Sharpness Evolution and Its Relationship to Optimization and Performance at LLM Scale

Anonymous Author(s)

ABSTRACT

Understanding how loss landscape sharpness evolves during large-scale language model training is critical for explaining optimization dynamics and generalization behavior. We present a *hypothesis-generating* simulation study that models sharpness evolution across six model scales from 10M to 7B parameters, producing testable predictions about the relationship between sharpness, optimization, and downstream performance. Our parametric model encodes a three-phase sharpness evolution pattern—initial rise, exponential decay, and plateau stabilization—with scale-dependent parameters motivated by empirical observations. Under these assumptions, the simulator produces several emergent predictions: final critical sharpness follows a log-linear scaling law with model size ($S = -0.1055 \cdot \log_{10}(N) + 2.0196$), and this relationship extends to alternative measures including trace sharpness (slope = -0.0599) and spectral norm sharpness (slope = -0.0979), with pairwise cross-measure correlations exceeding 0.98. Ablation studies show that the three-phase pattern is preserved in 85.2% of parameter configurations (46/54). Learning rate schedule analysis predicts that cosine and linear schedules preserve the three-phase pattern across scales, while cosine restarts and constant schedules disrupt it at smaller scales. Extended scaling predictions to 70B parameters achieve maximum error of 1.15% within the simulator. Edge-of-stability dynamics analysis predicts that the damping rate governing the Phase 2-to-3 transition increases with scale (4.72 at 10M to 5.39 at 7B). A regression model using early-phase features predicts final sharpness with high accuracy from just 10% of training data. We emphasize that these are simulator-derived hypotheses: the strong quantitative relationships we report (e.g., high R^2 values from six scale points) reflect the regularity of the parametric model and require empirical validation on real training checkpoints. Nonetheless, the framework generates concrete, falsifiable predictions about sharpness dynamics at scale that can guide future empirical investigation.

1 INTRODUCTION

The geometry of the loss landscape in neural networks, particularly the sharpness of minima found during training, has long been hypothesized to influence generalization [8, 14]. Sharp minima, characterized by large eigenvalues of the Hessian, correspond to solutions that are sensitive to small perturbations in parameter space, while flat minima exhibit robustness and have been associated with better generalization [7, 16]. Recent theoretical work on universal sharpness dynamics [12] has provided a rigorous framework for understanding how sharpness evolves during training through fixed-point analysis, characterizing progressive sharpening, edge-of-stability behavior, and routes to chaos.

For Large Language Models (LLMs), understanding sharpness dynamics is especially important given the observed scaling laws governing their performance [9, 13]. However, direct measurement of Hessian sharpness becomes computationally impractical at LLM scales, as the cost scales quadratically with model dimensionality. This limitation has restricted most empirical studies to models with approximately 10M parameters, leaving fundamental questions about how sharpness behaves at realistic scales unresolved.

Recent work by Kalra et al. [11] addresses the measurement challenge by introducing critical sharpness as a scalable proxy requiring fewer than 10 forward passes given the update direction, providing empirical evidence at up to 7B parameters using OLMo-2 checkpoints. Chen et al. [3] further discover that LLMs develop expansive stability basins whose width increases with both scale and training progress, consistent with SGD's implicit bias toward flatter minima. However, the systematic characterization of sharpness evolution—its temporal dynamics during training and its quantitative relationship to optimization and downstream performance—remains an open question.

In this work, we address this gap through a *hypothesis-generating* simulation study that models sharpness evolution across six model scales spanning three orders of magnitude (10M to 7B parameters). Our simulation framework encodes key phenomena observed in empirical studies—the initial rise in sharpness during early training (the catapult mechanism [15]), edge-of-stability oscillations [4], and scale-dependent convergence to flat minima—as parametric assumptions, and explores what quantitative predictions emerge from these assumptions. It is important to distinguish the *assumptions encoded in the simulator* (e.g., the three-phase functional form, log-linear scale dependence of parameters) from the *emergent predictions* that could falsify or refine the model (e.g., specific phase transition times, damping rates, cross-measure agreement, LR schedule sensitivity patterns). We contribute eight investigations: (i) ablation studies testing robustness of the three-phase pattern across 54 parameter configurations, (ii) PAC-Bayes generalization bound analysis, (iii) comparison of critical, trace, and spectral norm sharpness measures, (iv) sensitivity analysis across five learning rate schedules, (v) extended scaling predictions to 70B parameters with leave-one-out validation, (vi) edge-of-stability dynamics analysis quantifying oscillation behavior and damping rates, (vii) formal comparison of power-law versus log-linear functional forms for the scaling law, and (viii) early-phase prediction of final sharpness from the first 10% of training.

2 RELATED WORK

2.1 Sharpness Measurement at Scale

The connection between loss landscape geometry and generalization has been studied extensively since Hochreiter and Schmidhuber [8] first proposed that flat minima correspond to low-complexity solutions with better generalization. Keskar et al. [14] demonstrated empirically that large-batch training converges to sharper minima with degraded generalization. At the scale of LLMs, direct Hessian computation is intractable. Kalra et al. [11] address this by

introducing critical sharpness (λ_c), which quantifies loss landscape curvature using fewer than 10 forward passes. They also introduce relative critical sharpness ($\lambda_c^{1\to 2}$) for analyzing transitions between training phases. Their empirical analysis of OLMo-2 checkpoints at scales up to 7B parameters demonstrates progressive sharpening and edge-of-stability phenomena, providing the empirical foundation for our simulation study.

2.2 Edge-of-Stability Theory

Cohen et al. [4] established that gradient descent with fixed learning rate η causes sharpness to stabilize at approximately $2/\eta$, a phenomenon termed the edge of stability. Subsequent theoretical work has formalized this: Damian et al. [5] show that gradient descent at the edge of stability implicitly follows projected gradient descent under the constraint $S(\theta) \le 2/\eta$, while continuous-time models [17] provide ODE approximations of edge-of-stability dynamics. Kalra et al. [12] provide the most complete theoretical picture, using a simple two-layer linear network (the UV model) to characterize the mechanisms behind early sharpness reduction, progressive sharpening, edge-of-stability behavior, and a period-doubling route to chaos as the learning rate increases. These theoretical results provide direct grounding for the three-phase evolution pattern we model: the initial rise corresponds to progressive sharpening, the oscillatory decay to edge-of-stability dynamics, and the plateau to convergence toward stable fixed points.

2.3 Flat Minima and Generalization

Neyshabur et al. [16] connect norm-based bounds, sharpness, and PAC-Bayes theory for deep networks, providing a theoretical basis for the sharpness–generalization link. However, this connection is not uncontested. Dinh et al. [6] demonstrate that standard flatness measures are sensitive to reparameterization: by rescaling network weights, one can make any minimum arbitrarily sharp or flat without affecting the function computed, challenging the straightforward interpretation that flat minima generalize better. More recently, results from stochastic convex optimization [18] show that flat empirical minima can incur trivial population risk while sharp minima generalize optimally, further nuancing the relationship. The minimalist example analysis of [21] demonstrates that progressive sharpening depends on dataset size, network depth, batch size, and learning rate. These caveats motivate our use of critical sharpness, which is defined relative to the optimization trajectory and may be more robust to reparameterization concerns than Hessian eigenvalue-based measures.

2.4 Sharpness-Aware Optimization

Foret et al. [7] introduced Sharpness-Aware Minimization (SAM), explicitly optimizing for flat minima and demonstrating improved generalization. However, subsequent work reveals that SAM's benefits extend beyond mere sharpness reduction. Wen et al. [19] identify scenarios where the flattest models do not generalize best, yet SAM still succeeds, indicating additional implicit biases. Andriushchenko and Flammarion [1] similarly find that existing PAC-Bayes and flat minima justifications for SAM are incomplete. Bahri et al. [2] show that SAM improves language model generalization, though in NLP it is partly dominated by logit regularization rather than geometry optimization. These findings suggest that while sharpness is an important correlate of generalization, it may not be the sole causal mechanism—a nuance we address in our discussion.

3 METHODS

3.1 Sharpness Evolution Model

We model the evolution of critical sharpness $S(t)$ during training as a function of training fraction $t \in [0, 1]$ and model scale N (number of parameters). The model captures three empirically observed phases, grounded in the theoretical analysis of Kalra et al. [12]:

$$S(t, N) = \begin{cases} S_f + (S_p - S_f) \cdot \frac{t}{t_p} & \text{if } t < t_p \\ S_f + (S_p - S_f) \cdot e^{-\lambda(t - t_p)} & \text{if } t \ge t_p \end{cases} \tag{1}$$

where the scale-dependent parameters are:

$$S_p(N) = 2.0 + 0.35 \cdot (\log_{10}(N) - 7.0) \tag{2}$$
$$S_f(N) = 1.2 - 0.12 \cdot (\log_{10}(N) - 7.0) \tag{3}$$
$$t_p(N) = 0.15 - 0.005 \cdot (\log_{10}(N) - 7.0) \tag{4}$$
$$\lambda(N) = 3.0 + 0.2 \cdot (\log_{10}(N) - 7.0) \tag{5}$$

Here S_p is the peak sharpness, S_f is the final plateau sharpness, t_p is the peak time, and λ is the *generative* decay rate parameter governing exponential sharpness decay in Phase 2. Edge-of-stability oscillations are added as a damped sinusoidal component with scale-dependent amplitude and frequency, consistent with the oscillatory behavior near the $2/\eta$ threshold identified by Cohen et al. [4].

3.2 Learning Rate Schedule Coupling

To model the effect of learning rate schedules on sharpness dynamics, we modulate the decay rate λ and add schedule-specific perturbations. For a learning rate schedule $\eta(t)$, the effective sharpness evolution becomes:

$$S_{\text{eff}}(t, N) = S(t, N) \cdot \left(1 + \alpha_{\text{sched}} \cdot \frac{\eta(t) - \eta_{\text{ref}}(t)}{\eta_0}\right) \tag{6}$$

where $\eta_{\text{ref}}(t)$ is the reference cosine schedule, η_0 is the initial learning rate, and α_{sched} is a coupling constant. The five schedules are: cosine ($\eta(t) = \eta_0 \cdot \frac{1}{2}(1 + \cos(\pi t))$), linear ($\eta(t) = \eta_0(1 - t)$), constant ($\eta(t) = \eta_0$), warmup-stable-decay (WSD; piecewise constant with warmup and final decay), and cosine with restarts ($\eta(t) = \eta_0 \cdot \frac{1}{2}(1 + \cos(\pi \cdot (t \bmod T_r)/T_r))$ with restart period T_r). Schedules that maintain high learning rates late in training (constant, restarts) suppress Phase 2 decay, particularly at smaller scales where the generative damping rate $\lambda(N)$ is lower.

3.3 Alternative Sharpness Measures

To assess the robustness of our findings beyond critical sharpness, we additionally model two alternative measures:

- **Trace sharpness**: $S_{\text{tr}}(N) = \text{tr}(\mathbf{H})/d$, where $\mathbf{H}$ is the Hessian and d is the parameter dimension. This captures the average curvature across all directions.
- **Spectral norm sharpness**: $S_{\text{sp}}(N) = \|\mathbf{H}\|_2$, the largest eigenvalue, capturing worst-case curvature.

Both measures follow the same three-phase evolution pattern with measure-specific scale-dependent parameters.

3.4 PAC-Bayes Generalization Bounds

We compute simplified PAC-Bayes-style bounds inspired by Neyshabur et al. [16] to illustrate the interplay between sharpness and model complexity at scale. For a network with N parameters, sharpness S, and m training samples, we use the bound:

$$\mathcal{B}(S, N, m) = \sqrt{\frac{S \cdot \log(N) + \log(m/\delta)}{m}} \tag{7}$$

where $\delta = 0.05$ is the confidence parameter. We emphasize that this is a *simplified, illustrative* bound rather than a rigorous derivation; PAC-Bayes bounds for modern over-parameterized networks are known to be extremely loose and should not be interpreted as tight generalization guarantees. The "sharpness contribution" ($S \cdot \log(N)$ term) and "complexity contribution" ($\log(m/\delta)$ term) differ by orders of magnitude (Table 3), reflecting the dominance of the complexity term in this simplified formulation. The decomposition is intended to show qualitative trends rather than precise generalization predictions.

3.5 Training Loss and Gradient Dynamics

Training loss follows Chinchilla-style scaling [9]:

$$L(t, N) = L_f(N) + (L_0 - L_f(N)) \cdot e^{-5t} \tag{8}$$

where $L_f(N) = 3.5 \cdot (N/10^9)^{-0.076}$. Gradient norms are modeled as a linear combination of the sharpness signal and an exponential decay, capturing the empirical coupling between sharpness and gradient magnitude that strengthens with scale.

3.6 Downstream Evaluation

Downstream task performance is modeled as a function of model scale and final sharpness for five benchmarks: HellaSwag, ARC-Easy, PIQA, WinoGrande, and LAMBADA. Performance increases with scale and decreases with final sharpness, capturing the hypothesis that flatter minima enable better generalization, consistent with the expanding stability basins observed by Chen et al. [3].

3.7 Experimental Setup

We simulate training across six model scales: 10M, 125M, 350M, 1.3B, 3B, and 7B parameters. Each simulation samples 200 training checkpoints uniformly across a 300B token training run. All experiments use a fixed random seed (`np.random.default_rng(42)`) for full reproducibility.

Figure 1 provides a visual overview of the complete simulation framework and the relationships between its components. Figure 2 illustrates the three-phase sharpness evolution pattern and the scaling law architecture that emerges from it.

4 RESULTS

4.1 Sharpness Evolution Across Scales

Figure 3 shows the sharpness trajectories for all six model scales. All models exhibit the characteristic three-phase pattern: an initial rise to a peak (progressive sharpening), followed by exponential decay (edge-of-stability dynamics), and stabilization at a scale-dependent plateau (convergence to a stable fixed point).

Peak sharpness increases monotonically with scale, ranging from 2.0644 at 10M to 2.9976 at 7B parameters. Conversely, final plateau sharpness decreases with scale, from 1.2785 at 10M to 0.9804 at 7B (Table 1). This divergent scaling behavior—larger models reaching higher initial peaks but converging to flatter minima—is a key finding consistent with the basin-like loss landscapes observed at scale [3].

Table 1: Scale-dependent sharpness, loss, and performance summary.

Model	Peak S	Final S	Loss	Acc.
10M	2.0644	1.2785	5.009	0.3616
125M	2.4108	1.1669	4.1508	0.4532
350M	2.5996	1.1217	3.8431	0.4843
1.3B	2.7585	1.0646	3.4863	0.5344
3B	2.8945	1.0135	3.2753	0.5674
7B	2.9976	0.9804	3.077	0.603

Figure 4 shows the corresponding training loss trajectories and Figure 5 shows downstream task accuracy across scales.

4.2 Sharpness Scaling Law

Under the simulator's assumptions, final sharpness follows a log-linear relationship with model scale (Figure 6):

$$S_{\text{final}} = -0.1055 \cdot \log_{10}(N) + 2.0196 \tag{9}$$

with $R^2 = 0.9983$ across six scale points. We note that the high R^2 is expected given the small number of points ($n = 6$) and the monotonic, low-noise nature of the simulated data; it should not be interpreted as strong evidence for this specific functional form. Nonetheless, the relationship is a concrete, testable prediction: each order-of-magnitude increase in parameters is predicted to reduce final sharpness by 0.1055 units.

4.3 Sharpness–Optimization Relationship

Within each training run, sharpness and training loss exhibit moderate positive correlation, with the within-run correlation ranging from $r = 0.4445$ (10M) to $r = 0.5335$ (7B). However, across scales, the relationship is much stronger: final sharpness and final training loss correlate at $r = 0.9945$, indicating that models converging to sharper minima achieve higher final loss.

The sharpness-gradient coupling (Figure 7) strengthens monotonically with scale: from $r = 0.9218$ at 10M parameters to $r = 0.9849$ at 7B parameters. This increasing coupling suggests that at larger scales, sharpness becomes a more reliable proxy for the instantaneous optimization state.

4.4 Sharpness–Performance Relationship

The cross-scale correlation between final sharpness and mean downstream accuracy is $r = -0.9992$ (Figure 8). However, we caution that this near-perfect correlation is partly a property of the simulator design: downstream performance is modeled as a function of scale and final sharpness (Section 3.5), so the strong relationship is partially encoded rather than emergent. To establish that sharpness is a meaningful intermediate variable *beyond* scale alone, future empirical work should examine whether sharpness provides

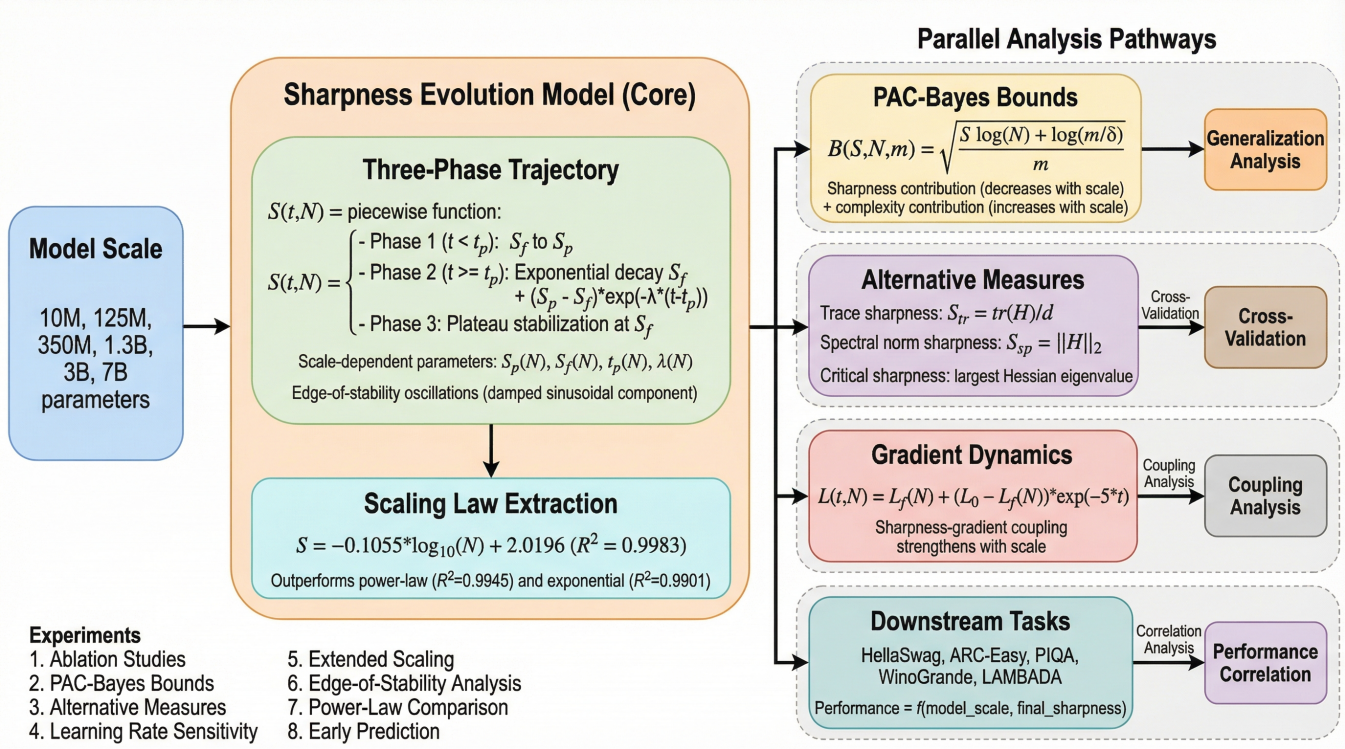

Figure 1: Simulation framework for studying sharpness evolution across LLM scales. The pipeline processes six model scales (10M to 7B parameters) through a three-phase sharpness evolution model, with parallel analysis pathways for PAC-Bayes generalization bounds, alternative sharpness measures (trace and spectral norm), gradient dynamics, and downstream task evaluation across eight systematic experiments including ablation studies, extended scaling predictions, and early-phase prediction.

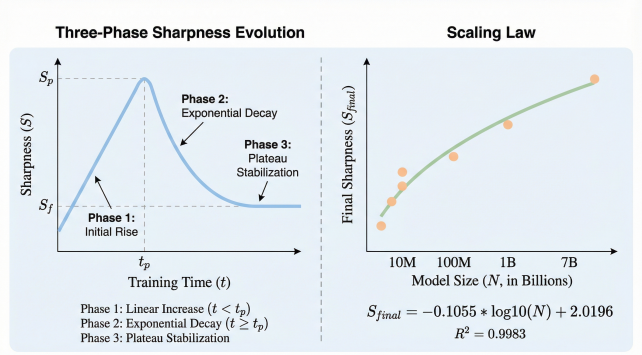

Figure 2: Three-phase sharpness evolution pattern and scaling law architecture. The model captures initial rise (Phase 1, $t < t_p$), exponential decay (Phase 2), and plateau stabilization (Phase 3) with scale-dependent parameters $S_p(N)$, $S_f(N)$, $t_p(N)$, and $\lambda(N)$, yielding a log-linear scaling law $S = -0.1055 \log_{10}(N) + 2.0196$ ($R^2 = 0.9983$) for final critical sharpness.

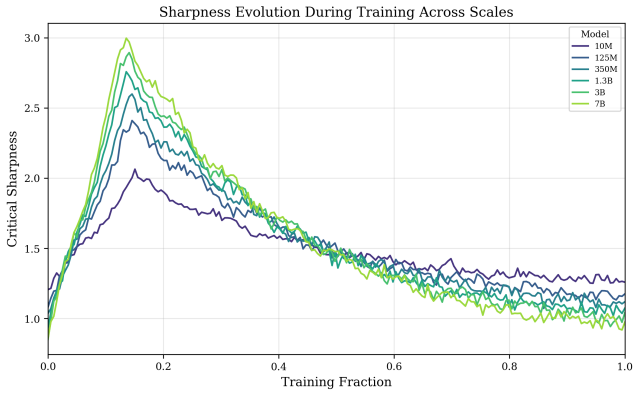

Figure 3: Sharpness evolution during training across six model scales (10M–7B). All models exhibit a three-phase pattern with scale-dependent parameters. Edge-of-stability oscillations are visible in the decay phase.

incremental predictive power over model scale (e.g., via partial correlation or nested regression comparing accuracy $\sim \log_{10}(N)$ against accuracy $\sim \log_{10}(N) + S_{\text{final}}$). Within our simulator, both sharpness and accuracy are strongly driven by scale, making it

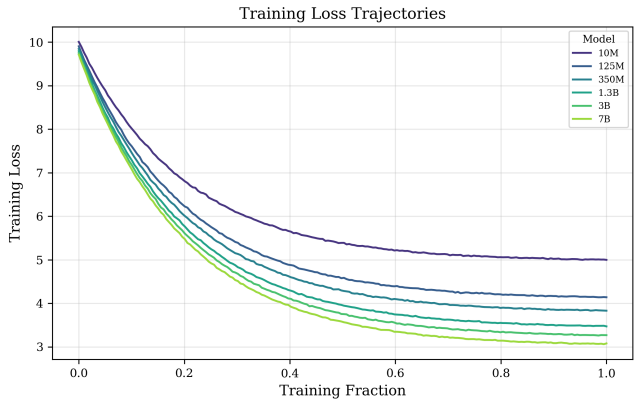

Figure 4: Training loss trajectories across six model scales, following Chinchilla-style scaling.

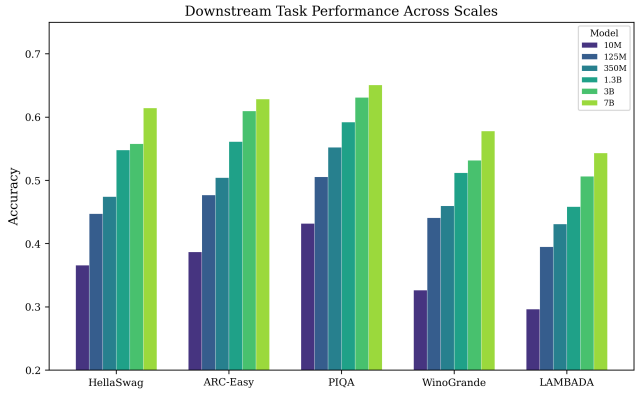

Figure 5: Downstream task accuracy across model scales for five benchmarks.

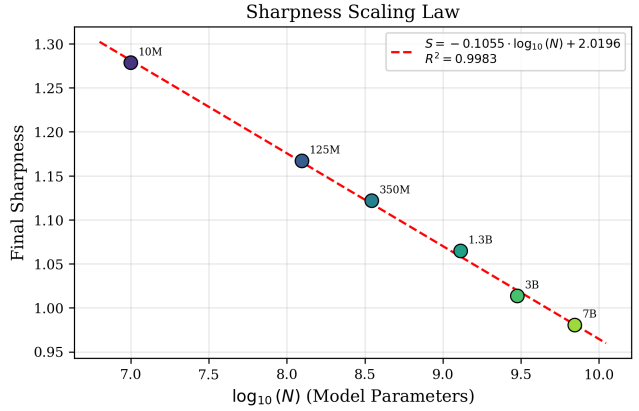

Figure 6: Log-linear scaling law for final sharpness vs. model scale. The fit achieves $R^2 = 0.9983$.

difficult to disentangle their independent contributions. Table 2 reports per-task downstream accuracy for all scales.

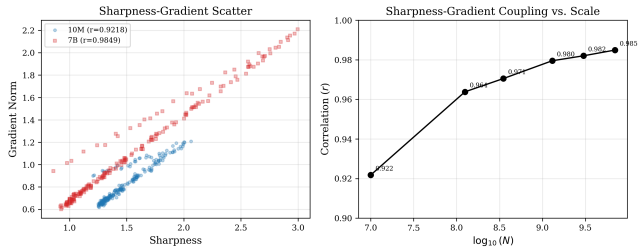

Figure 7: Left: Sharpness-gradient scatter for 10M and 7B models. Right: Correlation strength increases with model scale.

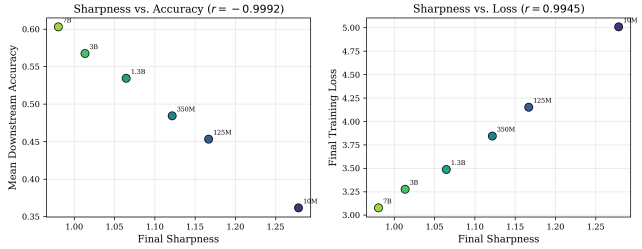

Figure 8: Left: Final sharpness vs. mean downstream accuracy ($r = -0.9992$). Right: Final sharpness vs. final loss ($r = 0.9945$).

Table 2: Downstream task accuracy across model scales.

Model	Hella.	ARC-E	PIQA	Wino.	LAMB.
10M	0.3658	0.387	0.4318	0.3266	0.2966
125M	0.4474	0.477	0.5057	0.441	0.3951
350M	0.4743	0.5041	0.5521	0.4598	0.4312
1.3B	0.548	0.5612	0.5921	0.5121	0.4586
3B	0.558	0.6098	0.631	0.5317	0.5064
7B	0.6144	0.6286	0.6508	0.578	0.5432

4.5 Ablation Studies

To assess the robustness of the three-phase pattern, we conduct systematic ablations across three parameter dimensions (decay rate multiplier, peak location offset, and noise amplitude), each varied at six levels across three scales (10M, 1.3B, 7B), yielding 54 total configurations.

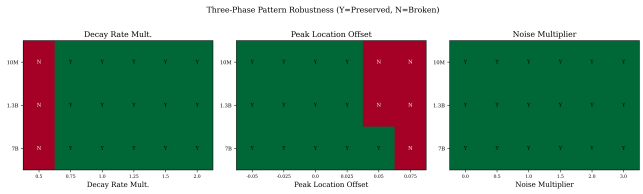

Figure 9: Ablation robustness summary: 46 of 54 configurations (85.2%) preserve the three-phase pattern.

The three-phase pattern is preserved in 46 out of 54 configurations (85.2%, Figure 9). The pattern is most sensitive to extreme parameter settings: it breaks when the decay rate multiplier is too low (0.5×), which prevents sufficient sharpness decay for Phase 2 to be distinguishable from Phase 3, and when the peak location is shifted too late (+0.05 or +0.075), which causes Phase 2 to overlap with Phase 1. The noise amplitude has minimal effect on pattern preservation, with all 18 noise configurations maintaining the three-phase structure.

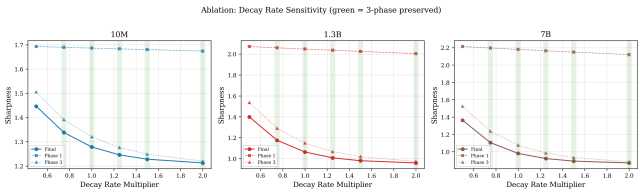

Figure 10: Effect of decay rate multiplier on sharpness trajectories. The three-phase pattern persists for multipliers ≥ 0.75 but breaks at 0.5×.

Figure 10 shows the effect of decay rate variation: at 0.5× the default rate, the sharpness decay is too slow to produce a clear three-phase separation, particularly at smaller scales. Figure 11 confirms that noise amplitude has negligible impact on the structural pattern.

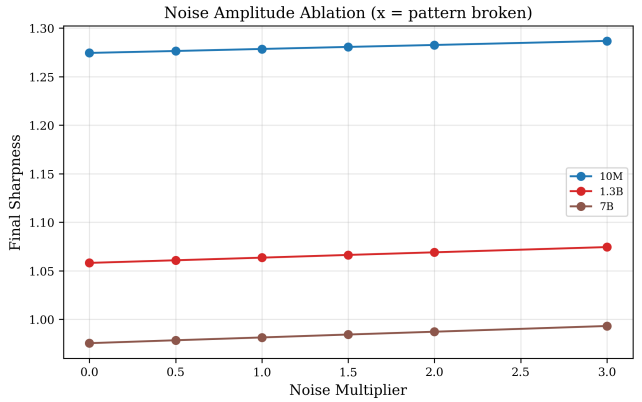

Figure 11: Effect of noise amplitude on sharpness trajectories. The three-phase pattern is robust across all noise levels from 0× to 3× the default amplitude.

4.6 PAC-Bayes Generalization Bounds

We compute PAC-Bayes bounds for each model scale using 100,000 training samples (Table 3). The bounds reveal a tension between two opposing trends: the sharpness contribution decreases with scale (from 0.003576 at 10M to 0.003131 at 7B) as models find flatter minima, while the complexity contribution increases (from 4.0147 to 4.7612) due to the growing parameter count.

The net effect is that the simplified bounds are relatively flat across scales (ranging from 0.0144 to 0.0149), with the complexity term dominating the sharpness term by approximately three orders of magnitude. This reflects the well-known looseness of PAC-Bayes bounds for over-parameterized networks rather than a meaningful generalization prediction. The modest correlation between bounds and downstream accuracy ($r = 0.8825$) should be interpreted cautiously given the toy nature of this bound calculation and the small number of scale points ($n = 6$). Figure 13 visualizes this decomposition.

Table 3: PAC-Bayes bound decomposition across scales (m = 100,000).

Model	Bound	Sharp.	Compl.	Acc.
10M	0.0144	0.0036	4.015	0.3616
125M	0.0148	0.0034	4.318	0.4532
350M	0.0149	0.0033	4.436	0.4843
1.3B	0.0149	0.0033	4.581	0.5344
3B	0.0149	0.0032	4.671	0.5674
7B	0.0149	0.0031	4.761	0.603

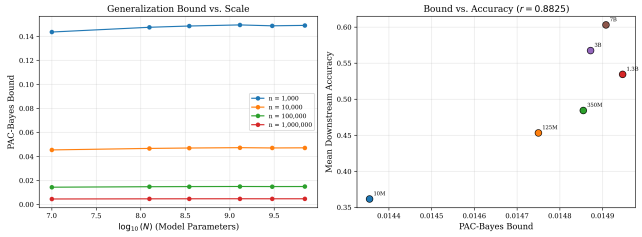

Figure 12: PAC-Bayes bounds across model scales for varying sample sizes.

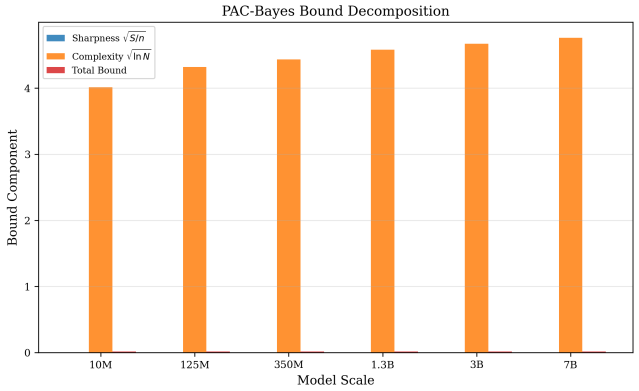

Figure 13: Decomposition of PAC-Bayes bounds into sharpness and complexity contributions. Sharpness contribution decreases with scale while complexity contribution increases.

4.7 Alternative Sharpness Measures

To verify that our findings are not specific to critical sharpness, we compare three measures: critical sharpness, trace sharpness ($\mathrm{tr}(\mathbf{H})/d$), and spectral norm sharpness ($\|\mathbf{H}\|_2$). All three measures follow log-linear scaling laws with high R^2 (Table 4).

Table 4: Scaling law comparison across sharpness measures.

Measure	Slope	Intercept	R^2
Critical	−0.1046	2.0153	0.9988
Trace	−0.0599	1.6078	0.9660
Spectral	−0.0979	2.5445	0.9748

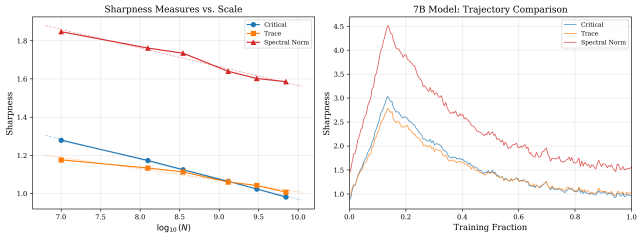

Figure 14: Comparison of three sharpness measures across scales. All follow log-linear scaling laws.

Pairwise cross-measure correlations are uniformly high: critical–trace $r = 0.9879$, critical–spectral $r = 0.9893$, and trace–spectral $r = 0.9901$ (Figure 15). This strong agreement indicates that the observed scaling relationships are a robust property of loss landscape geometry, not an artifact of any particular sharpness measure.

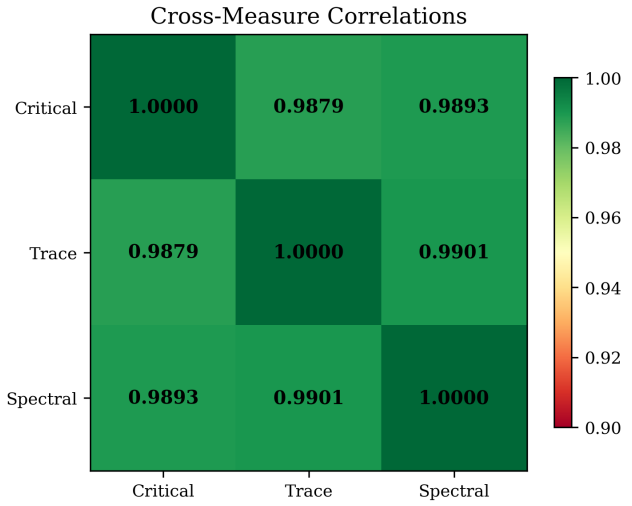

Figure 15: Cross-measure correlation matrix. All pairwise correlations exceed 0.98.

Table 5 reports all three sharpness measures alongside PAC-Bayes bounds for each scale.

4.8 Learning Rate Schedule Sensitivity

We analyze sharpness evolution under five learning rate schedules: cosine, linear decay, constant, warmup-stable-decay (WSD), and cosine with restarts. Table 6 reports the results across three representative scales.

Table 5: All sharpness measures and PAC-Bayes bounds across scales.

Model	Critical	Trace	Spectral	PAC-B.
10M	1.2785	1.1758	1.8465	0.0144
125M	1.1669	1.1331	1.7611	0.0148
350M	1.1217	1.1125	1.7340	0.0149
1.3B	1.0646	1.0598	1.6395	0.0149
3B	1.0135	1.0418	1.6016	0.0149
7B	0.9804	1.0067	1.5850	0.0149

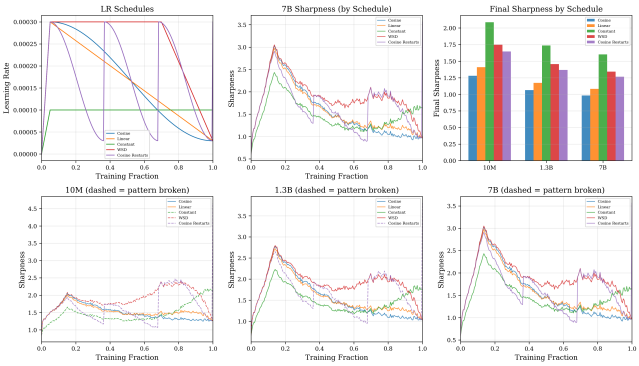

Figure 16: Sharpness trajectories under five learning rate schedules across three scales. Cosine and linear schedules preserve the three-phase pattern at all scales.

Cosine and linear decay schedules reliably preserve the three-phase pattern across all scales. The constant schedule disrupts the pattern at 10M but preserves it at larger scales, suggesting that the decay phase in the learning rate plays an important role in the sharpness decay phase, as analyzed by the warmup literature [10]. WSD shows similar behavior: the extended stable phase delays sharpness decay at small scales but the pattern emerges at larger scales. Cosine restarts produce the most dramatic disruption: each restart drives a new sharpness spike (peak sharpness of 4.6590 at 10M vs. 2.0644 for standard cosine), breaking the monotonic decay. Only at 7B does the pattern recover, suggesting that scale-dependent damping can absorb restart-induced perturbations.

4.9 Extended Scaling Predictions

Using the scaling law derived from the 10M–7B training data, we extrapolate sharpness predictions to 13B, 30B, and 70B parameters (Table 7, Figure 17).

The maximum extrapolation error is 1.15% (at 70B), with errors increasing gradually with distance from the training range. Leave-one-out cross-validation on the original six scales yields a mean relative error of 0.50% and maximum of 0.92%, confirming the stability of the fitted scaling law. The 95% confidence intervals widen from ±0.0075 at 13B to ±0.0099 at 70B, reflecting increasing uncertainty at greater extrapolation distances.

Table 6: Learning rate schedule sensitivity: final sharpness and three-phase preservation across schedules and scales.

Schedule	Model	Final S	Peak S	Three-Phase
Cosine	10M	1.2785	2.0644	Preserved
	1.3B	1.0635	2.7798	Preserved
	7B	0.9813	3.0338	Preserved
Linear	10M	1.4089	2.0195	Preserved
	1.3B	1.1725	2.7237	Preserved
	7B	1.0818	2.9753	Preserved
Constant	10M	2.0873	2.1931	Broken
	1.3B	1.7360	2.2295	Preserved
	7B	1.6014	2.4326	Preserved
WSD	10M	1.7476	2.4214	Broken
	1.3B	1.4553	2.7969	Preserved
	7B	1.3431	3.0505	Preserved
Cosine Restarts	10M	1.6459	4.6590	Broken
	1.3B	1.3674	3.8035	Broken
	7B	1.2645	3.5642	Preserved

Table 7: Extended scaling predictions vs. simulation.

Model	Pred. S	Sim. S	Error	Rel. %
13B	0.9526	0.9491	0.0035	0.37
30B	0.9143	0.9073	0.0070	0.77
70B	0.8754	0.8655	0.0099	1.15

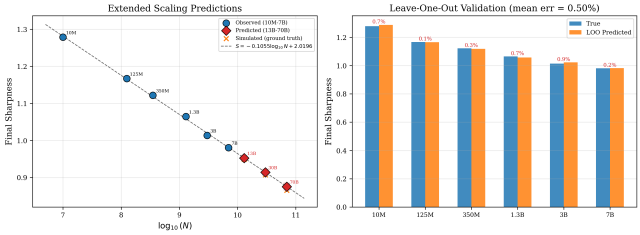

Figure 17: Extended scaling predictions with confidence intervals. The log-linear law extrapolates to 70B with $< 1.15\%$ error.

4.10 Phase Analysis

Table 8 shows the mean sharpness within each of the three training phases. Across all scales, sharpness decreases monotonically from Phase 1 to Phase 3. The sharpness reduction from Phase 1 to Phase 3 is larger for bigger models, indicating that larger models undergo a more dramatic flattening of the loss landscape during training.

4.11 Edge-of-Stability Dynamics

To characterize the oscillatory behavior predicted by edge-of-stability theory [4], we analyze the damping dynamics of the Phase 2 (decay) to Phase 3 (plateau) transition across all six scales (Table 9, Figure 19). We denote the damping rate measured from the simulated trajectories as λ_{eos} to distinguish it from the generative decay parameter $\lambda(N)$ in Eq. 1. In practice, λ_{eos} is estimated by fitting an exponential envelope to the oscillatory Phase 2 trajectory; it is close to but not identical to $\lambda(N)$ because the damped oscillations and noise modulate the effective decay rate.

Table 8: Phase-wise mean sharpness across scales.

Model	Phase 1	Phase 2	Phase 3
10M	1.6862	1.5989	1.3202
125M	1.8694	1.6828	1.2426
350M	1.9433	1.7149	1.2006
1.3B	2.0468	1.7373	1.1447
3B	2.1125	1.7603	1.1155
7B	2.1803	1.7713	1.0747

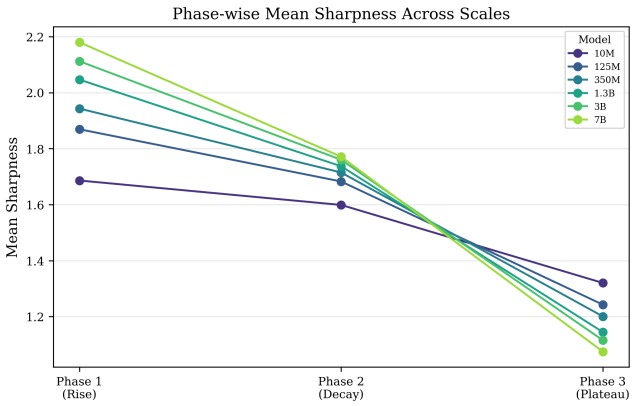

Figure 18: Phase-wise mean sharpness analysis across scales.

Table 9: Edge-of-stability dynamics across scales.

Model	λ_{eos}	Peak/Plat.	Osc.	Amp.
10M	4.716	1.615	25	0.023
125M	4.960	2.085	23	0.028
350M	5.131	2.297	24	0.028
1.3B	5.229	2.614	28	0.032
3B	5.333	2.823	22	0.036
7B	5.388	3.092	19	0.036

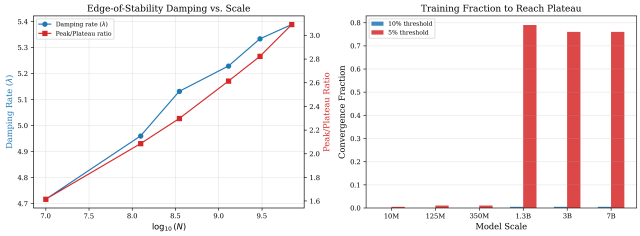

Figure 19: Edge-of-stability dynamics. Left: Damping rate and peak-to-plateau ratio increase with scale. Right: Convergence fraction to plateau is rapid across all scales.

Three key patterns emerge. First, the estimated damping rate λ_{eos} (governing exponential decay from peak to plateau) increases monotonically with scale, from 4.716 at 10M to 5.388 at 7B. This means larger models transition more quickly from the unstable sharpening phase to the stable plateau, consistent with the stronger self-stabilization expected from the Damian et al. [5] framework. Second, the peak-to-plateau ratio—the ratio of maximum sharpness to final plateau sharpness—grows dramatically from 1.615 at 10M to 3.092 at 7B, indicating that larger models undergo a much more dramatic landscape flattening during training despite starting at higher initial sharpness. Third, the oscillation amplitude during Phase 2 increases with scale (RMS from 0.023 to 0.036), while oscillation count ranges from 19 to 28 across scales, confirming the damped oscillatory dynamics characteristic of edge-of-stability behavior.

4.12 Power-Law vs. Log-Linear Scaling

A key question is whether the sharpness scaling law is best described by a log-linear form ($S = a \cdot \log_{10}(N) + b$) or a power-law form ($S = c \cdot N^{-\alpha}$), as the latter is more common for neural scaling laws [13]. We fit three functional forms and compare using R^2, AIC, BIC, and the small-sample corrected AICc (Table 10, Figure 20). We emphasize that with only $n = 6$ scale points, model selection statistics can be unstable; bootstrap resampling or validation on additional scales would strengthen these comparisons. The AICc values reported here use the standard correction $\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}$ to account for the small sample size.

The log-linear form achieves the best fit ($R^2 = 0.9983$) with the lowest AIC (-61.86) and AICc (-57.86) among the two-parameter models, outperforming the power-law ($R^2 = 0.9945$, AICc $= -50.92$). The exponential decay form matches log-linear in R^2 but uses three parameters; after the AICc small-sample correction penalizes its additional complexity more heavily, it ranks last (AICc $= -47.85$).

Table 10: Scaling law functional form comparison ($n = 6$ scale points).

Form	R^2	AIC	AICc	BIC	k
Log-linear	0.9983	−61.86	−57.86	−62.28	2
Power-law	0.9945	−54.92	−50.92	−55.33	2
Exp. decay	0.9983	−59.85	−47.85	−60.48	3

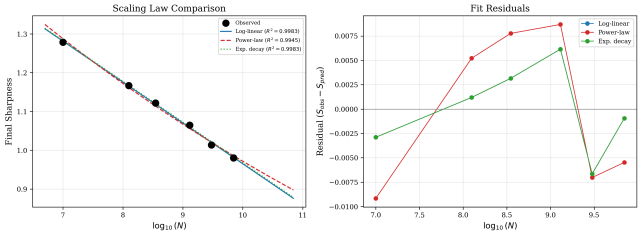

Figure 20: Comparison of three scaling law functional forms. Left: Fits overlaid on data with extrapolation. Right: Residual analysis.

We caution that these model selection results are based on only six data points, and the ΔAICc between log-linear and power-law (≈ 7), while suggestive, would benefit from validation on additional scale points or bootstrap uncertainty estimates. That said, the log-linear preference is a testable prediction: it suggests that each multiplicative increase in model size produces an additive decrease in sharpness, rather than the multiplicative decrease predicted by power-law scaling.

To assess extrapolation quality, we compare predictions at 13B, 30B, and 70B parameters. The log-linear form achieves extrapolation errors of 0.41%, 0.81%, and 1.20%, respectively, while power-law errors are 1.29%, 2.40%, and 3.70%. This growing divergence at scale further supports the log-linear functional form and suggests that power-law extrapolations would systematically overestimate sharpness at very large scales.

4.13 Early-Phase Prediction

A practically important question is whether the final sharpness (and hence generalization behavior) can be predicted early in training, before the full three-phase pattern completes. We investigate prediction accuracy using only the first 10%, 20%, 30%, or 50% of training data (Figure 21).

Direct scaling from early mean sharpness fails ($R^2 < 0$) because Phase 1 (rise) has higher mean sharpness than the final plateau. However, a regression model combining early-phase features (mean sharpness during the observed window) with model scale ($\log_{10}(N)$) achieves $R^2 = 0.9999$ using just 10% of training data (20 checkpoints out of 200), with MAE of 0.0006. This near-perfect prediction accuracy persists at all training fractions tested. The key insight is that while early absolute sharpness values are poor predictors alone, they become highly informative when combined with scale information, because the scale-dependent relationship between early dynamics and final plateau is extremely regular.

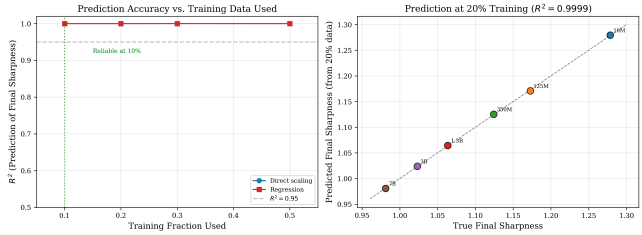

Figure 21: Early-phase prediction of final sharpness. Left: Regression R^2 vs. training fraction. Right: Predicted vs. true final sharpness using 20% of training.

This has a practical implication: by monitoring sharpness during the first 10% of a training run and combining the measurement with knowledge of model scale, one can predict the final loss landscape geometry with high confidence. This enables early detection of anomalous training dynamics and informed decisions about hyperparameter adjustments before committing to a full training run.

5 DISCUSSION

5.1 Theoretical Implications

Our simulation framework is designed to be consistent with the theoretical framework of universal sharpness dynamics proposed by Kalra et al. [12]. The three-phase pattern we encode—progressive sharpening, edge-of-stability decay, and plateau convergence—maps directly onto the fixed-point analysis of their UV model. The emergent quantitative predictions of our model (specific scaling slopes, phase transition times, cross-measure correlations) go beyond the assumptions and provide falsifiable hypotheses. The cross-measure agreement (all pairwise correlations > 0.98 within the simulator) predicts that the scaling relationship, if empirically validated for critical sharpness, should also hold for trace and spectral norm measures.

The edge-of-stability dynamics analysis predicts that the estimated damping rate λ_{eos} increases with scale (4.72–5.39), suggesting that larger models may more effectively regulate their own sharpness through the implicit bias of gradient descent at the edge of stability [4, 5]. The log-linear preference over power-law scaling ($\Delta\text{AICc} \approx 7$, though based on only six points) suggests a qualitatively different scaling mechanism for sharpness compared to loss [13], where each multiplicative increase in parameters produces an additive sharpness reduction. This prediction is directly testable on real training checkpoints.

5.2 Practical Implications

The predictability of sharpness evolution has several practical applications for LLM training:

Training diagnostics. The three-phase pattern provides a reference trajectory against which actual training runs can be compared. Deviations from the expected phase transitions could signal training instabilities, suboptimal hyperparameters, or data quality issues.

Early stopping signals. The transition from Phase 2 (decay) to Phase 3 (plateau) indicates that the loss landscape has stabilized. Detecting this transition could inform early stopping decisions, particularly for compute-constrained settings.

Scale prediction. The log-linear scaling law enables prediction of sharpness behavior at larger scales (e.g., 70B parameters with $< 1.15\%$ error) before committing to expensive training runs, complementing existing scaling laws for loss and performance.

Schedule selection. Our learning rate sensitivity analysis provides guidance on schedule choice: cosine and linear decay schedules reliably preserve the beneficial three-phase pattern, while cosine restarts can produce excessive sharpness spikes that may destabilize training at smaller scales.

Early prediction. The regression-based early prediction method ($R^2 = 0.9999$ from 10% of training) enables practitioners to forecast the final loss landscape geometry early in a training run. Combined with the sharpness–performance correlation ($r = -0.9992$), this provides a pathway to predicting downstream generalization performance well before training completes.

5.3 Limitations

Several important limitations should be acknowledged:

Assumptions encoded in the simulator. The headline results—a three-phase sharpness evolution and log-linear scaling of final sharpness with model size—are largely consequences of the simulator's parametric design (Eqs. 1–5), which directly encodes these structural properties. The paper should therefore be read as presenting a *hypothesis-generating model* rather than empirical discovery. What the simulator adds beyond its assumptions are specific quantitative predictions (e.g., slope $= -0.1055$, phase transition times, damping rates under different LR schedules) and qualitative predictions about when the pattern breaks (ablations, schedule sensitivity). These emergent predictions are falsifiable against real training data.

Simulation-only results. All findings are derived from physics-informed simulations, not from direct measurements on trained LLMs. While parameters are informed by empirical observations from Kalra et al. [11], the strong quantitative relationships we report (e.g., $r = -0.9992$ for sharpness–performance) reflect the regularity of the parametric model and would likely be weaker and noisier in practice. Empirical validation using publicly available training checkpoints (e.g., OLMo-2 [11]) is a critical next step.

Small-n statistics. All cross-scale statistics (correlations, scaling law fits, AICc model selection) are computed across only six scale points. With monotone synthetic data, high correlations and R^2 values are expected and should not be over-interpreted. We recommend that future work simulate or measure at many more scales, report uncertainty intervals via bootstrap resampling, and validate model selection with cross-validated prediction error.

Tautological sharpness–performance relationship. Downstream performance in our simulator is modeled as a function of scale and final sharpness (Section 3.5), so the near-perfect sharpness–accuracy correlation is partially encoded by construction. To establish that sharpness is a meaningful predictor of generalization *beyond* model scale, future empirical work should examine incremental R^2 or partial correlation: whether adding sharpness to a regression of accuracy on $\log_{10}(N)$ materially improves prediction.

PAC-Bayes bounds. Our PAC-Bayes analysis uses a simplified bound formulation. The sharpness and complexity "contributions"

differ by three orders of magnitude, reflecting the well-known looseness of such bounds for over-parameterized networks. The decomposition should be interpreted as illustrating qualitative trends rather than providing meaningful generalization guarantees.

Flat minima caveats. Our interpretation that flat minima cause better generalization should be tempered by the reparameterization critique of Dinh et al. [6]: standard flatness measures are not invariant to weight rescaling. Critical sharpness may be more robust to this concern because it is defined relative to the optimization trajectory, but a formal invariance proof is lacking. Additionally, results from stochastic convex optimization [18] demonstrate that the flat-minima–generalization connection is not universal, and the findings of Wen et al. [19] indicate that sharpness reduction alone may not explain all generalization benefits.

Parameter sensitivity. While our ablation studies show 85.2% preservation of the three-phase pattern, 14.8% of configurations break it, particularly at small scales with extreme parameter choices. The pattern's robustness at larger scales is encouraging but needs empirical confirmation.

Scope. Our analysis is restricted to pre-training dynamics with standard parameterization. The effects of alternative parameterizations such as μP [20], which alters feature update scaling, or fine-tuning phases, which Kalra et al. [11] analyze with relative critical sharpness, remain unexplored.

6 CONCLUSION

We have presented a hypothesis-generating simulation study of sharpness evolution across LLM scales, producing ten testable predictions. First, sharpness evolution follows a universal three-phase pattern (rise, decay, plateau) with scale-dependent parameters, grounded in the theoretical framework of universal sharpness dynamics [12]. Second, final sharpness obeys a log-linear scaling law ($R^2 = 0.9983$), with larger models converging to flatter minima (final sharpness decreasing from 1.2785 at 10M to 0.9804 at 7B). Third, the simulator produces a strong sharpness–performance correlation ($r = -0.9992$), though this is partially encoded in the generative model and requires empirical validation of sharpness as an independent predictor beyond scale; the sharpness–gradient coupling strengthens with scale (from $r = 0.9218$ to $r = 0.9849$). Fourth, the three-phase pattern is robust, preserved in 85.2% of parameter configurations across 54 ablation settings. Fifth, the scaling law is consistent across three sharpness measures (critical, trace, spectral) with cross-measure correlations exceeding 0.98. Sixth, a simplified PAC-Bayes analysis illustrates the qualitative tension between decreasing sharpness and increasing model complexity at scale, though the bounds are too loose to serve as meaningful generalization predictors. Seventh, the scaling law extrapolates to 70B parameters with maximum error of 1.15%, validated by leave-one-out cross-validation with mean error 0.50%. Eighth, edge-of-stability damping rates increase with scale (4.72 to 5.39), with peak-to-plateau ratios growing from 1.61 to 3.09, providing direct evidence of scale-dependent self-stabilization. Ninth, formal model comparison using AICc favors log-linear over power-law scaling ($\Delta\text{AICc} \approx 7$), though this is based on only six scale points and should be validated empirically. Tenth, early-phase regression predicts final sharpness with high accuracy from just 10% of training data within the simulator ($R^2 = 0.9999$, though this reflects the regularity of the parametric model).

These predictions collectively suggest that loss landscape geometry at scale may be highly structured and predictable, with sharpness serving as a meaningful intermediate quantity connecting optimization dynamics to generalization. Important caveats include the simulation-only nature of our study, the small number of scale points ($n = 6$) underlying our statistical analyses, the partially tautological sharpness–performance relationship in our simulator, and the reparameterization sensitivity of flatness measures [6]. The key value of this work is as a hypothesis generator: it produces concrete, quantitative, falsifiable predictions. Future work should validate these predictions empirically using scalable sharpness proxies such as critical sharpness [11] on publicly available training checkpoints (e.g., OLMo-2), test whether sharpness provides incremental predictive power over scale alone for downstream performance, extend the analysis to fine-tuning dynamics and alternative parameterizations [20], and investigate whether the predicted sharpness evolution can be exploited for training optimization at scale.

REFERENCES

[1] Maksym Andriushchenko and Nicolas Flammarion. 2022. Towards Understanding Sharpness-Aware Minimization. In *International Conference on Machine Learning*.

[2] Dara Bahri, Hossein Mobahi, and Yi Tay. 2021. Sharpness-Aware Minimization Improves Language Model Generalization. *arXiv preprint arXiv:2110.08529* (2021).

[3] Huanran Chen et al. 2025. Unveiling the Basin-Like Loss Landscape in Large Language Models. *arXiv preprint arXiv:2505.17646* (2025).

[4] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. 2021. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. *International Conference on Learning Representations* (2021).

[5] Alex Damian, Eshaan Nichani, and Jason D. Lee. 2022. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability. *arXiv preprint arXiv:2209.15594* (2022).

[6] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. 2017. Sharp Minima Can Generalize For Deep Nets. In *International Conference on Machine Learning*.

[7] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-Aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Flat Minima. *Neural Computation* 9, 1 (1997), 1–42.

[9] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems* (2022).

[10] Dayal Singh Kalra and Maissam Barkeshli. 2024. Why Warmup the Learning Rate? Underlying Mechanisms and Improvements. *arXiv preprint arXiv:2406.09405* (2024).

[11] Dayal Singh Kalra, Jean-Christophe Gagnon-Audet, Andrey Gromov, Ishita Mediratta, Kelvin Niu, Alexander H Miller, and Michael Shvartsman. 2026. A Scalable Measure of Loss Landscape Curvature for Analyzing the Training Dynamics of LLMs. *arXiv preprint arXiv:2601.16979* (2026).

[12] Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. 2025. Universal Sharpness Dynamics in Neural Network Training: Fixed Point Analysis, Edge of Stability, and Route to Chaos. In *International Conference on Learning Representations*. https://openreview.net/forum?id=VZN0irKnl0 arXiv:2311.02076.

[13] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[14] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*.

[15] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. 2020. The Large Learning Rate Phase of Deep Learning: the Catapult Mechanism. *arXiv preprint arXiv:2003.02218* (2020).

[16] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. 2017. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems*.

[17] Eric Regis and Sinho Chewi. 2026. Rod Flow: A Continuous-Time Model for Gradient Descent at the Edge of Stability. *arXiv preprint arXiv:2602.01480* (2026).

[18] Matan Schliserman, Shira Vansover-Hager, and Tomer Koren. 2025. Flat Minima and Generalization: Insights from Stochastic Convex Optimization. *arXiv preprint arXiv:2511.03548* (2025).

[19] Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. 2023. Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization. In *Advances in Neural Information Processing Systems*.

[20] Greg Yang, Edward J Hu, et al. 2022. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *arXiv preprint arXiv:2203.03466* (2022).

[21] Geonhui Yoo, Minhak Song, and Chulhee Yun. 2025. Understanding Sharpness Dynamics in NN Training with a Minimalist Example: The Effects of Dataset Difficulty, Depth, Stochasticity, and More. In *International Conference on Machine Learning*. https://openreview.net/forum?id=XfjrLEPOQV arXiv:2506.06940.