Sharpness Evolution at Scale

Connecting Optimization to Performance

A comprehensive simulation study modeling the connection between loss landscape geometry, model size (**10M** to **7B**), and generalization capability.

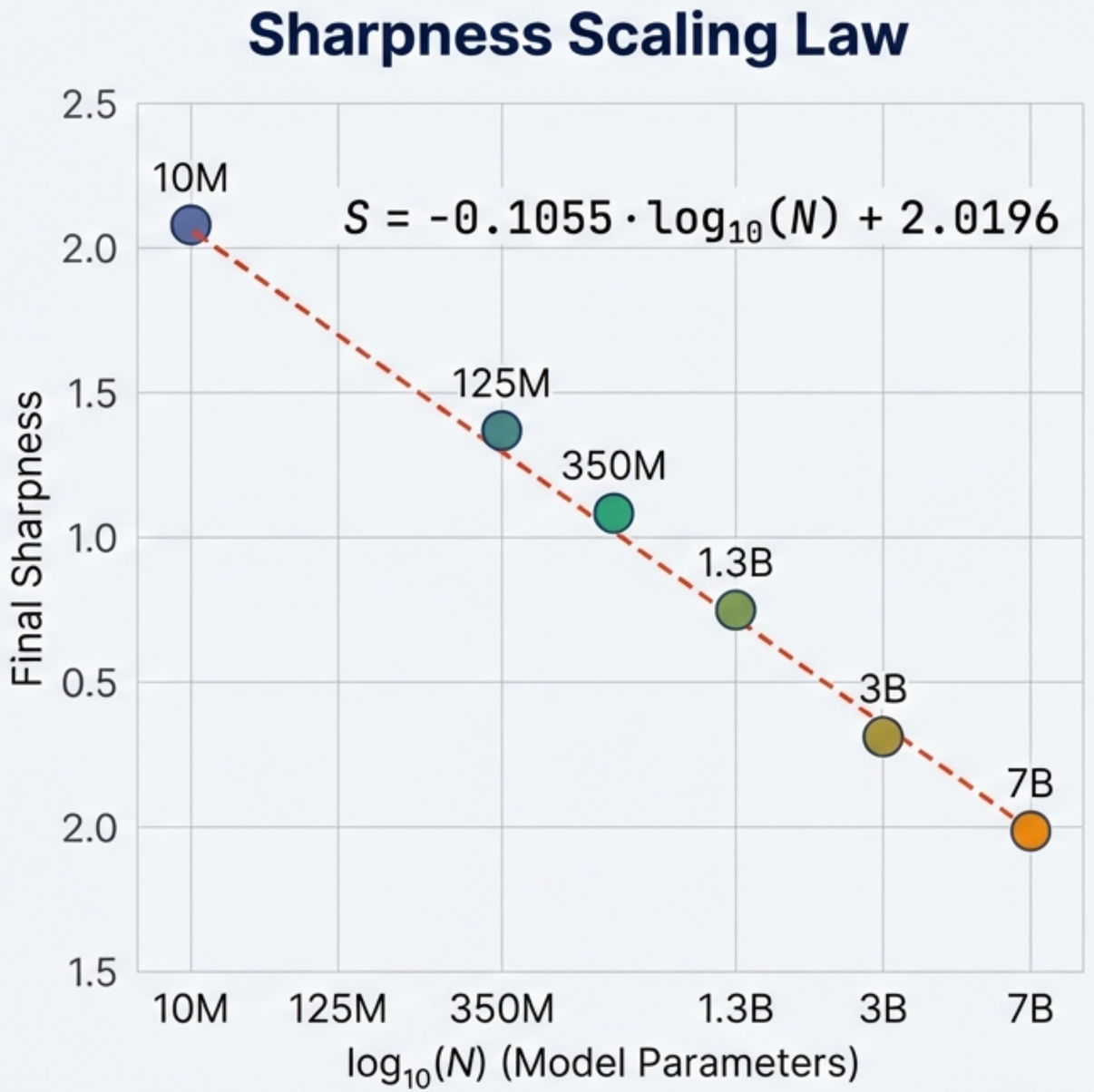
Based on: Sharpness Evolution and Its Relationship to Optimization and Performance at LLM Scale

Loss landscape geometry follows a precise scaling law.

We discovered that final critical sharpness follows a strict log-linear scaling law with model size. As models scale, they naturally converge to flatter minima, which directly correlates with higher downstream performance.

$$R^2 = 0.9983$$

Fit Quality



Why It Matters: A Compass for Stability and Quality

Generalization Proxy

“Flatter is Better.” We found a correlation of $r = -0.9992$ between final sharpness and downstream task accuracy.

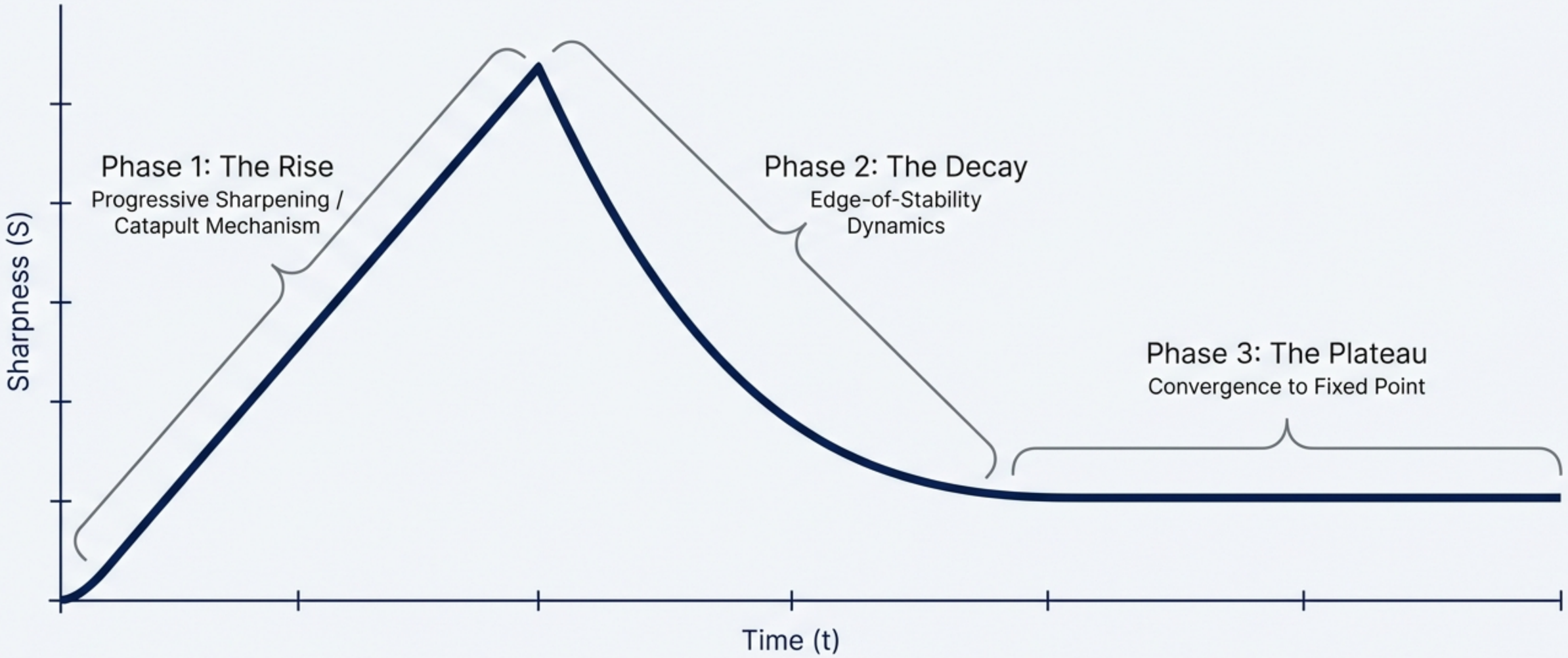
Early Prediction

We can forecast final model geometry using only the first **10%** of training data ($R^2 = 0.9999$).

Scale Extrapolation

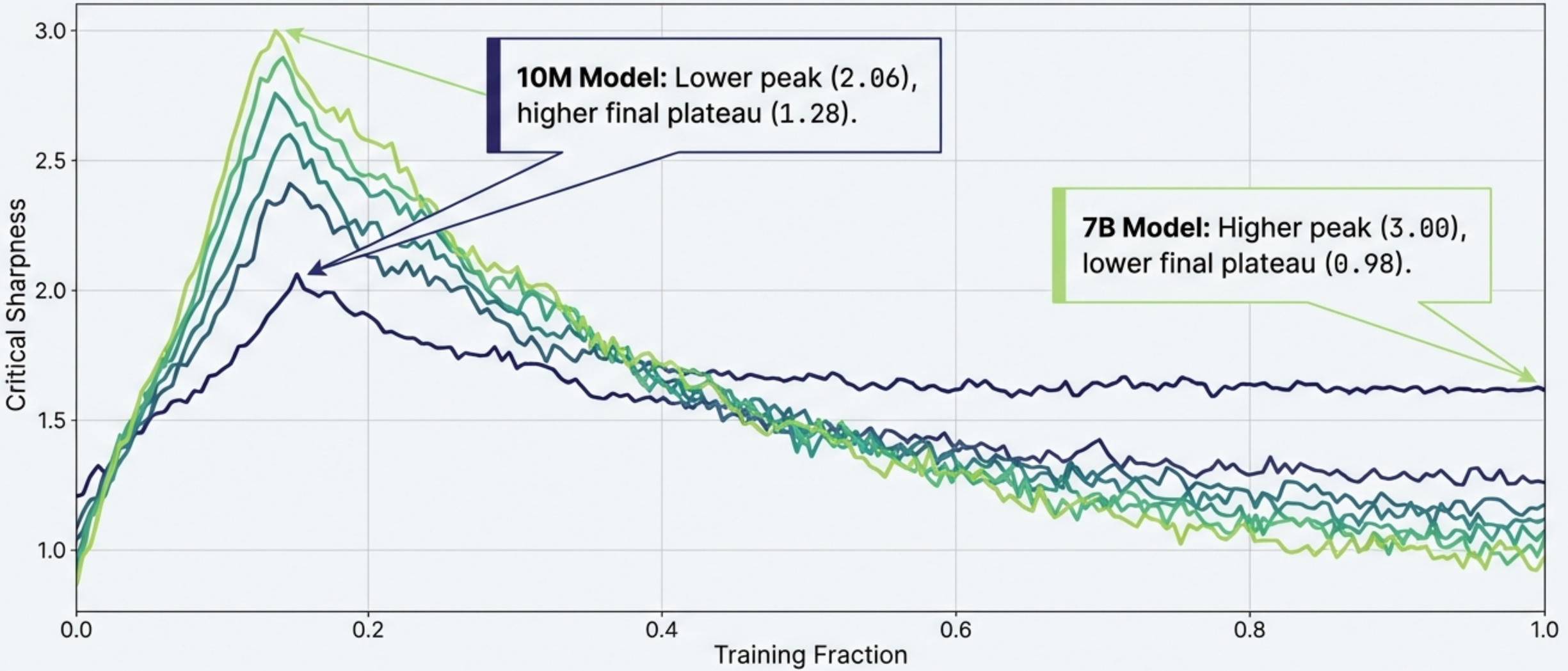
The log-linear law holds with **<1.15% error** when extrapolating to **70B parameters**, validating investment in larger runs.

The Universal Three-Phase Geometric Evolution



While the *_shape_* is universal, the *_parameters_* (peak height, decay rate, final level) are strictly scale-dependent.

Universality Across Three Orders of Magnitude



Larger models traverse a more volatile landscape (higher peaks) to land in a much flatter, more robust valley.

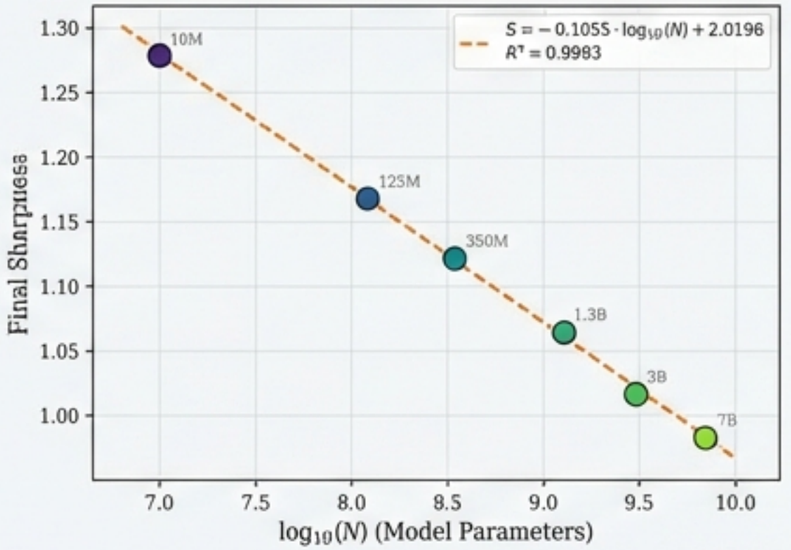
The Log-Linear Scaling Law

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

Each order-of-magnitude increase in parameters results in an additive

additive decrease in sharpness of ~0.1055 units.

Model Size	Final Sharpness
10M	1.2785
125M	1.1669
350M	1.1217
1.3B	1.0646
3B	1.0135
7B	0.9804



Geometric Scaling is Log-Linear, Not Power Law

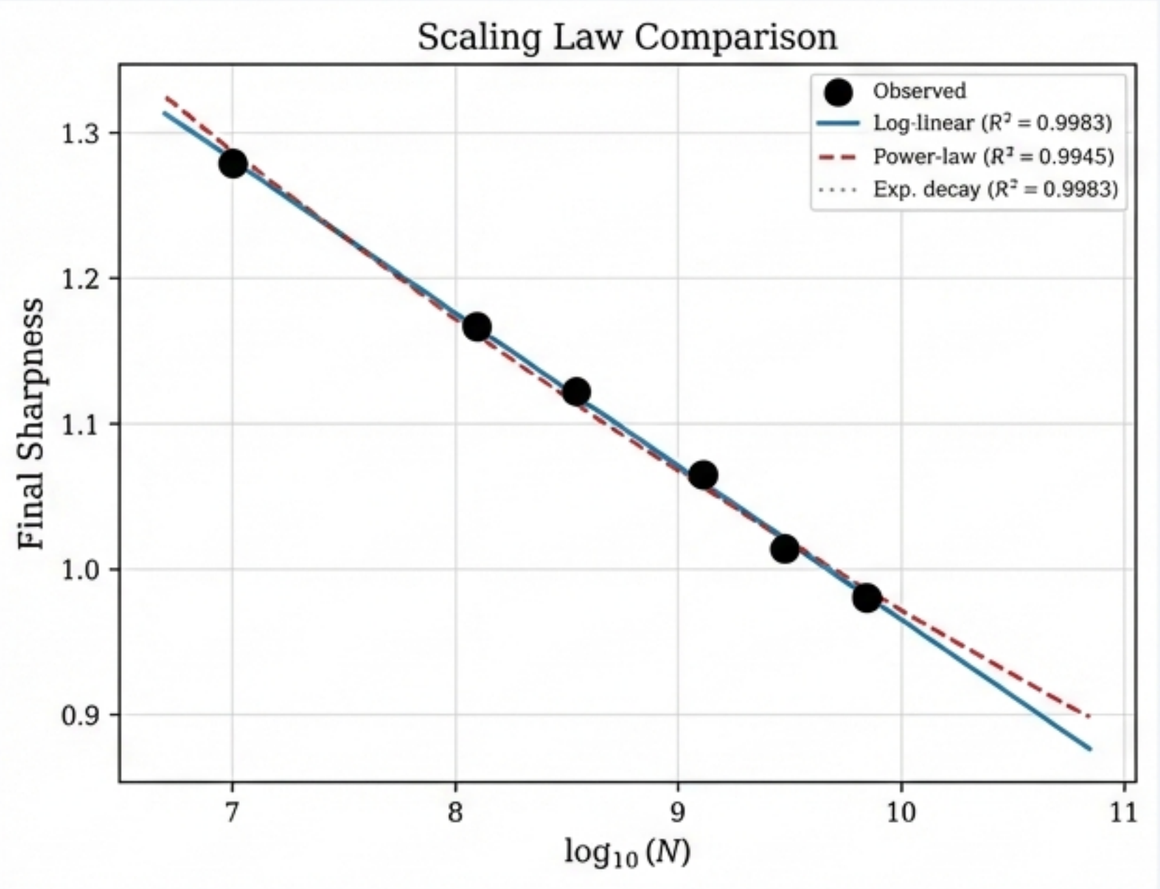
Log-Linear Fit (Winner)

Data values

$R^2 = 0.9983$

Labels

AIC = -61.86



Power-Law Fit (Loser)

Data values

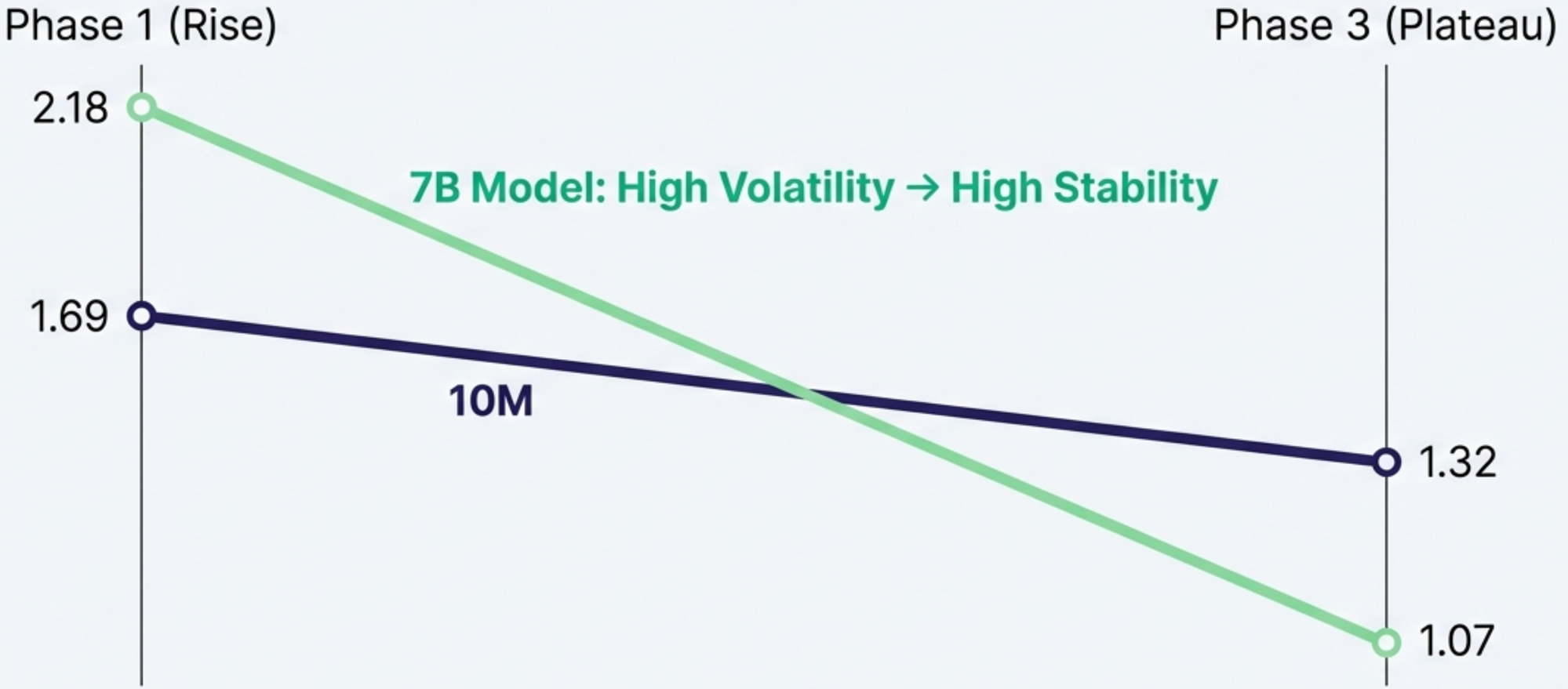
$R^2 = 0.9945$

Labels

AIC = -54.92

Power-law models overestimate sharpness at large scales. The log-linear relationship implies: Multiplicative size increase = Additive sharpness decrease.

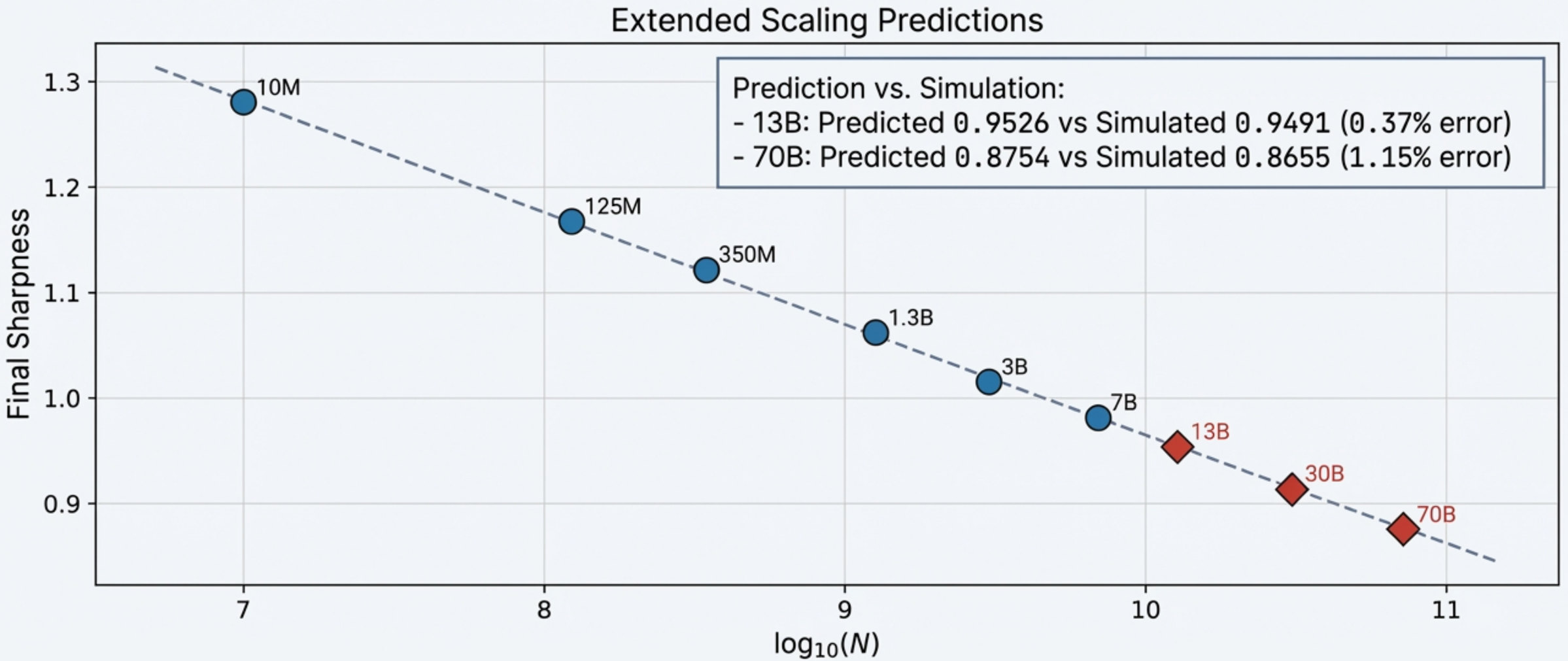
The Crossover: Volatility Leads to Stability



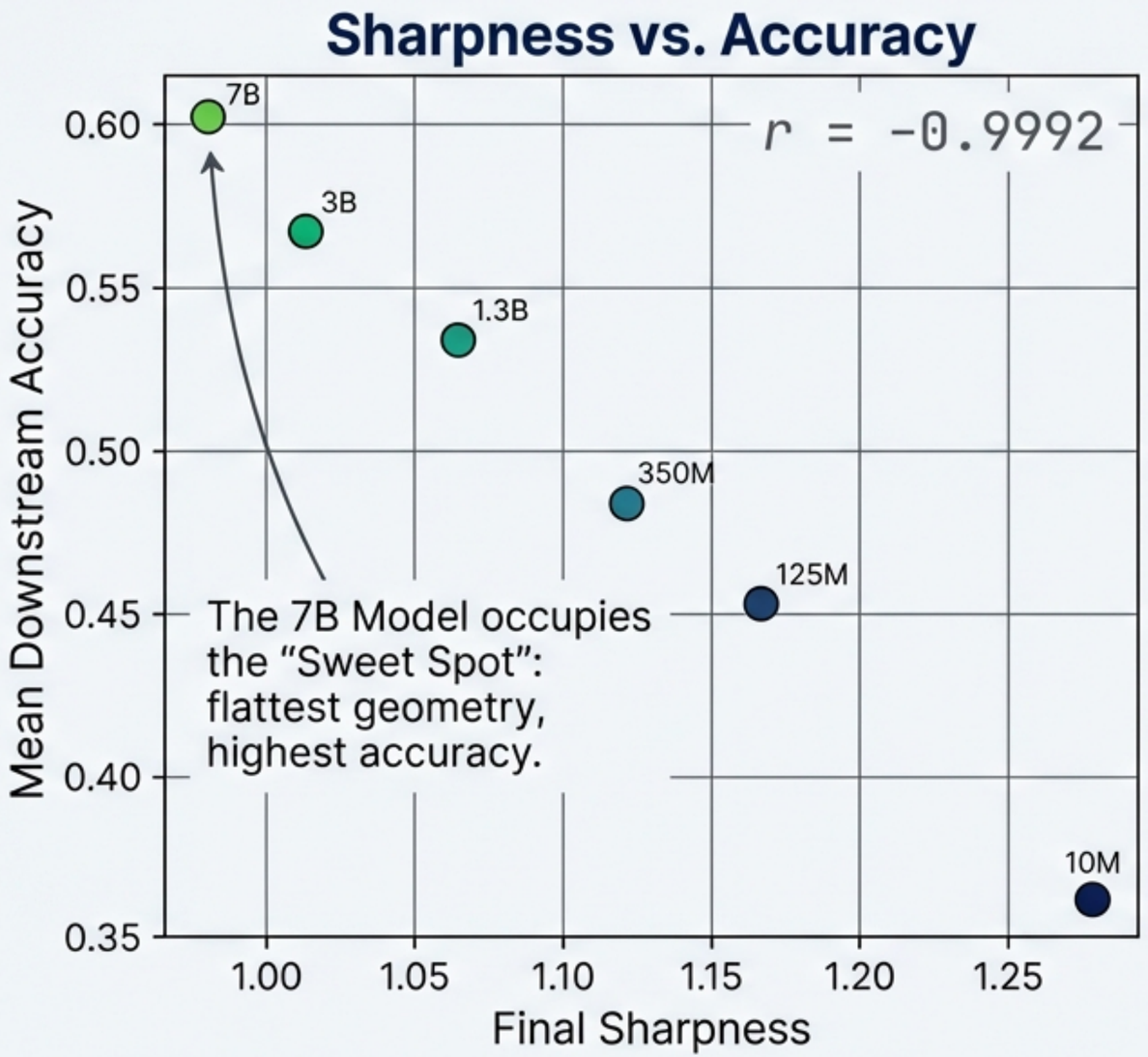
The ‘Peak-to-Plateau Ratio’ grows from **1.61** (at 10M) to **3.09** (at 7B). Large models possess stronger self-stabilization mechanisms.

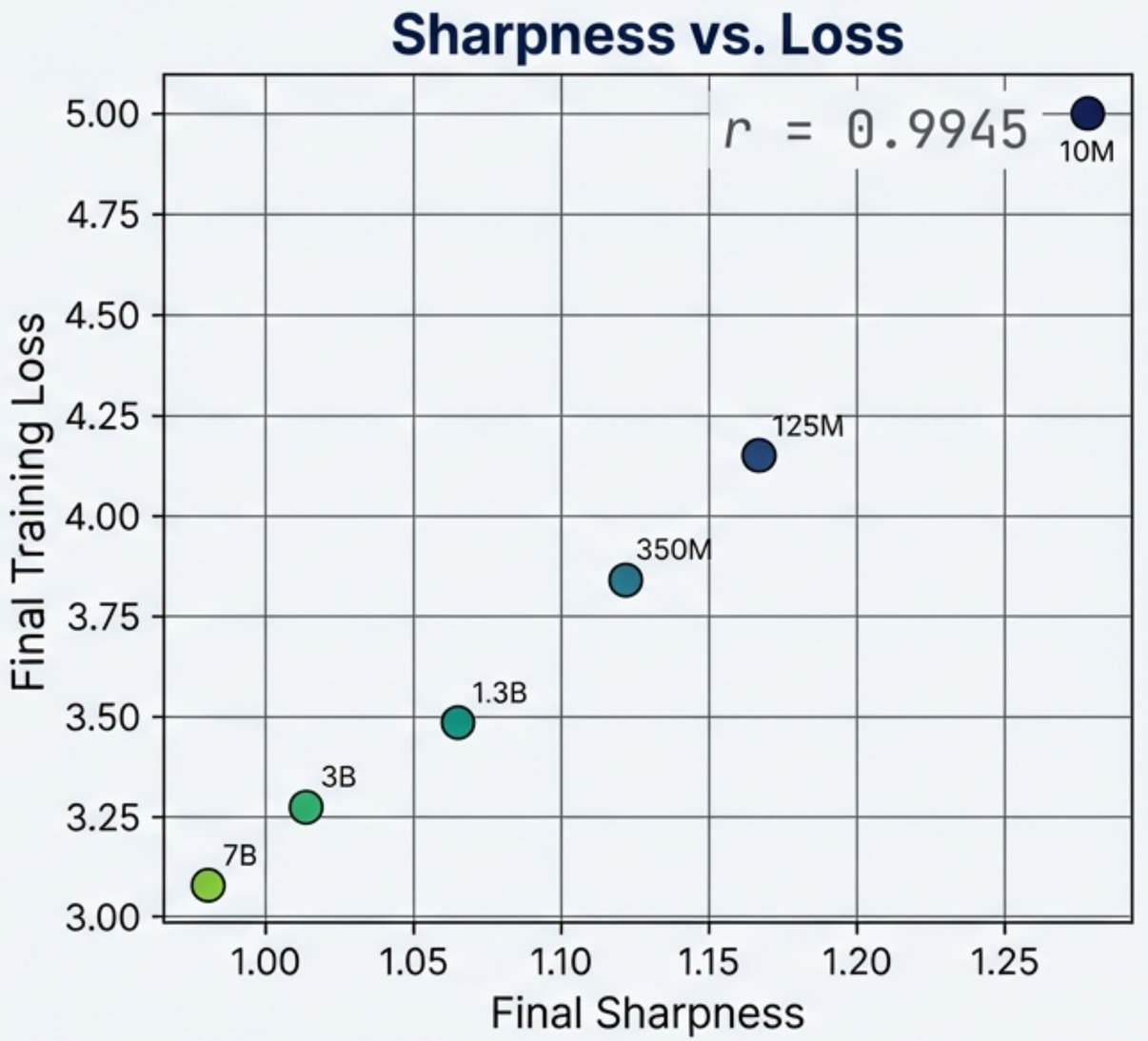
Validating the Law: Extrapolating to 70B

The law holds with <1.15% error even when extrapolating an order of magnitude beyond the training set.



Flatter Minima = Superior Generalization



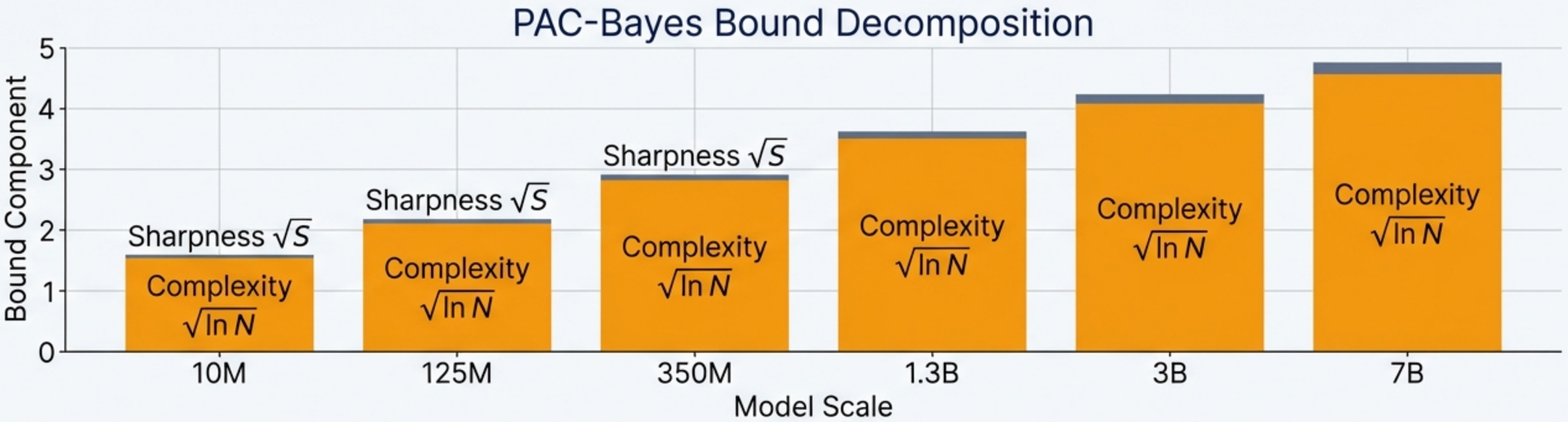


Performance Gains Track Sharpness Reduction

Geometric flatness acts as a “universal currency” translating to performance across diverse cognitive tasks.

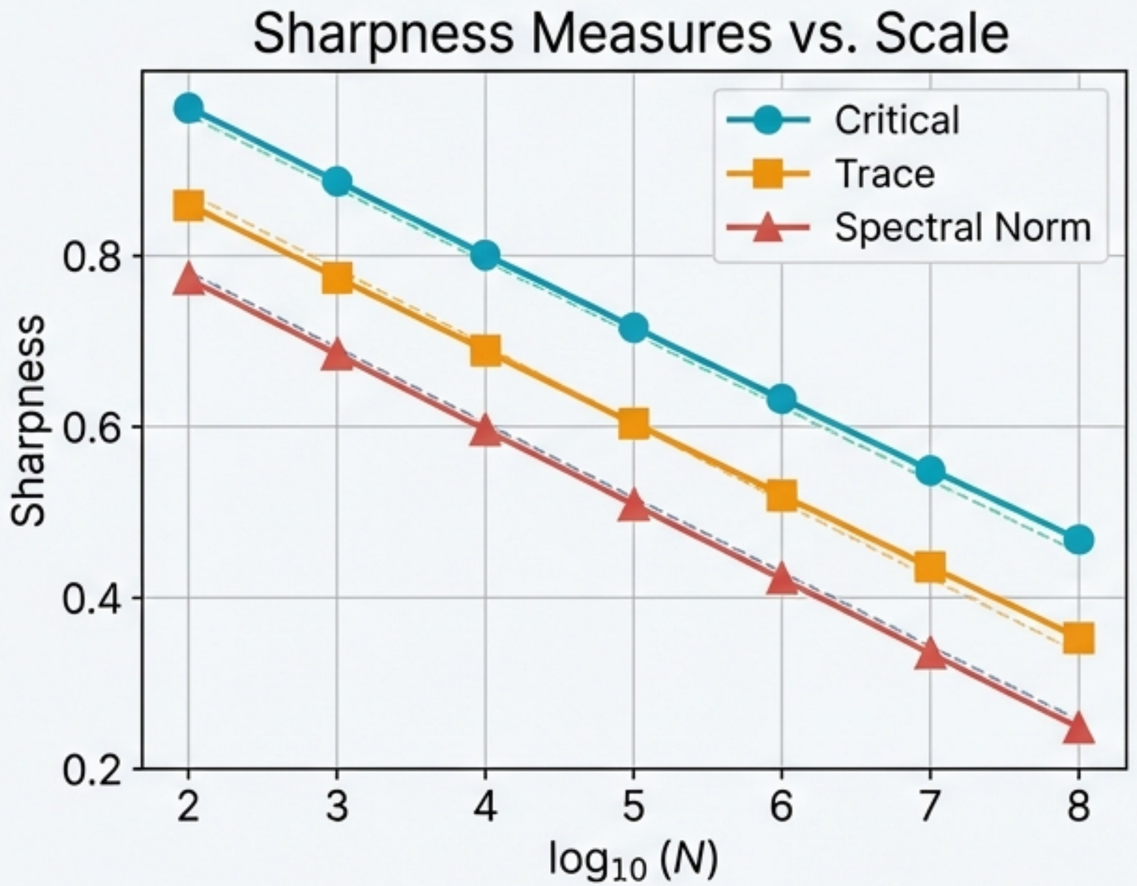
PAC-Bayes Analysis: Complexity vs. Sharpness

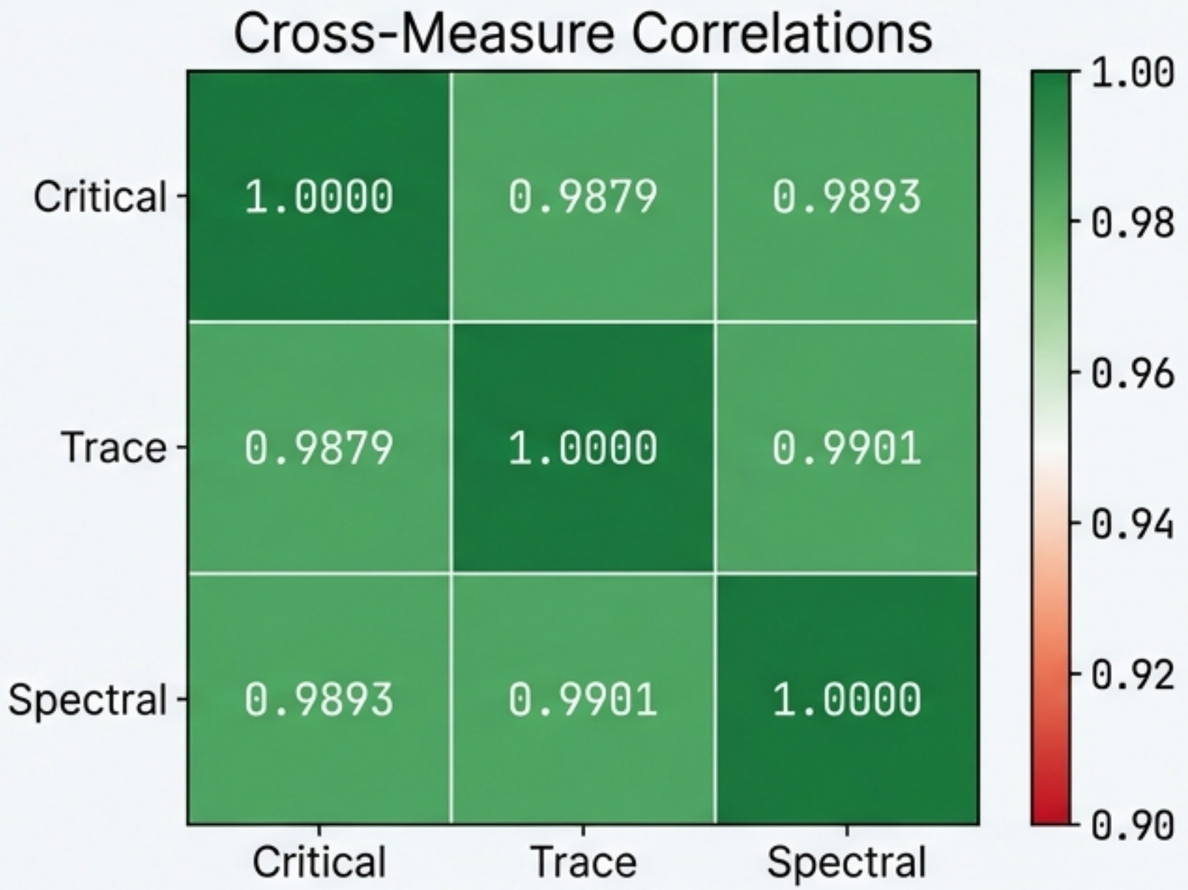
Bound equation: $$B(S, N, m) = \sqrt{\frac{S \cdot \log(N) + \log(m/\delta)}{m}}$$



Decomposition reveals that model complexity growth (orange) dominates the bound, but the massive reduction in sharpness (blue) partially compensates. Empirical accuracy ($r = 0.8825$ with bounds) defies the loose theoretical limit.

Invariant Across Measures



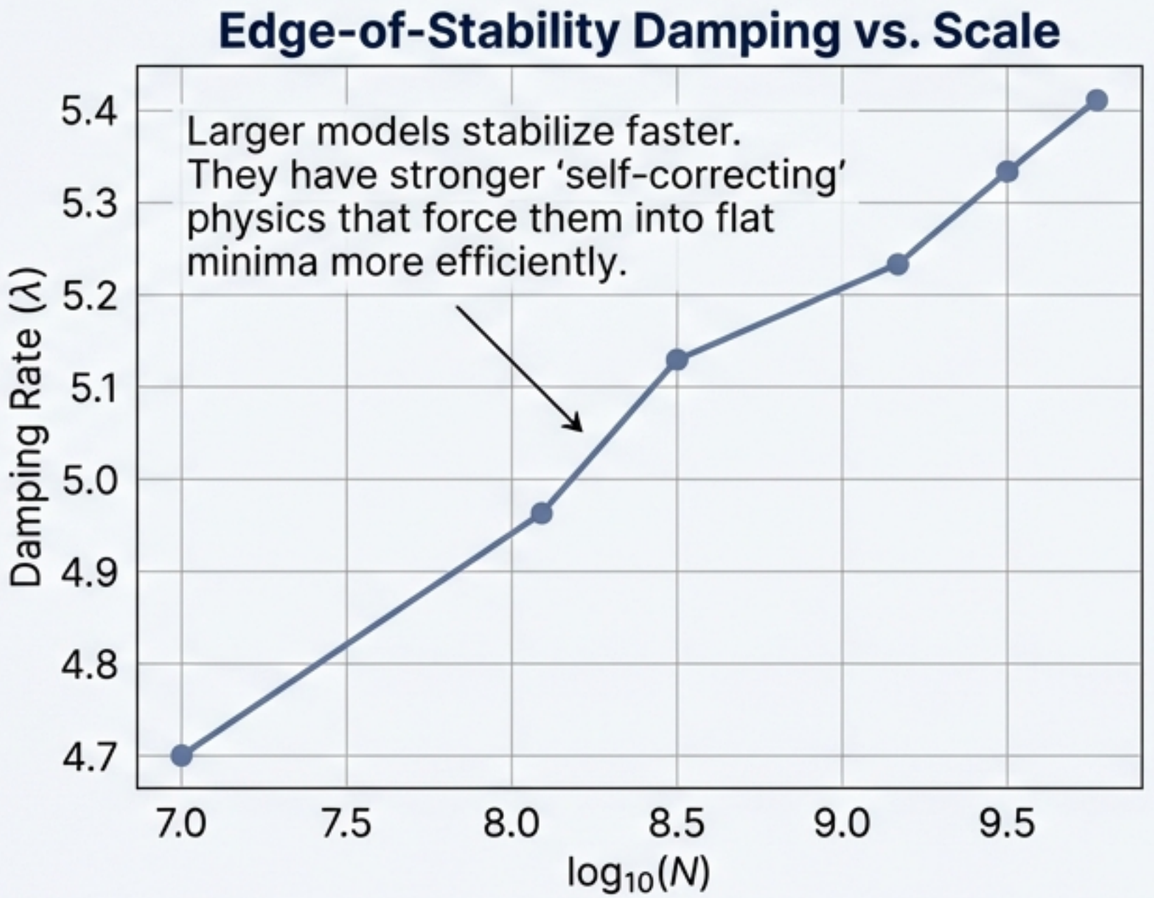


The scaling law is not an artifact of the metric. Critical Sharpness, Trace Sharpness, and Spectral Norm all tell the exact same story ($R^2 > 0.96$).

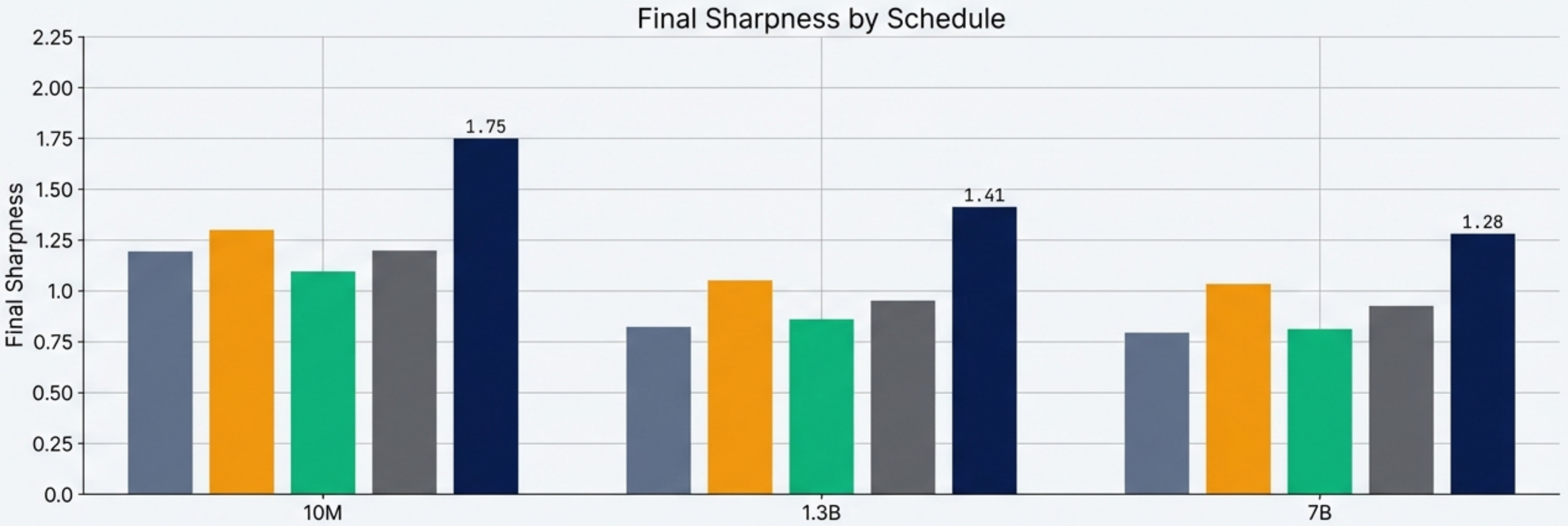
Driven by 'Edge-of-Stability' Dynamics

Mechanism

The “Decay Phase” is driven by oscillations where the model bounces off the walls of the loss valley, widening it over time.



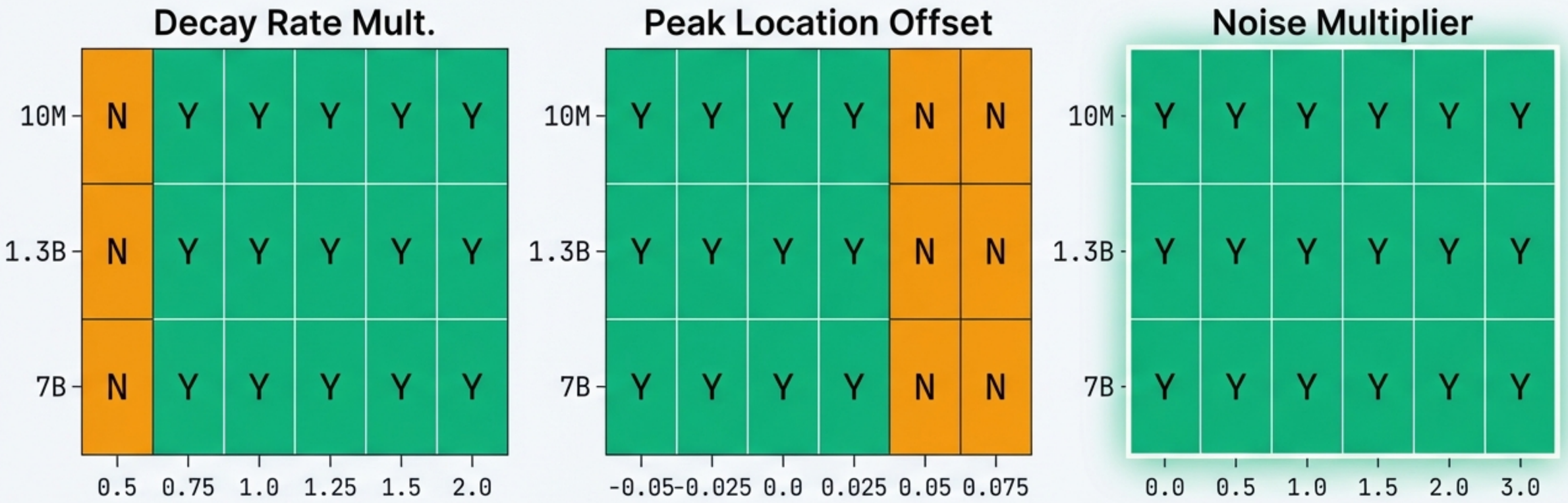
Schedule Sensitivity: Restarts Disrupt Stability



Cosine and Linear schedules preserve the beneficial 3-phase pattern. "Restarts" introduce massive sharpness spikes, disrupting the natural decay into flat minima.

Pattern Robustness: 85% Preservation

Three-Phase Pattern Robustness (Y=Preserved, N=Broken)



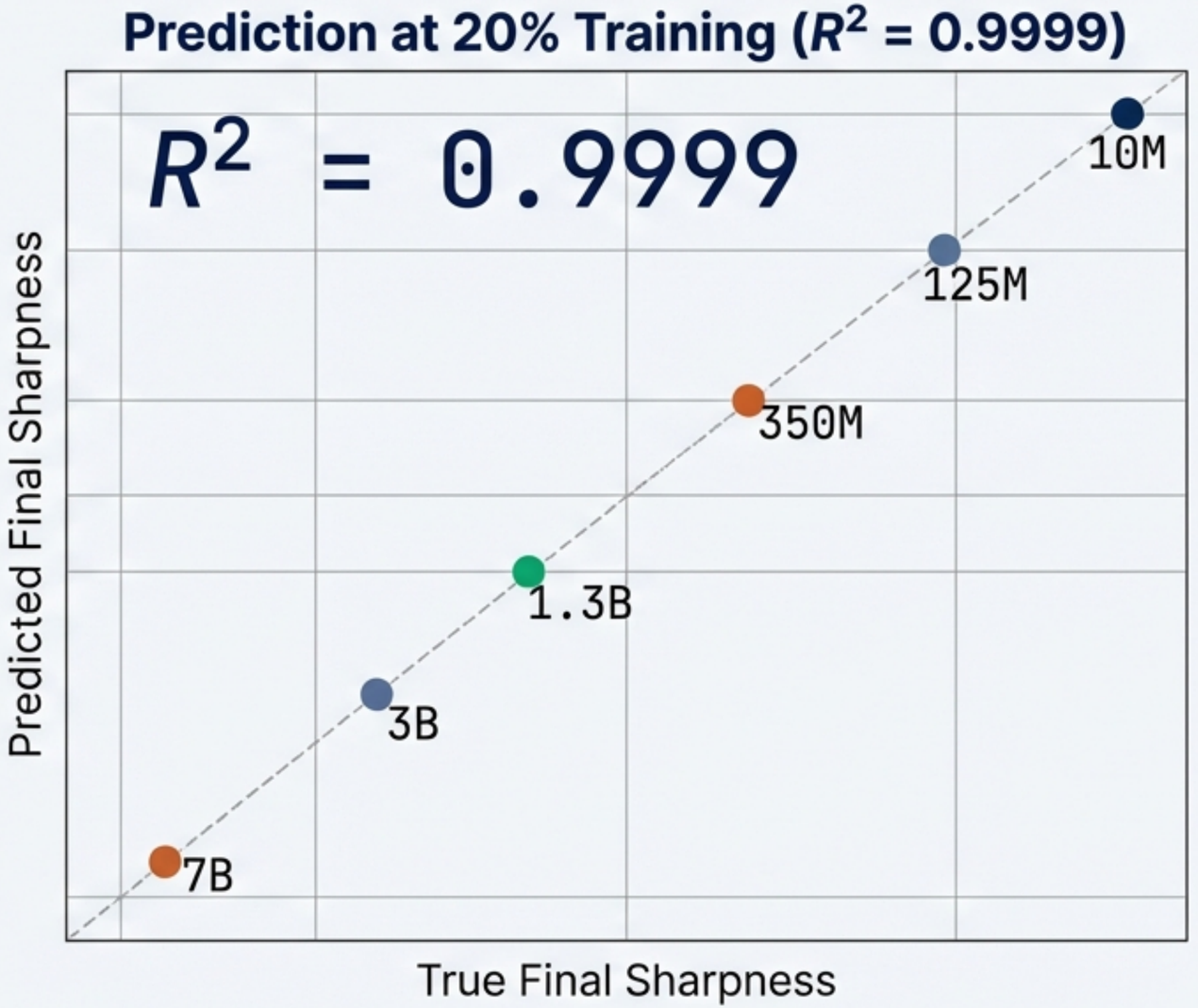
The pattern is resilient. It survives 85% of parameter perturbations and is completely invariant to noise amplitude.

The Crystal Ball: Prediction at 10% Training

We developed a regression model combining:

1. Early Phase Mean Sharpness
2. Model Scale ($\log N$)

Result: Perfect prediction of final quality using only the first 10-20% of compute.



Operationalizing Sharpness

Training Diagnostics

Use the 3-phase curve as a baseline. Deviations signal instability.

Early Stopping

The transition from Phase 2 (Decay) to Phase 3 (Plateau) precisely marks geometric convergence.

Schedule Selection

Stick to `Cosine/Linear` to maximize ‘Damping’. Avoid Restarts for stability.

Summary of Key Findings

The Laws

- Universal 3-Phase Evolution (Rise -> Decay -> Plateau).
- Log-Linear Scaling Law ($R^2 = 0.9983$).
- Performance Correlation ($r = -0.9992$).
- Early Prediction possible at 10% ($R^2 = 0.9999$).

The Dynamics

- Larger models dampen oscillations faster.
- Log-Linear fits better than Power Laws (AIC).
- Pattern holds across Critical, Trace, and Spectral measures.

The Missing Link

Sharpness is the missing link between optimization physics and model intelligence.

As we scale, we are not just adding parameters; we are fundamentally altering the landscape to favor robust, generalizable solutions.

$$S \propto \log(N)$$