the reported samples were greedy, early-training text from a weak model will repeat. Holtzman et al. (2020) showed that likelihood-maximising decoding yields text that is bland and repetitive even from strong models, and that decoding choices alone can change text quality dramatically from one fixed model. Gopalani and Hu (2025) found that during loss plateaus small transformers produce outputs with a strong repetition bias, accompanied by hidden states for different tokens becoming nearly parallel, and they report the same repetition and collapse in the early pre-training of Pythia and OLMo. The anchor's pre-transition picture, incorrect words "dominated by short, recurring fragments like 'th' or 'yo'", is what either mechanism predicts, and its end is what the decline of repetition predicts. That does not make the end of repetition uninteresting; Gopalani and Hu (2025) tie it to the slow learning of attention maps, which is a T3 question. It does mean that the dispersion of incorrect words partly measures the decoder and the repetition bias, not the vocabulary.

Hong and Hong (2025) foresee some of this. They concede that decoding might shift the apparent transition region and that the link between their statistics and internal representations is open. The analysis here goes further in one respect: the agreement of the probes, which the anchor treats as the strongest part of its evidence, follows from their construction.

5. What aggregate losses hide and what they make

The anchor's other argument is that the transition is invisible in the loss and visible only to tailored probes. Two bodies of work bear on this, and they point in opposite directions.

The first shows that aggregate loss hides real breakthroughs. Kangaslahti et al. (2025) decompose changes in loss between checkpoints along directions of the low-rank training subspace and cluster examples by their decomposed trajectories. In a 3-layer, 40-million-parameter model trained on English Wikipedia, the exact loss curves were smooth, but clusters along individual directions showed sudden drops at different points in training, and the clusters corresponded to grammatical constructions such as predicting "to" and "from" after a first clause. In an arithmetic task, clustering on exact loss curves recovered digit positions but missed the skill of carrying, which the decomposition recovered. The authors read the result as support for the view that learning in high dimensions proceeds through many breakthroughs at different scales. The quantization model of Michaud et al. (2023) gives a theory of the same picture: if skills are learned as discrete chunks in order of how often they are used, and use frequencies follow a power law, then the aggregate loss falls as a smooth power law built entirely from sharp steps. Hoogland et al. (2024) supply a concrete example at the anchor's scale: in their stage analysis of a 3-million-parameter language model, the recorded change in loss over the two stages in which the induction circuit forms

was -0.18 and -0.40, against -2.33 and -1.22 over the earlier bigram and n-gram stages (Figure 1b).

The second body shows that aggregates also manufacture sharpness. Wu and Lo (2025) grouped the 14,042 MMLU questions into ten difficulty levels and evaluated 56 models. Hard questions showed U-shaped scaling and easy ones an inverted U followed by steady gains. Summed, the two offset each other, producing the long flat stretch and sudden rise that is reported as emergence, and the soar began roughly where easy questions returned to standard scaling. Schaeffer et al. (2024) traced the unpredictability of multiple-choice benchmarks to the sequence of transformations from log-likelihood to accuracy, and to the fact that accuracy depends on how probability is distributed over a few specific wrong answers. Arnold et al. (2024) adapted physics methods for detecting transitions to the public training checkpoints of the Pythia suite (Biderman et al., 2023). Peaks in their distributional dissimilarity for short prompts, at around 20K, 40K and 80K training steps (which the authors call epochs), turned out to be outliers: the output distribution changed at one checkpoint and returned almost at once, and the authors state that these peaks "do not mark transitions between two macroscopic phases of behavior". Baherwani et al. (2026) evaluated single-sample capabilities across ten seeds of Pythia models from 14 to 410 million parameters and found that the same capability appeared at different points in training for different seeds, never appeared for some seeds at small scale, and appeared earlier on average in larger models. Any average over seeds blurs such events, and any single seed can show a step that its siblings place elsewhere.

Together these results remove the evidential weight from the phrase "invisible in the loss". A breakthrough in a subpopulation would be invisible in the loss, so the anchor's observation is compatible with a T3 event. A threshold readout of a smoothly improving share would be invisible in the loss for the reasons of Section 4, so it is equally compatible with T1. What would discriminate them is a decomposition of the loss itself, such as the per-word or per-position loss on correct vocabulary continuations, or the decomposition of Kangaslahti et al. (2025), applied in the anchor's band. If a cluster of examples shows a drop there, the event is in the model. If the per-character loss falls smoothly everywhere, the event is in the probes.

6. Fast crossovers and hidden progress

The second alternative to a phase transition is a crossover: fast change along a smooth path. The literature on learning dynamics shows that this is the generic shape of feature learning, and that it often hides continuous progress beneath an apparently discontinuous curve.

Saxe et al. (2014) solved the gradient-descent dynamics of deep linear networks exactly. Although the input-output map is linear, the weight dynamics are not, and they display long plateaus followed by rapid transitions to lower error. Each mode of the input-output correlation is learned along a sigmoid in time whose onset depends on the strength of the mode and the initialisation. Saxe et al. (2019) used the same solutions to explain rapid developmental transitions in the hierarchical differentiation of concepts, with "semantic illusions" between them. Everything in these models is analytic. There is no singularity at any size, yet the curves show plateaus and jumps that would satisfy most operational definitions of a transition.

Nam et al. (2024) built a solvable multilinear model in which each skill is a basis function, and derived the emergence of skills as sigmoidal functions of training time, data and model size. A single fit parameter matched the emergence of multiple skills in a two-layer network trained on multitask sparse parity. Reddy (2023) traced the abrupt formation of an induction head in a minimal attention network to the sequential learning of three nested logits, and proposed that sharp transitions in attention networks arise from chains of nested nonlinearities that must each be learned in turn. Gopalani et al. (2024) found a long loss plateau followed by a sudden drop in a BERT model trained on matrix completion, with the model moving from copying its input to predicting masked entries and its attention heads moving to task-relevant patterns. Gopalani and Hu (2025) examined such plateaus more closely and found that the optimal attention pattern begins forming during the plateau: in their words, "though the loss dynamics are abrupt, attention progress measure as well as repetitions and representation collapse are not". Biasing attention toward or away from the optimal pattern shortened or lengthened the plateau.

Two results on algorithmic tasks show the same structure from the side of hidden progress. Barak et al. (2022) found that neural networks learning sparse parities show discontinuous jumps in their training curves, at iteration counts near the computational limit, but proved that stochastic gradient descent makes continual progress throughout, amplifying the right features through a gap in the Fourier spectrum of the population gradient that loss and error do not register. Nanda et al. (2023) reverse-engineered a one-layer transformer that groks modular addition, trained for 40,000 epochs on 30 percent of the pairs modulo 113, and defined progress measures from the learned Fourier algorithm. Training divided into memorisation (epochs 0 to 1.4k), circuit formation (1.4k to 9.4k) and cleanup (9.4k to 14k), and the progress measures improved continuously before the test loss dropped. Their conclusion is that grokking, viewed through the right instrument, is gradual amplification followed by removal of the memorising components (Figure 1a).

Grokking has been explained in several ways, and they share this conclusion about shape. Levi et al. (2024) showed that grokking occurs in linear networks learning linear tasks with Gaussian inputs, and that the sharp rise in test accuracy there can be an artefact of the accuracy measure rather than a shift from memorisation to understanding. Kumar et al. (2024) attributed grokking to a transition from lazy dynamics, in which the network fits with its initial features, to rich feature learning, with the delay controlled by the rate of feature learning and the alignment of the initial kernel with the target. Liu et al. (2023) traced it to a mismatch between training and test loss as functions of weight norm, and induced grokking on image, language and molecule tasks while removing it from algorithmic ones. Varma et al. (2023) explained it by a competition between a memorising circuit and a more efficient generalising one, predicted a critical dataset size, and produced "ungrokking" and "semi-grokking". Huang et al. (2024) extended the circuit-competition account to double descent and to emergent abilities in multi-task settings. Each of these mechanisms produces a steep curve from smooth dynamics; none requires a singularity.

The relevance to the anchor is direct. A character model learning word-level structure is learning higher-order statistics after lower-order ones, and Section 7 shows that this progression has been documented as a sequence of stages. Each stage, viewed in one statistic, can look like a jump. Whether the anchor's band is T2, T3 or T1 depends on measurements the paper did not take; that some quantity changes quickly in the band is not in doubt; that the change is T4 has no support in the paper.

7. Developmental stages are real, and they come with mechanisms

The strongest evidence that training-time transitions are more than readouts comes from studies that found a specific mechanism forming in a bounded window, and then manipulated it.

The induction head is the canonical case. Olsson et al. (2022) found that these heads take shape in the same narrow stretch of training in which in-context learning improves abruptly, and offered six lines of preliminary evidence for the hypothesis that they account for most in-context learning; the evidence was causal for small attention-only models and only correlational for larger models with MLPs. Tigges et al. (2024), tracking circuits through 300 billion tokens of training in models from 70 million to 2.8 billion parameters, report that task abilities and the components supporting them emerge at similar token counts across scale, consistent with the earlier finding that induction heads emerge at roughly 2 to 5 billion tokens in models of all sizes. Singh et al. (2024) clamped activations during training on synthetic data and identified three interacting subcircuits that must form for an induction head to appear; their interaction yields the phase change, and the subcircuits explain why its timing depends on the data. Edelman et al. (2024) trained transformers on sequences from random Markov chains and found a uniform stage, a stage using in-context unigram statistics, and a rapid transition to the correct bigram solution, with evidence that the simpler unigram solution delays the bigram one.

Chen et al. (2024) is the most thoroughly controlled case. In BERT-base, trained with three seeds and analysed over 300K steps, syntactic attention structure appears abruptly, together with a precipitous loss drop that the authors place between steps 20K and 30K across seeds, and grammatical capability measured on BLiMP improves just after. The authors then addressed each alternative that Section 3 raises. They replaced accuracy with continuous alternatives, the attention weight on the correct syntactic neighbour and the relative likelihood of the correct member of a minimal pair, and the transition remained. They replaced the step axis with weight norm, distance from initialisation and optimisation path length, and the transition remained. They suppressed syntactic attention during training and showed that it is necessary for the later grammatical capability, and that briefly suppressing it improves final quality. They also note that, across random pretraining seeds, syntactic attention metrics and model capabilities are not correlated, which is part of why they argued from the training trajectory rather than from finished models. This is what a T3 claim looks like when its authors have tried to break it.

Loss-landscape geometry provides a second, mechanism-agnostic instrument. Lau et al. (2024) introduced the local learning coefficient as a complexity measure from singular learning theory, with a scalable estimator. Hoogland et al. (2024) estimated it through training for a two-layer attention-only language model of about 3 million parameters, trained on a subset of the Pile, and for an in-context linear regression model of about 50 thousand parameters. Plateaus in the coefficient, found after light smoothing and differentiation with respect to log training time, divided the language model's 50k steps into five stages ending at 900, 6.5k, 8.5k, 17k and 50k steps. In the first the model learns bigram statistics, reaching a bigram score 0.3 nats above the irreducible entropy of the empirical bigram distribution; in the second it learns common n-grams; in the third and fourth it forms the induction circuit. The regression model's stages ended at 1k, 40k, 126k, 320k and 500k steps. The authors call these "approximate developmental stages", note that the resolution of the boundaries depends on checkpoint density and estimation precision, and report that similar divisions appear across seeds. Wang et al. (2024) refined the coefficient to individual components and traced the differentiation and specialisation of attention heads over training, finding a previously unidentified multigram circuit. Carroll et al. (2025) used it to explain a transient phase in which regression transformers first behave like ridge regression and then specialise to their training tasks, through an evolving

trade-off between loss and complexity.

A broader literature documents the order in which language models acquire statistics. Belrose et al. (2024) found that early in training, networks handle maximum-entropy samples that share only the low-order moments of the training data about as well as real data, and handle them worse later; they also found that Pythia models' loss on sequences sampled from unigram and bigram models reaches a trough between 2 to the 6 and 2 to the 8 steps, rises until about 2 to the 10, and then falls again as in-context learning takes over. Rende et al. (2024) showed that transformers trained on natural language learn interactions among tokens in order of increasing degree, saturating the low-degree ones while continuing on higher ones. Nguyen (2024) measured how transformer predictions move from simple to more complex n-gram rules during training. Choshen et al. (2022) found that language models with different initialisations, architectures and data acquire linguistic phenomena in a similar order, along a trajectory that is approximately one-dimensional. Achille et al. (2019) showed that early training has critical periods during which a deficit has lasting effects.

Transience complicates the picture. Singh et al. (2023) found that in-context learning, once emerged, can disappear in favour of in-weights learning while the training loss keeps falling, and Chan et al. (2022) showed that whether in-context learning emerges at all depends on distributional properties such as burstiness and a long tail of rare classes. A stage is not a permanent acquisition, and its existence depends on the data.

Set against this record, the anchor's three contributions are not new. Developmental transitions have been documented in models of a few million parameters (Hoogland et al., 2024; Olsson et al., 2022) and in small algorithmic models (Power et al., 2022; Nanda et al., 2023). They have been located on linear step axes (Chen et al., 2024; Nanda et al., 2023), and the stage boundaries of Hoogland et al. (2024), though found by differentiating with respect to log time, are reported in steps. And they occur early: the bigram and n-gram stages of Hoogland et al. end within the first 13 percent of the run, and the anchor's band sits at roughly 38 to 42 percent of its 600 epochs (Figure 1a). The anchor's transition, from character fragments to short words, is in content a passage from low-order to higher-order statistics of the kind these studies document. What would turn it from a plausible instance of T3 into a demonstrated one is the apparatus Chen et al. (2024) applied: a mechanism located in the weights or activations, an intervention on it, and the transition surviving a change of metric.

8. Where the strict sense has been earned

The strict sense, T4, has been demonstrated, but under conditions different from the anchor's. The pattern across these cases is instructive.

In solvable models, transitions are derived in a limit. Cui et al. (2024) analysed a single dot-product attention layer with tied, low-rank query and key matrices in the high-dimensional limit with a proportional number of samples. They characterised the global minimum of the empirical loss in closed form and found that it is either a positional mechanism or a semantic one, with a transition from the first to the second above a threshold in sample complexity. Lavie et al. (2026) derived a closed-form posterior over the attention matrix of a single-layer network trained on a copy task and reduced it to two order parameters. Softmax attention showed a first-order transition in the amount of training data, with two minima becoming equally deep at about 300 samples in their setting, while linear attention showed a second-order transition followed by a smooth crossover. Rubin et al. (2024) applied an adaptive-kernel theory of feature learning to teacher-student models and mapped grokking onto a first-order transition, with the post-grokking network analogous to a mixed phase. Zunkovic and Ilievski (2022) analysed solvable rule-learning models and derived critical exponents and the distribution of grokking times. Ziyin and Ueda (2022) proved that the competition between prediction error and model complexity in a regularised loss yields a second-order transition for networks with one hidden layer and a first-order one for deeper networks.

In empirical systems, the strict sense has been approached through control parameters other than time, with predictions that can be checked. Raventos et al. (2023) found a task-diversity threshold for the emergence of in-context learning in linear regression, between 2 to the 14 and 2 to the 15 pretraining tasks at dimension 8, which weight decay lowered and added layers or width raised. Lubana et al. (2024) modelled the acquisition of a formal language's type constraints as bond percolation on a bipartite graph of entities and properties, and the theory predicted that the onset of emergent performance scales with the square root of the number of properties, which their experiments confirmed. Gu et al. (2025) trained models on a synthetic biography dataset mixed with web data and found abrupt transitions in memorisation as model size or the mixing ratio crossed a critical value, with the critical ratio following a power law in model size, which they explain by a knapsack-like allocation of bounded capacity. Liu et al. (2022) mapped phase diagrams of learning performance over hyperparameters, finding comprehension, grokking, memorisation and confusion phases. Kirsch et al. (2022) characterised transitions between memorising, generalising and failing meta-learned algorithms induced by model size, number of tasks and meta-optimisation. Nakaishi et al. (2024)

found a transition in generated text as a temperature-like decoding parameter varies, with power-law statistics resembling natural language at the critical point.

Finite-size scaling, the standard empirical test for T4, has rarely been applied to neural network training at all. The only located attempts are two 2026 single-author preprints on grokking. Wang (2026a) reports finite-size scaling of gradient avalanche statistics across eight model scales, with an effective dimensionality crossing one at generalisation onset. Wang (2026b) reports the same crossing in transformers on modular addition and MLPs on XOR, approached from opposite sides, and that runs which do not grok stay on one side. These are interesting, but they concern a derived statistic of gradients rather than an order parameter of the model's behaviour, they are unreviewed, and they concern algorithmic tasks.

Two observations follow. First, every case above in which T4 has been derived, or confirmed by a prediction that held, has a control parameter that is a property of the problem or the training recipe: data amount, task count, mixing ratio, number of properties, regularisation strength, model size or decoding temperature. Training time is the axis in only one of them: in Lubana et al. (2024) the transition is located in training iterations, but training is single-pass, so iterations equal data seen. The percolation theory's control variable is the amount of data, and the number of properties sets where the transition falls. A transition in time can be the finite-time trace of a transition in one of these parameters, as when a run passes through a critical amount of seen data, but the time axis itself has no thermodynamic limit, and calling it a control parameter, as the anchor does, borrows the vocabulary without the content. Second, the empirical cases that come closest to T4 earned their status through prediction: Lubana et al. (2024) and Gu et al. (2025) predicted how the location of the transition moves with a structural parameter, and the prediction held. Nothing comparable has been shown for any training-time transition in a real language model. Tigges et al. (2024) report that induction heads and other components appear at similar token counts across model sizes, which concerns the location of the event; no located study reports how its width in tokens depends on model size, which is what a finite-size scaling analysis would need.

9. Emergence over scale and over training

The anchor connects its result to emergence over scale, citing Wei et al. (2022) as the source of the phase-transition reading. The connection needs care, because the two literatures have reached different points.

Over scale, the metric critique has been partly confirmed and partly answered. Hu et al. (2024) introduced an evaluation with effectively unlimited resolution, by sampling until a correct answer appears, and found that small models make consistent progress on tasks where conventional metrics barely register them. With it they predicted the code-generation performance of a 2.4-billion-parameter model to within 0.05 percent before training. But they also found tasks whose curves, though smooth, accelerate in a way the standard scaling form cannot fit, which they call "accelerated emergence"; they showed that composing reasoning steps produces the opposite curvature, and proposed that competition among several circuits produces it, while describing their own validation of that hypothesis as "superficial". Du et al. (2024) trained models of 1.5, 6 and 32 billion parameters and found that, on MMLU, C-Eval, GSM8K and a Chinese translation of GSM8K, performance stayed at chance until the pre-training loss fell to about 2.2 and then improved; the threshold held under continuous metrics, and the curves for the three sizes overlapped when plotted against loss. Across intermediate checkpoints of OPT models from 125 million to 175 billion parameters, Xia et al. (2023) found that a checkpoint's perplexity forecast its few-shot accuracy on 74 multiple-choice BIG-Bench tasks whatever the model's size. Snell et al. (2024) showed that fine-tuning shifts the point of emergence toward smaller models and used the shift to predict, in some cases, emergence in models trained with up to four times more compute. Caballero et al. (2023) proposed a broken power law that fits delayed, sharp inflections such as those in arithmetic. Lu et al. (2024) argued from more than 1,000 experiments that many purported emergent abilities reduce to in-context learning, model memory and linguistic knowledge.

Two lessons carry over to training time. The first is that sharpness surviving a change to a continuous metric is real but is not yet a phase transition. Du et al. (2024) show a threshold in loss that survives continuous metrics, and Hu et al. (2024) show acceleration that survives unlimited resolution; both are evidence against pure T1, and neither is evidence for T4, since a kink in performance as a function of loss is fully compatible with a smooth crossover. The second is that the right coordinate matters. If downstream behaviour is better predicted by loss than by size or compute, then a capability's arrival in training time is the time at which the loss reaches the relevant level, and a curve against loss is the cleaner object of study. Muckatira et al. (2024) make a related point from the other side: 36 models of 1 to 165 million parameters pretrained on simplified language reached zero-shot performance, on tasks posed in that simplified language, comparable to models six times larger pretrained on unrestricted text, so the scale at which an ability appears depends on the data as much as on the model.

The complex-systems view adds a definitional check. Krakauer et al. (2025) argue that a necessary, though not sufficient, mark of emergence is that the system's long-run behaviour admits a new, compressed description in

coarse-grained variables. A sharp curve is neither necessary nor sufficient for that. A survey of emergent abilities, last updated in August 2026, catalogues definitions, the metric debate, loss thresholds and prediction methods across scale, compression and reasoning models (Berti et al., 2025). It does not set out what evidence would distinguish a training-time phase transition from a fast crossover or a threshold readout, which is the gap this paper addresses.

10. The measured values

Table 1 collects the measured values on which the argument rests, one row per cited result, with the sense of Section 3 that each bears on most directly. The column "Axis" names the variable along which the reported change occurs.

Three patterns stand out. The rows that come closest to T4 all have a control parameter other than training time, or, in the one case located in iterations (Lubana et al., 2024), a single pass in which iterations stand for data seen, and they come from solvable models or tightly controlled synthetic data. The rows that establish T3 come from studies that found a mechanism and, in the best cases, intervened on it; they include models as small as the anchor's and transitions located on linear step axes. The rows that bear on T1 show that the most common threshold metrics account for almost all reported emergence over scale and that aggregation can manufacture or conceal a sharp turn. The anchor's rows sit with T1 on the evidence it reports, and could move to T3 with the controls of Section 11.

Figure 1(b) makes a further point about loss. In the stage analysis at the anchor's scale, the first induction-circuit stage (LM3) changed the loss least of all five stages (-0.18), and the two induction stages together (-0.58) changed it less than either of the first two stages (-2.33, -1.22). A transition that matters to the model's computation need not show in the loss, which is why "invisible in the loss" cannot discriminate between the senses.

11. A procedure for classifying a reported transition

Separating the senses of Section 3 needs no new instrument: each control below has been used by at least one published study, and most are inexpensive. Algorithm 1 states them as a procedure. Its input is a training run, its checkpoints and a statistic reported to change sharply; its output is the strongest sense the evidence supports.

Every step is there because some published analysis needed it. Step 1 exists because seed-level timing can vary widely (Baherwani et al., 2026), because averaging over seeds can blur or misplace a step, and because single-checkpoint outliers have been mistaken for transitions (Arnold et al., 2024). Step 2 is the test implied by Schaeffer et al. (2023), with the addition that the reported statistic's behaviour under a smooth null be derived, as Section 4 does for dispersion and divergence; a statistic whose turn follows from its construction cannot confirm anything. Step 3 adapts the decomposition of Kangaslahti et al. (2025) and the per-subpopulation analysis of Wu and Lo (2025). Step 4 reports width, which no T1 or T3 claim needs but every T4 claim does, and adopts the alternative axes of Chen et al. (2024). Step 5 is the apparatus of Chen et al. (2024), Singh et al. (2024) and Olsson et al. (2022). Step 6 is the finite-size scaling the anchor itself proposes as future work, with the condition that the control parameter be a property of the problem rather than the clock.

Applied retrospectively, the anchor reaches Step 1 in part: it reports five seeds, but as means with deviation bands, and its checkpoint spacing within the band is not stated, and the text itself places the cusps both at 230 to 250 and at step 350. It does not reach Step 2: its statistics are functions of a binary vocabulary test, two of them are fixed by the share of correct words, and no linear score in the band is reported beyond the smooth loss. The remaining steps were not attempted. The procedure would classify the result as unresolved between T1 and a T3 instance of the bigram-to-word progression that Section 7 describes, which is less than the paper claims and more than nothing.

Clean verdicts are not guaranteed. A T3 stage can be revealed only by a T1 readout, and a T2 crossover in time can be the trace of a T4 transition in seen-data amount. Its value is that it records which controls were run, so that a reader can see what the word "transition" is carrying.

12. The strongest case against this paper

The strongest objection is that the distinctions drawn here are pedantic: practitioners use "phase transition" loosely to mean a sudden, qualitative change in behaviour, everyone knows it is a metaphor, and the anchor's observation of a sudden qualitative change in its model's output stands whatever it is called.

The reply is that the anchor does not use the term loosely. It names an order parameter and a control variable, appeals to a first-order mapping and barrier crossing, cites Landau and the renormalisation group, and concludes that the phenomenon is a general feature of training and scale-invariant in form. Those are T4 claims, and they invite T4 tests. More practically, the senses differ in what a reader should do next. If the event is T1, a better metric dissolves it and no one should look for a mechanism. If it is T3, there is a circuit to find, and the event is a target for interventions such as the suppression experiments of Chen et al. (2024). If it is T4, there are scaling laws to measure, and the location can be predicted from structural parameters as Lubana et al. (2024) and Gu et al. (2025) did. A paper that reports

Table 1. Published measurements bearing on what training-time and scale transitions establish. Values are as reported by the named source; "sense" is the author's classification.

Source	System	Axis	Reported change	Value	Sense
Hong and Hong 2025	3.6M char model, 5 seeds	epoch	band of all probes	230 to 250 of 600	T1 or T3
Hong and Hong 2025	same	epoch	mean correct-word length	~1.5 to ~2.5 chars	T1
Schaeffer et al. 2023	BIG-Bench annotations	scale	emergence under 2 metrics	>92% of cases	T1
Schaeffer et al. 2023	BIG-Bench	scale	metrics showing emergence	at most 5 of 39	T1
Wu and Lo 2025	56 LLMs, MMLU	compute	difficulty groups	10 groups, 14,042 items	T1
Hu et al. 2024	PassUntil, 2.4B	scale	prediction error, code	0.05%	T1
Du et al. 2024	1.5B, 6B, 32B	loss	MMLU leaves chance	loss ~2.2	T2
Xia et al. 2023	OPT 125M to 175B	perplexity	ICL tasks predicted	74 tasks	T2
Muckatira et al. 2024	36 models, 1M to 165M	scale	zero-shot parity	~6x larger models	T2
Kangaslahti et al. 2025	3-layer 40M, Wikipedia	step	basis vectors with labels	22 of 26	T3
Nanda et al. 2023	1-layer, mod 113	epoch	memorise, form, clean up	0-1.4k, -9.4k, -14k	T2/T3
Barak et al. 2022	sparse parity nets	iteration	jump near limit	~nˆO(k)	T2
Gopalani and Hu 2025	shallow transformers	step	hidden-state cosine	often >0.9	T2
Hoogland et al. 2024	2-layer 3M LM	step	stage ends	0.9k to 50k (5)	T3
Hoogland et al. 2024	same	step	loss change, LM3 and LM4	-0.18, -0.40	T3
Hoogland et al. 2024	same	step	bigram score gap, LM1	0.3 nats	T3
Hoogland et al. 2024	50k ICL regression	step	stage ends	1k to 500k (5)	T3
Chen et al. 2024	BERT-base, 3 seeds	step	loss drop window	steps 20K to 30K	T3
Olsson et al. 2022, via Tigges et al. 2024	models of all sizes	tokens	induction heads appear	~2B to 5B tokens	T3
Belrose et al. 2024	Pythia suite	step	n-gram loss trough	2^6 to 2^8 steps	T3
Baherwani et al. 2026	Pythia, 10 seeds per size	step	seeds solving repetition, 160M and 410M	10 of 10	T3
Arnold et al. 2024	Pythia checkpoints	step	outlier peaks	~20K, 40K, 80K	T1
Raventos et al. 2023	ICL regression, D=8	task count	diversity threshold	2^{14} to 2^{15}	T4 candidate
Lavie et al. 2026	1-layer softmax copy	samples	first-order point	~300 samples	T4
Lubana et al. 2024	formal language	data seen (= iterations, online)	onset scaling	sqrt of properties	T4 (location predicted)
Gu et al. 2025	biographies plus web	mix ratio	critical ratio vs size	power law	T4 (location predicted)
Cui et al. 2024	dot-product attention	sample ratio	positional to semantic	threshold alpha_c	T4
Wang 2026a	grokking, 8 scales	epoch	gradient-cascade dimension	crosses 1	T4?
Snell et al. 2024	fine-tuned LLMs	scale	emergence predicted	up to 4x compute	T2
Hu et al. 2024	PassUntil	scale	accelerated emergence	concave curve	T2

a sudden change without saying which sense it has earned leaves the reader unable to choose.

A second objection is that Section 4 rests on a reading of the anchor's methods. If its windows were defined by characters or by time rather than by a fixed number of words, the counts of correct and incorrect words would not be complementary and the identity would not hold. That is correct, and the paper states the dependence. The anchor's text says the generated text was divided "into non-overlapping windows of W = 21 words" and that the count of interest is "the number of words N per window"; under any other reading, a binomial or similar bounded-count null still applies, and the Poisson reference remains misspecified for a count whose maximum is fixed by the window. A one-line report of the share of correct words at each checkpoint would settle the matter: if the flip lies where that share crosses one half, the arithmetic accounts for it.

A third objection is that the derivations here are stylised. Real characters are not independently correct, and words are not independent within a window. That is also correct. The dispersion identity does not need independence; it needs only that the two counts sum to a constant. The binomial values and the exact-match argument for word length do use independence, and they are offered as the null against which a claim of

Algorithm 1. Classifying a reported training-time transition.

```
Input: checkpoints C over training, seeds S, reported statistic f,
       candidate band B, model sizes N (if available).
Step 1, resolution and seeds:
   densify checkpoints in and around B; plot f per seed, not averaged.
   if the turn is one checkpoint wide and reverts: report outlier, stop.
Step 2, threshold readout (T1):
   write f as g(latent); compute a score linear in per-token success
   (per-token log-likelihood, expected matches) on held-out data.
   derive f's behaviour under the null that the latent rises smoothly.
   if the linear score is smooth in B and f's turn is predicted
     by the null: report T1, stop.
Step 3, decomposition:
   decompose the loss by example, position or subspace direction;
   if no subpopulation drops in B: report T1, stop.
Step 4, crossover (T2):
   fit a smooth sigmoid to the decomposed loss; report width in steps
   and in tokens; check width under alternative x-axes (weight norm,
   distance from initialisation, loss).
Step 5, mechanism (T3):
   locate a component whose formation coincides with B; ablate or clamp
   it during training; require the drop to move or vanish and to
   recur across seeds in S. If so: report T3.
Step 6, singularity (T4):
   identify a control parameter other than time (data, tasks, mix, size);
   sweep N; require width, in absolute units, to shrink with N under a
   scaling collapse, or a derivation in a limit of size or of a
   structural parameter. If so: report T4.
Output: the highest step passed, with its measured width and seeds.
```

reorganisation should be tested, not as a model of the anchor's text.

A fourth objection is that the developmental literature cited in Section 7 makes the anchor's interpretation more plausible, not less: if small models pass through stages from bigrams to n-grams to induction, a character model passing from fragments to words is probably undergoing a real stage. That is fair, and it is why this paper classifies the anchor as unresolved between T1 and T3 rather than as T1. The point is that its evidence does not discriminate, and that the discriminating controls exist and have been applied by others.

13. What is not known

It is not known whether the anchor's band reflects a change in the model or in the probes. No linear score in per-character success, and no decomposition of the loss, has been reported for that band.

It is not known how the anchor's statistics depend on decoding. The methods state both sampling at temperature 1.0 and greedy decoding, and greedy decoding of weak models produces repetitive text that would drive the pre-transition statistics.

It is not known whether any training-time transition in a real language model sharpens with model size. Locations have been compared across sizes; widths have not.

It is not known what the right control parameter is for a training-time transition. Seen-data amount, loss level and weight norm have each been used as alternative axes, but no study has shown that a training-time transition collapses onto a single curve under one of them across sizes.

It is not known whether the accelerated emergence identified over scale has a training-time counterpart, or whether its proposed cause, competition among circuits, is correct; its proposers describe their validation as superficial.

It is not known how often developmental stages are transient. In-context learning has been shown to emerge and then vanish while loss falls, and no survey reports how many documented stages persist to the end of training.

It is not known whether finite-size scaling of gradient statistics, the only located application of that method to training, measures anything about the model's behaviour. The two studies are unreviewed and confined to algorithmic tasks.

14. Conclusion

The anchor shows that a small character-level transformer's generated words change character in a narrow band of training: errors stop being repeated fragments, correct words get longer, and the two dispersion indices of word counts exchange which one sits near the Poisson value. That much is established, and it is consistent with a real developmental stage. What the result does not show is a phase transition in any sense stronger than that the curves turn. Its four probes are functions of one share of correct words, two of them by arithmetic, and their agreement is what a single smoothly rising share would produce. The smoothness of the loss, which the anchor reads as concealment, is equally what a threshold readout predicts.

The wider record says that the anchor's question has better answers than its own. Developmental stages in

small transformers are real and have been located, timed and manipulated, sometimes at the same scale and on linear axes. Transitions in the strict sense have been derived and measured, but always along an axis that belongs to the problem, such as data, tasks or mixture, and always with a theory that predicted where the transition would move. Between these lies a large class of fast crossovers that look like transitions on a training curve and are not. The word "phase transition" has been asked to cover all three, and the classification in Algorithm 1, with a derived null for each reported statistic, a linear score, a decomposition, a mechanism and a size sweep, is the measurement that the next claim in this line should carry.

References

Achille, A., Rovere, M., & Soatto, S. (2019). Critical Learning Periods in Deep Neural Networks. arXiv:1711.08856. doi:10.48550/arXiv.1711.08856. [PREPRINT]

Arnold, J., Holtorf, F., Schafer, F., & Lorch, N. (2024). Phase Transitions in the Output Distribution of Large Language Models. arXiv:2405.17088. doi:10.48550/arXiv.2405.17088. [PREPRINT]

Baherwani, V., Chen, Z., Qiu, S., Wilson, A. G., & Izmailov, P. (2026). Emergent Capabilities Arise Randomly from Learning Sparse Attention Patterns. arXiv:2606.25010. doi:10.48550/arXiv.2606.25010. [PREPRINT]

Barak, B., Edelman, B. L., Goel, S., Kakade, S., Malach, E., & Zhang, C. (2022). Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit. arXiv:2207.08799. doi:10.48550/arXiv.2207.08799. [PREPRINT]

Belrose, N., Pope, Q., Quirke, L., Mallen, A., & Fern, X. (2024). Neural Networks Learn Statistics of Increasing Complexity. arXiv:2402.04362. doi:10.48550/arXiv.2402.04362. [PREPRINT]

Berti, L., Giorgi, F., & Kasneci, G. (2025). Emergent Abilities in Large Language Models: A Survey. arXiv:2503.05788. doi:10.48550/arXiv.2503.05788. [PREPRINT]

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., et al. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. arXiv:2304.01373. doi:10.48550/arXiv.2304.01373. [PREPRINT]

Caballero, E., Gupta, K., Rish, I., & Krueger, D. (2023). Broken Neural Scaling Laws. arXiv:2210.14891. doi:10.48550/arXiv.2210.14891. [PREPRINT]

Carroll, L., Hoogland, J., Farrugia-Roberts, M., & Murfet, D. (2025). Dynamics of Transient Structure in In-Context Linear Regression Transformers. arXiv:2501.17745. doi:10.48550/arXiv.2501.17745. [PREPRINT]

Chan, S. C. Y., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., McClelland, J., & Hill, F. (2022). Data Distributional Properties Drive Emergent In-Context Learning in Transformers. arXiv:2205.05055. doi:10.48550/arXiv.2205.05055. [PREPRINT]

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. (2024). Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs. arXiv:2309.07311. doi:10.48550/arXiv.2309.07311. [PREPRINT]

Choshen, L., Hacohen, G., Weinshall, D., & Abend, O. (2022). The Grammar-Learning Trajectories of Neural Language Models. arXiv:2109.06096. doi:10.48550/arXiv.2109.06096. [PREPRINT]

Cui, H., Behrens, F., Krzakala, F., & Zdeborova, L. (2024). A phase transition between positional and semantic learning in a solvable model of dot-product attention. arXiv:2402.03902. doi:10.48550/arXiv.2402.03902. [PREPRINT]

Du, Z., Zeng, A., Dong, Y., & Tang, J. (2024). Understanding Emergent Abilities of Language Models from the Loss Perspective. arXiv:2403.15796. doi:10.48550/arXiv.2403.15796. [PREPRINT]

Edelman, B. L., Edelman, E., Goel, S., Malach, E., & Tsilivis, N. (2024). The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains. arXiv:2402.11004. doi:10.48550/arXiv.2402.11004. [PREPRINT]

Ganguli, D., Hernandez, D., Lovitt, L., DasSarma, N., Henighan, T., Jones, A., et al. (2022). Predictability and Surprise in Large Generative Models. arXiv:2202.07785. doi:10.48550/arXiv.2202.07785. [PREPRINT]

Gopalani, P., Lubana, E. S., & Hu, W. (2024). Abrupt Learning in Transformers: A Case Study on Matrix Completion. arXiv:2410.22244. doi:10.48550/arXiv.2410.22244. [PREPRINT]

Gopalani, P., & Hu, W. (2025). What Happens During the Loss Plateau? Understanding Abrupt Learning in Transformers. arXiv:2506.13688. doi:10.48550/arXiv.2506.13688. [PREPRINT]

Gu, X., Lyu, K., Li, J., & Zhang, J. (2025). Data Mixing Can Induce Phase Transitions in Knowledge Acquisition. arXiv:2505.18091. doi:10.48550/arXiv.2505.18091. [PREPRINT]

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The Curious Case of Neural Text Degeneration. arXiv:1904.09751. doi:10.48550/arXiv.1904.09751. [PREPRINT]

Hong, N., & Hong, T. (2025). Evidence of Phase Transitions in Small Transformer-Based Language Models. arXiv:2511.12768. doi:10.48550/arXiv.2511.12768. [PREPRINT]

Hoogland, J., Wang, G., Farrugia-Roberts, M., Carroll, L., Wei, S., & Murfet, D. (2024). Loss Landscape Degeneracy and Stagewise Development in Transformers. arXiv:2402.02364. doi:10.48550/arXiv.2402.02364. [PREPRINT]

Hu, S., Liu, X., Han, X., Zhang, X., He, C., Zhao, W., et al. (2024). Predicting Emergent Abilities with Infinite Resolution Evaluation. arXiv:2310.03262. doi:10.48550/arXiv.2310.03262. [PREPRINT]

Huang, Y., Hu, S., Han, X., Liu, Z., & Sun, M. (2024). Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition. arXiv:2402.15175. doi:10.48550/arXiv.2402.15175. [PREPRINT]

Kangaslahti, S., Rosenfeld, E., & Saphra, N. (2025). Hidden Breakthroughs in Language Model Training. arXiv:2506.15872. doi:10.48550/arXiv.2506.15872. [PREPRINT]

Kirsch, L., Harrison, J., Sohl-Dickstein, J., & Metz, L. (2022). General-Purpose In-Context Learning by Meta-Learning Transformers. arXiv:2212.04458. doi:10.48550/arXiv.2212.04458. [PREPRINT]

Krakauer, D. C., Krakauer, J. W., & Mitchell, M. (2025). Large Language Models and Emergence: A Complex Systems Perspective. arXiv:2506.11135. doi:10.48550/arXiv.2506.11135. [PREPRINT]

Kumar, T., Bordelon, B., Gershman, S. J., & Pehlevan, C. (2024). Grokking as the Transition from Lazy to Rich Training Dynamics. arXiv:2310.06110. doi:10.48550/arXiv.2310.06110. [PREPRINT]

Lau, E., Furman, Z., Wang, G., Murfet, D., & Wei, S. (2024). The Local Learning Coefficient: A Singularity-Aware Complexity Measure. arXiv:2308.12108. doi:10.48550/arXiv.2308.12108. [PREPRINT]

Lavie, I., Fischer, K., Lekov, A., Van Maele, F., Ringel, Z., & Helias, M. (2026). Phase Transitions in Attention: A Bayesian Theory of Copy Head Emergence. arXiv:2606.12058. doi:10.48550/arXiv.2606.12058. [PREPRINT]

Levi, N., Beck, A., & Bar-Sinai, Y. (2024). Grokking in Linear Estimators - A Solvable Model that Groks without Understanding. arXiv:2310.16441. doi:10.48550/arXiv.2310.16441. [PREPRINT]

Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., & Williams, M. (2022). Towards Understanding Grokking: An Effective Theory of Representation Learning. arXiv:2205.10343. doi:10.48550/arXiv.2205.10343. [PREPRINT]

Liu, Z., Michaud, E. J., & Tegmark, M. (2023). Omnigrok: Grokking Beyond Algorithmic Data. arXiv:2210.01117. doi:10.48550/arXiv.2210.01117. [PREPRINT]

Lu, S., Bigoulaeva, I., Sachdeva, R., Tayyar Madabushi, H., & Gurevych, I. (2024). Are Emergent Abilities in Large Language Models just In-Context Learning? arXiv:2309.01809. doi:10.48550/arXiv.2309.01809. [PREPRINT]

Lubana, E. S., Kawaguchi, K., Dick, R. P., & Tanaka, H. (2024). A Percolation Model of Emergence: Analyzing Transformers Trained on a Formal Language. arXiv:2408.12578. doi:10.48550/arXiv.2408.12578. [PREPRINT]

Michaud, E. J., Liu, Z., Girit, U., & Tegmark, M. (2023). The Quantization Model of Neural Scaling. arXiv:2303.13506. doi:10.48550/arXiv.2303.13506. [PREPRINT]

Muckatira, S., Deshpande, V., Lialin, V., & Rumshisky, A. (2024). Emergent Abilities in Reduced-Scale Generative Language Models. arXiv:2404.02204. doi:10.48550/arXiv.2404.02204. [PREPRINT]

Nakaishi, K., Nishikawa, Y., & Hukushima, K. (2024). Phase transition in large language models and the criticality of natural languages. arXiv:2406.05335. doi:10.48550/arXiv.2406.05335. [PREPRINT]

Nam, Y., Fonseca, N., Lee, S. H., Mingard, C., & Louis, A. A. (2024). An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem. arXiv:2404.17563. doi:10.48550/arXiv.2404.17563. [PREPRINT]

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. arXiv:2301.05217. doi:10.48550/arXiv.2301.05217. [PREPRINT]

Nguyen, T. (2024). Understanding Transformers via N-gram Statistics. arXiv:2407.12034. doi:10.48550/arXiv.2407.12034. [PREPRINT]

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In-context Learning and Induction Heads. arXiv:2209.11895. doi:10.48550/arXiv.2209.11895. [PREPRINT]

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177. doi:10.48550/arXiv.2201.02177. [PREPRINT]

Raventos, A., Paul, M., Chen, F., & Ganguli, S. (2023). Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. arXiv:2306.15063. doi:10.48550/arXiv.2306.15063. [PREPRINT]

Reddy, G. (2023). The mechanistic basis of data dependence and abrupt learning in an in-context classification task. arXiv:2312.03002. doi:10.48550/arXiv.2312.03002. [PREPRINT]

Rende, R., Gerace, F., Laio, A., & Goldt, S. (2024). A distributional simplicity bias in the learning dynamics of transformers. arXiv:2410.19637. doi:10.48550/arXiv.2410.19637. [PREPRINT]

Rubin, N., Seroussi, I., & Ringel, Z. (2024). Grokking as a First Order Phase Transition in Two Layer Networks. arXiv:2310.03789. doi:10.48550/arXiv.2310.03789. [PREPRINT]

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120. doi:10.48550/arXiv.1312.6120. [PREPRINT]

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. arXiv:1810.10531. doi:10.48550/arXiv.1810.10531. [PREPRINT]

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004. doi:10.48550/arXiv.2304.15004. [PREPRINT]

Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S., & Koyejo, S. (2024). Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive? arXiv:2406.04391. doi:10.48550/arXiv.2406.04391. [PREPRINT]

Singh, A. K., Chan, S. C. Y., Moskovitz, T., Grant, E., Saxe, A. M., & Hill, F. (2023). The Transient Nature of Emergent In-Context Learning in Transformers. arXiv:2311.08360. doi:10.48550/arXiv.2311.08360. [PREPRINT]

Singh, A. K., Moskovitz, T., Hill, F., Chan, S. C. Y., & Saxe, A. M. (2024). What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. arXiv:2404.07129. doi:10.48550/arXiv.2404.07129. [PREPRINT]

Snell, C., Wallace, E., Klein, D., & Levine, S. (2024). Predicting Emergent Capabilities by Finetuning. arXiv:2411.16035. doi:10.48550/arXiv.2411.16035. [PREPRINT]

Tigges, C., Hanna, M., Yu, Q., & Biderman, S. (2024). LLM Circuit Analyses Are Consistent Across Training and Scale. arXiv:2407.10827. doi:10.48550/arXiv.2407.10827. [PREPRINT]

Varma, V., Shah, R., Kenton, Z., Kramar, J., & Kumar, R. (2023). Explaining grokking through circuit efficiency. arXiv:2309.02390. doi:10.48550/arXiv.2309.02390. [PREPRINT]

Wang, G., Hoogland, J., van Wingerden, S., Furman, Z., & Murfet, D. (2024). Differentiation and Specialization of Attention Heads via the Refined Local Learning Coefficient. arXiv:2410.02984. doi:10.48550/arXiv.2410.02984. [PREPRINT]

Wang, P. (2026a). Grokking as Dimensional Phase Transition in Neural Networks. arXiv:2604.04655. doi:10.48550/arXiv.2604.04655. [PREPRINT]

Wang, P. (2026b). Dimensional Criticality at Grokking Across MLPs and Transformers. arXiv:2604.16431. doi:10.48550/arXiv.2604.16431. [PREPRINT]

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent Abilities of Large Language Models. arXiv:2206.07682. doi:10.48550/arXiv.2206.07682. [PREPRINT]

Wu, T.-Y., & Lo, P.-Y. (2025). U-shaped and Inverted-U Scaling behind Emergent Abilities of Large Language Models. arXiv:2410.01692. doi:10.48550/arXiv.2410.01692. [PREPRINT]

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., & Stoyanov, V. (2023). Training Trajectories of Language Models Across Scales. arXiv:2212.09803. doi:10.48550/arXiv.2212.09803. [PREPRINT]

Ziyin, L., & Ueda, M. (2022). Exact Phase Transitions in Deep Learning. arXiv:2205.12510. doi:10.48550/arXiv.2205.12510. [PREPRINT]

Zunkovic, B., & Ilievski, E. (2022). Grokking phase transitions in learning local rules with gradient descent. arXiv:2210.15435. doi:10.48550/arXiv.2210.15435. [PREPRINT]

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24355741