Training Once Is Enough: Activation Fingerprint Convergence Reveals Ensemble Collapse in Language Model Training

Jeremy McEntire

Abstract

We demonstrate that neural language models trained from independent random seeds converge to nearly identical activation fingerprints during training, with phase boundaries synchronized to within ±50 gradient steps at most scales. Using 100 fixed probe inputs to capture activation snapshots at 40 checkpoints across GPT-2 124M (5 seeds), Qwen 2.5-1.5B [Team, 2025] (4 seeds), and Qwen 2.5-7B (5 seeds), we find cross-seed cosine similarity exceeding 0.999 in stable regimes at all scales. Final validation loss coefficient of variation is 0.41% at 124M, 2.43% at 1.5B, and 1.53% at 7B—all well below typical stochastic training noise. The convergence exhibits three phases: (1) representational turbulence during the chaotic regime, where similarity *decreases* with model scale (0.994 at 124M vs. 0.976 at 7B), (2) rapid alignment during transition, and (3) effective fingerprint collapse in stable regime, where similarity *increases* with scale (0.9997 at 124M, 0.9995 at 1.5B, 0.9999 at 7B). Regime boundaries are synchronized across seeds to within ±50 steps at most scales, with cross-seed regime agreement reaching 97% at 1.5B. These findings imply that multi-seed training—a standard practice costing 3−5× compute—provides diminishing returns as models scale. The loss landscape contains a dominant solution manifold that all initializations converge toward, and the convergence tightens with model size.

1 Introduction

Training neural networks from multiple random seeds is standard practice. Researchers report mean ± standard deviation across 3–5 seeds to demonstrate robustness; practitioners train ensembles for uncertainty estimation. At scale, this is enormously expensive: training GPT-3 once cost an estimated $4.6M in compute; five seeds would cost $23M.

But do different seeds actually produce meaningfully different models? Prior work on mode connectivity [Garipov et al., 2018, Draxler et al., 2018] has shown that independently trained networks can be connected by paths of low loss in weight space. Ainsworth et al. [2023] demonstrated that after permutation alignment, models from different seeds are often linearly mode connected. These results operate in *weight space*—they show that different parameterizations are functionally interchangeable.

We ask a stronger question: do models from different seeds converge to the same *functional behavior*, as measured directly in activation space? Using activation fingerprinting—a technique we introduced for regime detection in speculative weight prediction [McEntire, 2026b]—we track how models' internal representations evolve during training across independent seeds.

Our answer is striking: yes, they converge, and the convergence tightens with scale. At 7B parameters, cross-seed cosine similarity reaches 0.9999 in the stable training regime, and final validation losses agree to within 1.53% coefficient of variation. Phase boundaries—the steps at which models transition between chaotic, transition, and stable training regimes—are synchronized across seeds to within ±50 gradient steps at most scales (wider at 7B transition onset, ±108 steps), despite entirely independent random initializations.

This **ensemble collapse** phenomenon has practical implications. If models trained from different seeds converge to the same functional behavior, multi-seed training provides diminishing returns with scale. A single training run, combined with post-hoc confidence estimation (e.g., using the regime detector to assess trajectory stability), may be sufficient. At 7B scale, this would save 60–80% of ensemble training compute.

Contributions.

1. **Activation-space convergence**: We demonstrate that independently trained models converge to near-identical activation fingerprints, with convergence tightening with scale.
2. **Phase boundary synchronization**: Training regime transitions are deterministic properties of the optimization landscape, not artifacts of initialization.
3. **Three-scale evidence**: Consistent findings across GPT-2 124M, Qwen 1.5B, and Qwen 7B on WikiText-103.
4. **Practical recommendation**: Multi-seed training is redundant at scale; single-seed training with regime-based confidence estimation is sufficient.

2 Related Work

Mode connectivity. Garipov et al. [2018] and Draxler et al. [2018] independently demonstrated that the loss landscape contains smooth, low-loss paths connecting independently trained networks. Ainsworth et al. [2023] showed that after permutation alignment, networks are often linearly mode connected. These results establish functional equivalence in weight space; we demonstrate convergence directly in activation (representation) space.

Lottery tickets and subnetwork convergence. Frankle and Carlin [2019] showed that sparse subnetworks ("winning tickets") can match full-network performance. Frankle et al. [2020] demonstrated that networks become "stable to SGD noise" early in training, after which models sharing a common prefix converge to linearly connected solutions. Our work extends this: even without a shared prefix, models converge to the same activation fingerprint.

Training dynamics. Cohen et al. [2021] identified progressive sharpening and edge-of-stability phases. Lewkowycz et al. [2020] documented the catapult phase. Nanda et al. [2023] identified three mechanistic phases in grokking. We use activation-space cosine similarity as a computationally cheap proxy for these dynamics, following McEntire [2026b].

Neural scaling laws. Kaplan et al. [2020] established power-law relationships between model size, data, and loss. Hoffmann et al. [2022] refined compute-optimal scaling. Our finding that cross-seed variance decreases with scale is consistent with the broader observation that larger models exhibit more predictable training dynamics.

3 Method: Activation Fingerprinting

We adopt the activation fingerprinting framework from McEntire [2026b]. At each training checkpoint (every 50 steps), we compute an activation fingerprint by forward-passing 100 fixed probe sentences through the model and concatenating the mean-pooled final hidden states. The probe set spans arithmetic, linguistic, factual, and random inputs.

3.1 Regime Classification

The cosine similarity between consecutive fingerprints serves as a proxy for the local Lyapunov exponent:

$$s_t = \frac{\mathbf{a}_t \cdot \mathbf{a}_{t-\Delta}}{\|\mathbf{a}_t\| \|\mathbf{a}_{t-\Delta}\|} \tag{1}$$

where $\Delta = 50$ steps. We classify training into three regimes using thresholds calibrated on initial runs:

$$\text{regime}(t) = \begin{cases} \text{stable} & \text{if } s_t > \tau_{\text{high}} \\ \text{chaotic} & \text{if } s_t < \tau_{\text{low}} \\ \text{transition} & \text{otherwise} \end{cases} \tag{2}$$

3.2 Cross-Seed Analysis

For ensemble collapse, the key measurements are:

- **Cross-seed loss CV**: Coefficient of variation of final validation loss across seeds.
- **Phase boundary synchronization**: Standard deviation of the training step at which each seed enters each regime.
- **Per-regime cosine similarity**: Mean similarity within each regime, computed separately per seed.
- **Cross-seed regime agreement**: Fraction of checkpoints where all seeds are classified into the same regime.

4 Experimental Setup

We evaluate three language models spanning two orders of magnitude:

- **GPT-2 124M**: 12 layers, 768 hidden, 12 heads. 5 seeds (42–46).
- **Qwen 2.5-1.5B**: 28 layers, 1536 hidden, 12 heads. 4 seeds (42–45).
- **Qwen 2.5-7B**: 28 layers, 3584 hidden, 28 heads. 5 seeds (42–46).

All models are randomly initialized and trained from scratch on WikiText-103 [Merity et al., 2017] for 2000 steps with AdamW ($\eta = 5 \times 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01), cosine learning rate schedule, 100-step warmup, batch size 4, sequence length 256. Checkpoints are saved every 50 steps (40 per seed), yielding 560 total fingerprint observations. GPT-2 and Qwen 1.5B were trained on A100 40GB; Qwen 7B on A100 80GB. All code is available at `https://github.com/jmcentire/leap-verify`.

5 Results

5.1 Cross-Seed Loss Convergence

Table 1 summarizes final validation losses across seeds.

Table 1: Final validation loss across seeds. CV = coefficient of variation.

Model	Seeds	Mean	Std	Range	CV (%)
GPT-2 124M	5	1.616	0.007	1.609–1.625	**0.41**
Qwen 2.5-1.5B	4	0.814	0.020	0.801–0.843	2.43
Qwen 2.5-7B	5	0.985	0.015	0.970–1.002	1.53

All three scales exhibit extremely low cross-seed variance. GPT-2 124M achieves the tightest convergence (CV = 0.41%), likely because its small parameter count constrains the solution space. Qwen 1.5B shows the highest CV (2.43%), but this is still far below the level at which ensemble diversity would provide meaningful benefit. The 7B model falls between (1.53%), consistent with its non-monotonic position in the regime distribution (Section 5.2).

5.2 Phase Boundary Synchronization

Table 2 shows when each seed enters each training regime.

The synchronization is remarkable. All five GPT-2 seeds enter transition at *exactly* step 100—zero variance. At 1.5B, transition onset varies by only ±29 steps across 4 seeds. At 7B, transition onset varies more (±108 steps, driven by seed 42 being late at step 1250) but stable entry is tight (±42 steps). Given that these models start from entirely independent random initializations, this synchronization implies that regime boundaries are deterministic properties of the optimization landscape—the learning rate schedule, data distribution, and architecture jointly determine *when* the model transitions, regardless of initialization.

Table 2: Phase boundary synchronization: training step at which each seed first enters each regime. Std computed across seeds.

Model	Transition	Mean	Std	Range
GPT-2 124M	First transition	100	0	[100, 100]
	First stable	160	22	[150, 200]
Qwen 2.5-1.5B	First transition	1375	29	[1350, 1400]
	First stable	1775	35	[1750, 1800]
Qwen 2.5-7B	First transition	1060	108	[1000, 1250]
	First stable	1340	42	[1300, 1400]

5.3 Regime Distribution Across Scales

Table 3: Fraction of training spent in each regime (mean across seeds). Percentages sum to <100% because the first checkpoint has no predecessor for similarity computation. The regime distribution shifts non-monotonically with scale.

Model	Chaotic	Transition	Stable	Regime Agreement
GPT-2 124M	4.0%	60.0%	**33.5%**	74.5%
Qwen 2.5-1.5B	63.7%	31.9%	1.9%	**96.9%**
Qwen 2.5-7B	48.0%	16.5%	**33.0%**	93.0%

Three notable patterns emerge from Tables 3 and 4. First, the regime distribution is *non-monotonic* with scale: GPT-2 spends 33.5% in stable, Qwen 1.5B barely reaches stable (1.9%), and Qwen 7B recovers to 33.0% stable—matching the 124M model. Second, cross-seed regime agreement is *lowest* at 124M (74.5%), because GPT-2 oscillates between transition and stable rather than cleanly separating regimes. The larger models show cleaner, more deterministic regime progressions despite spending more time in chaotic. Third, the 7B model exhibits the clearest three-phase structure: a sustained chaotic phase (48%), brief transition (17%), then sustained stable (33%).

5.4 Activation Similarity by Regime

Table 4: Mean cosine similarity between consecutive activation fingerprints, by regime. Chaotic similarity *decreases* with scale; stable similarity *increases*.

Model	Chaotic	Transition	Stable
GPT-2 124M	0.9942	0.9987	0.9997
Qwen 2.5-1.5B	0.9851	0.9988	0.9995
Qwen 2.5-7B	0.9764	0.9982	0.9999

The regime classification is well-calibrated across scales. The signature finding is the *opposing trends* in chaotic vs. stable similarity:

- **Chaotic**: Similarity *decreases* with scale ($0.994 \rightarrow 0.985 \rightarrow 0.976$). Larger models undergo more violent representational changes during early training.
- **Transition**: Remarkably stable across scales ($\sim$0.998).
- **Stable**: Similarity *increases* with scale ($0.9997 \rightarrow 0.9995 \rightarrow 0.9999$). Larger models achieve tighter representational convergence.

This pattern is consistent with a funnel-shaped loss landscape: larger models explore more broadly during chaotic training but converge more tightly to a narrow solution manifold.

5.5 Ensemble Collapse in Val Loss Trajectory (7B)

To visualize ensemble collapse directly, we examine the evolution of cross-seed validation loss CV during 7B training. At step 50, CV is 0.73%—the baseline diversity from random initialization. At step 250, CV spikes to 35.8% as one seed (seed 44) hits a catastrophic early loss spike (val loss 2.01 vs. mean ∼1.0). From step 500 onward, CV monotonically decreases: 3.1% → 2.1% → 1.6% → **1.53%** at convergence.

This is the clearest quantitative signature of ensemble collapse: seeds that wildly disagree early in training progressively converge to the same loss basin. The catastrophic outlier (seed 44) is fully absorbed by step 500—the optimization landscape's attractor basin is strong enough to recover from 2× loss spikes.

6 Discussion

6.1 Practical Implication: Train Once

Our results suggest that multi-seed training—a standard practice costing 3−5× base compute—provides diminishing returns at scale. At 7B parameters, cross-seed loss CV is 1.53% and activation similarity exceeds 0.999 in the stable regime. The models are not merely producing similar losses; they are producing near-identical internal representations.

We recommend **single-seed training with post-hoc confidence estimation**: train once, use the regime detector to verify that the model has reached stable regime (similarity $> \tau_{\text{high}}$), and report the regime classification as a proxy for training stability. This provides the same information as multi-seed training—evidence that the model has converged to the dominant solution manifold—at $\frac{1}{5}$ the compute cost.

6.2 Functional vs. Parametric Convergence

Models trained from different seeds may have substantially different weights (weight-space divergence) yet produce nearly identical activations (function-space convergence). This is the definition of a *solution manifold*: a set of weight configurations that implement the same function. The manifold has a characteristic temporal structure—chaotic → transition → stable—that all initializations traverse in synchrony.

The activation fingerprint measures functional equivalence directly, without requiring the expensive permutation alignment needed for weight-space comparison [Ainsworth et al., 2023]. This makes it a practical tool for detecting convergence during training, not only in post-hoc analysis.

6.3 Scaling Prediction

The trends we observe across 124M, 1.5B, and 7B suggest a prediction for larger scales:

1. Cross-seed loss CV will continue to decrease (0.41% → 2.43% → 1.53% shows non-monotonic behavior at 1.5B, but the 7B recovery suggests the 1.5B anomaly is a regime-distribution artifact, not a trend reversal).

2. Stable-regime similarity will approach 1.0000 even more closely.

3. Phase boundary synchronization will remain tight (±50 steps).

In the limit, large models may have a *single effective solution* (up to weight permutation), making ensemble training strictly redundant.

6.4 Connection to the Activation Fingerprint Framework

This paper is part of a series using activation fingerprints as a unifying analytical tool:

- McEntire [2026b] (Paper 1): Uses fingerprint similarity for regime detection in speculative weight prediction.
- This paper: Uses cross-seed fingerprint convergence to demonstrate ensemble collapse.
- McEntire [2026a] (Paper 3): Uses fingerprints to index specialist models for query-driven parameter composition.

The same low-cost signal—100 probe sentences, one forward pass, cosine similarity—serves regime detection, convergence monitoring, and model indexing, demonstrating the generality of the activation fingerprint as a training diagnostic.

6.5 Limitations

Training was limited to 2000 steps on WikiText-103. The Qwen 1.5B model has only 4 seeds (one was lost to infrastructure failure). We measure within-fingerprint cosine similarity (consecutive checkpoints) rather than cross-seed fingerprint similarity (same step, different seeds)—the latter would require storing and comparing $O(n^2)$ fingerprint pairs. We do not measure weight-space divergence directly; functional convergence in activation space does not preclude parametric divergence, and the relationship between the two warrants further study.

7 Conclusion

We have demonstrated ensemble collapse in language model training: independently initialized models converge to near-identical activation fingerprints with synchronized phase boundaries. The convergence tightens with model scale, consistent with a dominant solution manifold that all initializations converge toward.

The practical implication is clear: as models scale, multi-seed training provides diminishing returns. At 7B parameters, a single training run—verified via regime detection—captures the same information as a 5-seed ensemble at $\frac{1}{5}$ the compute cost. For the growing number of organizations training models at scale, this represents a substantial cost saving.

The theoretical implication is equally important: the loss landscape of modern neural networks is simpler than it appears. Despite the astronomical dimensionality of the parameter space, training trajectories from different initializations converge to the same functional behavior, at the same pace, with phase transitions at the same training steps. The optimization landscape is not a wilderness—it is a funnel.

Acknowledgments

The author thanks Amos Waterland for creating the ASC architecture and for early discussions on manifold structure in computation that informed the solution-manifold interpretation. Brian Cremeans is acknowledged for discussions on dynamical systems analogies. Claude (Anthropic) provided extensive assistance with experimental implementation, GPU infrastructure management, data analysis, and manuscript preparation. This work used GPU compute from vast.ai on NVIDIA A100 40GB and 80GB GPUs.

Code availability. All code, experiment scripts, and paper source are available at `https://github.com/jmcentire/leap-verify` (DOI: 10.5281/zenodo.18739387).

References

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning (ICML)*, 2018.

Jonathan Frankle and Michael Carlin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Jeremy McEntire. Constellation-indexed model composition: Query-driven parameter mixing via activation fingerprints. *arXiv preprint*, 2026a.

Jeremy McEntire. Leap+verify: Regime-adaptive speculative weight prediction for accelerating neural network training. *arXiv preprint arXiv:2602.19580*, 2026b. URL `https://arxiv.org/abs/2602.19580`.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.