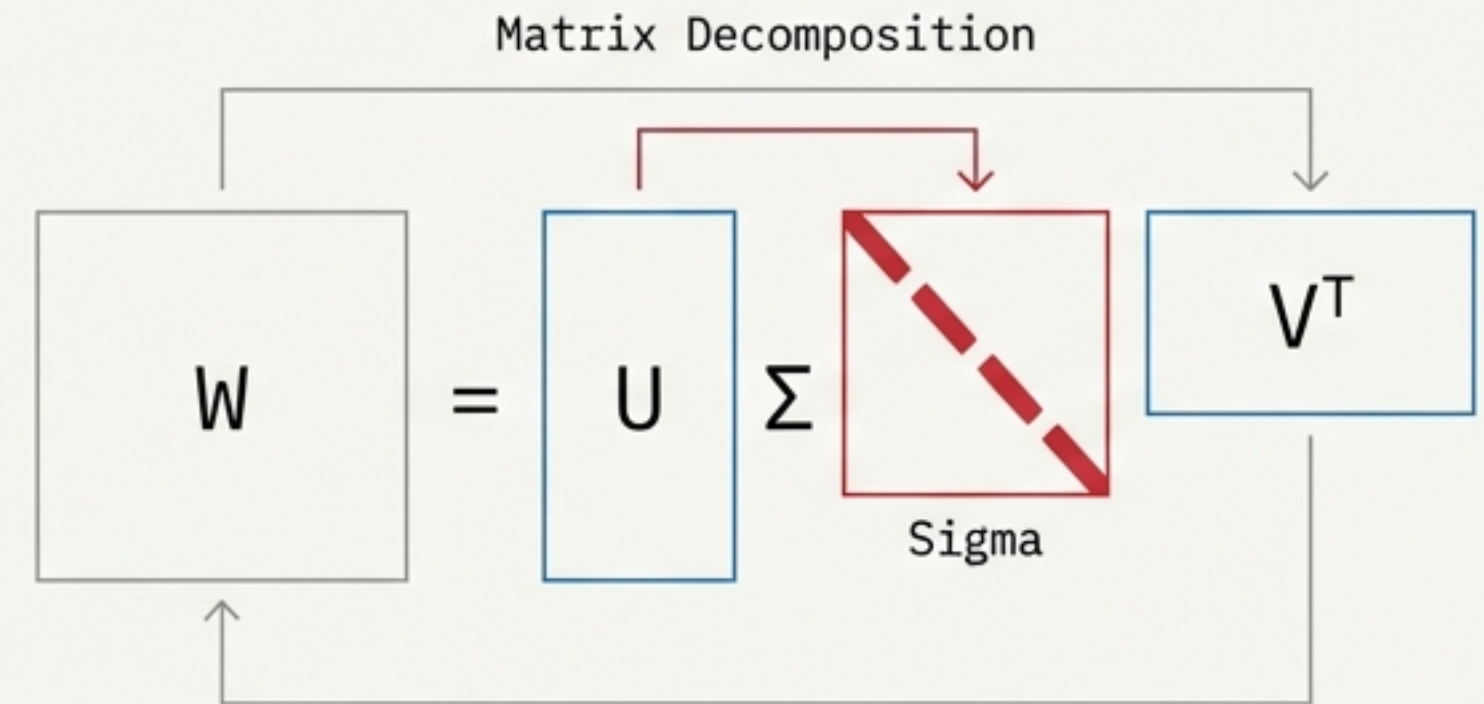


# The Unlearned Flaw: Identifying & Correcting Spectral Optimization Errors

Why Neural Networks Fail to Learn Condition Numbers and How to Fix It.



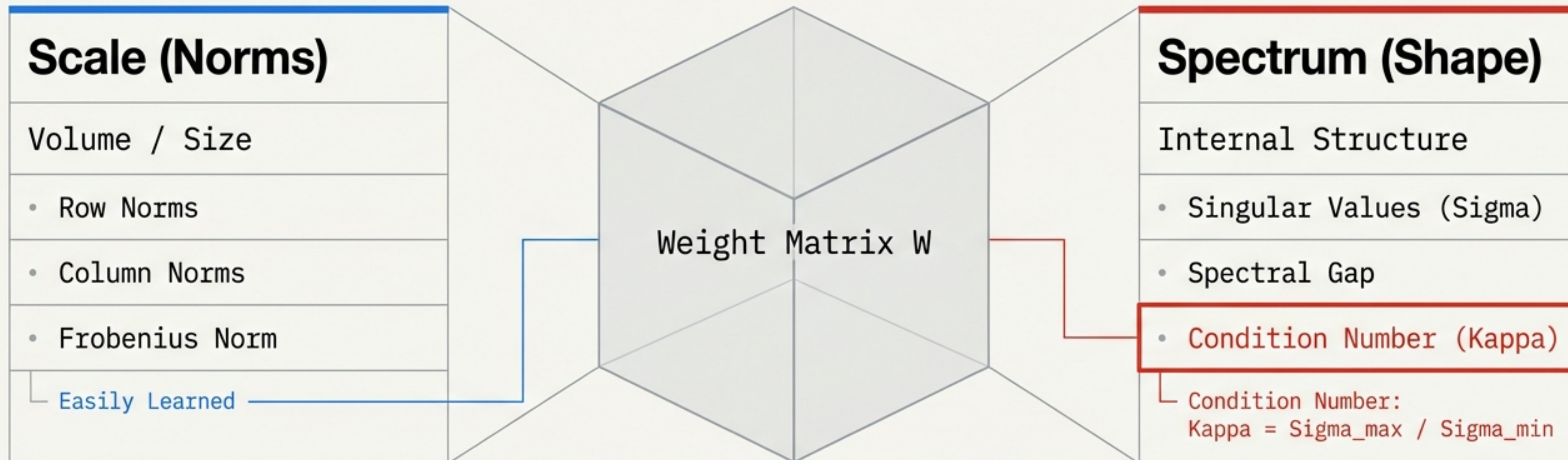


# Executive Summary: The ‘Silent Killer’ in Optimization

The Vitals	The Diagnosis - Visual
<ul style="list-style-type: none"><li>- <b>The Symptom:</b> Optimizers learn Scale (Norms) well (~0.13 error) but fail to learn Shape (Condition Number).</li><li>- <b>The Severity:</b> Errors amplify exponentially. A 3-layer linear net sees 13,167x error amplification.</li><li>- <b>The False Cure:</b> Adam performs worse than SGD (0.60 vs 0.14 overall error). Weight decay does not help.</li><li>- <b>The Root Cause:</b> Gradient imbalance. The optimizer is “deaf” to the smallest singular values (10-100x less signal).</li><li>- <b>The Solution:</b> Hybrid Regularization. Combine Learnable Multipliers (for Scale) + Spectral Regularization (for Shape).</li></ul>	<pre>graph LR; A[Weight Matrices (Trained via Gradient Descent)] --&gt; B[Well-learned (error &lt; 0.05)]; A --&gt; C[Moderately learned (error ~0.13)]; A --&gt; D[Poorly learned (error grows with dimension)]; B --&gt; E[Depth Amplification (Critical Finding)]; C --&gt; E; D --&gt; E; E --&gt; F[Root Cause - Gradient Imbalance]; E --&gt; G[Norm-Spectral Trade-off]; E --&gt; H[Extended Training Paradox]; F --&gt; I[Optimizer Comparison]; G --&gt; H; H --&gt; I;</pre> <p><b>Weight Matrices</b> (Trained via Gradient Descent)</p> <ul style="list-style-type: none"><li><b>Well-learned (error &lt; 0.05)</b><ul style="list-style-type: none"><li>• Details - error &lt; 0.05</li><li>• Condition number: error &lt; 0.05</li><li>• Weight matrix error &gt; 0.25</li></ul></li><li><b>Moderately learned (error ~0.13)</b><ul style="list-style-type: none"><li>• Details - error ~ 0.13</li><li>• Condition number: error ~0.13</li></ul></li><li><b>Poorly learned (error grows with dimension)</b><ul style="list-style-type: none"><li>• Details - error ~ 0.27</li><li>• Condition number: 0.27 (d=32) -&gt; 1.59 (d=128) -&gt; 34.9 (d=512)</li></ul></li></ul> <p><b>Depth Amplification (Critical Finding)</b> <b>13,167x error</b></p> <ul style="list-style-type: none"><li>• Single matrix: condition number error = 0.24x</li><li>• 2-layer deep linear: error = 653x (2,721x amplification)</li><li>• 3-layer deep linear: error = 13,167x (54,863x amplification)</li></ul> <p><b>Root Cause - Gradient Imbalance</b></p> <p>The power law of gradient magnitude is <math>\sigma = 10^{-\sigma^2}</math> is signed by optimizer error, the behavior in critical scenarios changes.</p> <ul style="list-style-type: none"><li>• The “toni law” shape of gradient magnitude is magnitude as 10-100x less signal).</li><li>• Optimizer behavior the optimization consists of all automates in manner and smallest singular values after dimension, another optimizer envelop.</li></ul> <p><b>Norm-Spectral Trade-off</b></p> <ul style="list-style-type: none"><li>• Data: error = 0.18</li><li>• Condition error = 0.24x - condition number = 0.24</li><li>• Insights: Roumless learn avoidance of the learning optimization-shuse without norm-spectral ranods.</li></ul> <p><b>Extended Training Paradox</b></p> <ul style="list-style-type: none"><li>• Data: matrix = 0.02x</li><li>• 2-layer error = 653x</li><li>• 3-layer matrix = 0,877x</li><li>• Insights: Cohievenhe gradient magnitude are used in drop onwmetad notumas training paradox</li></ul> <p><b>Optimizer Comparison</b></p> <ul style="list-style-type: none"><li>• Data: Adam = 0.14</li><li>• Condition nonor = 0.14 - mpolown error = 0.20</li><li>• Insights: Optimizer piactics eved decreased commonsen: fisms. values and ingredients.</li></ul>



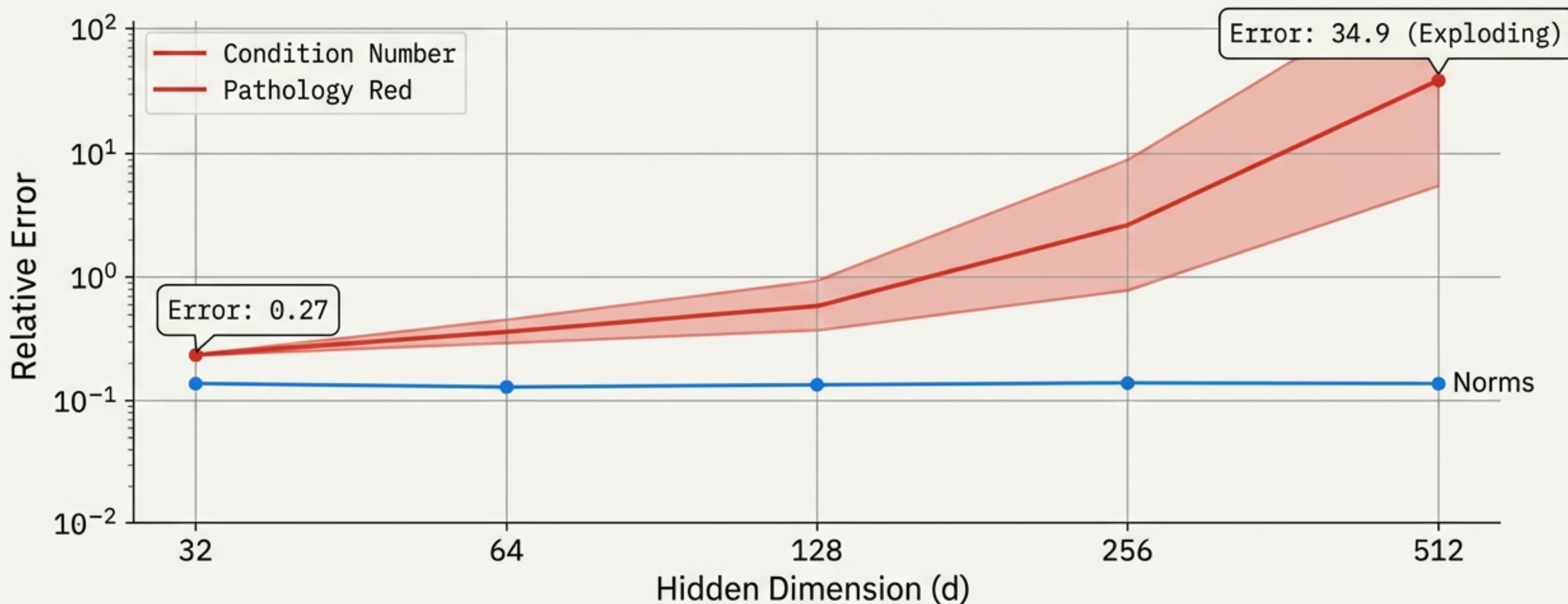
# Anatomy of a Weight Matrix



The Research Question: We know **Gradient Descent struggles with Scale**.  
Does it get the Internal Structure right?



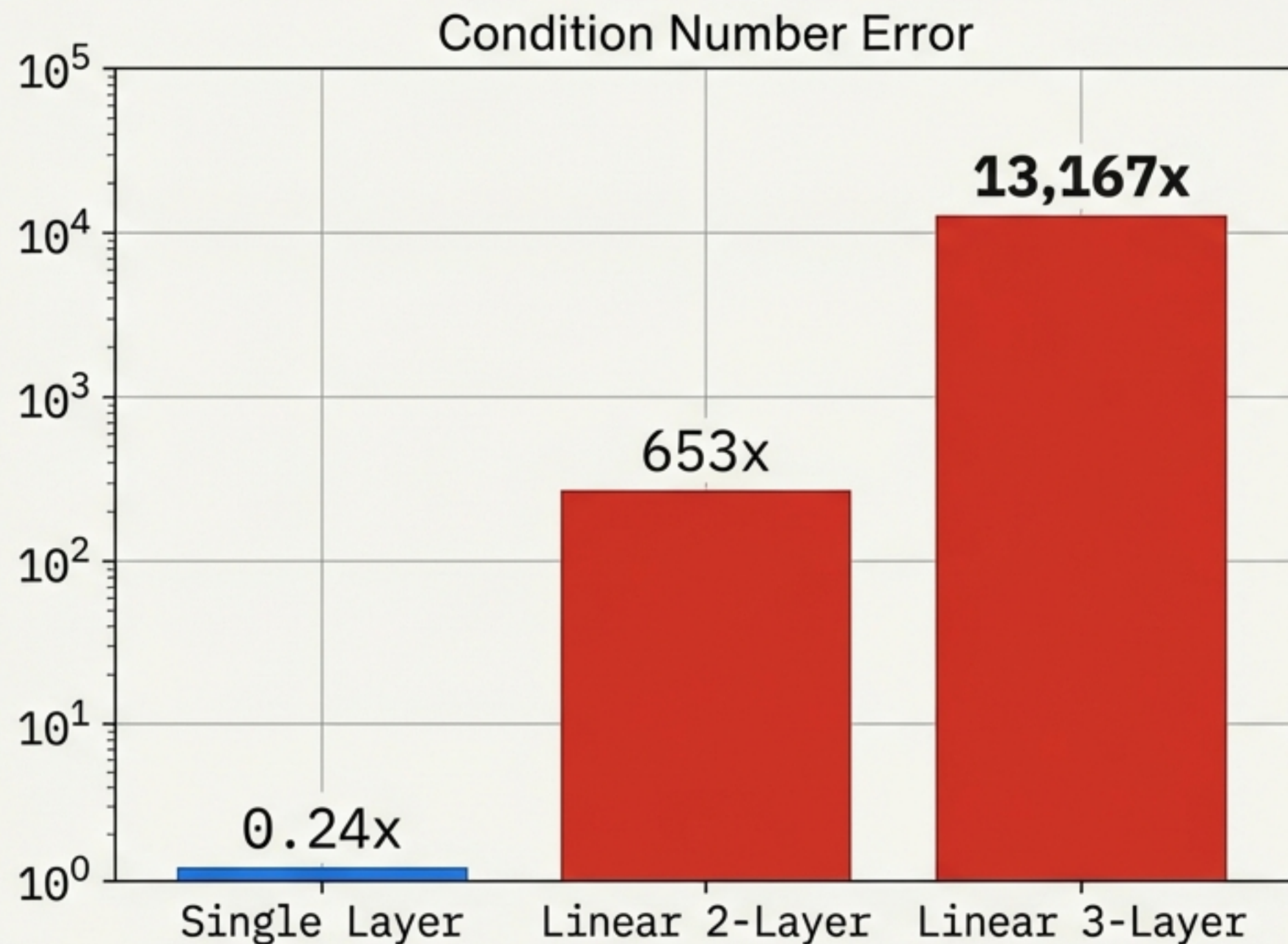
# The Diagnosis: Condition Number is Broken



While Norm errors stay low ( $\sim 0.13$ ), Condition Number errors explode super-linearly with dimension.



# The Progression: Errors Amplify Exponentially with Depth



## The Pathology of Depth:

Errors in individual layers do not add up; they multiply.

1-Layer: 0.24x

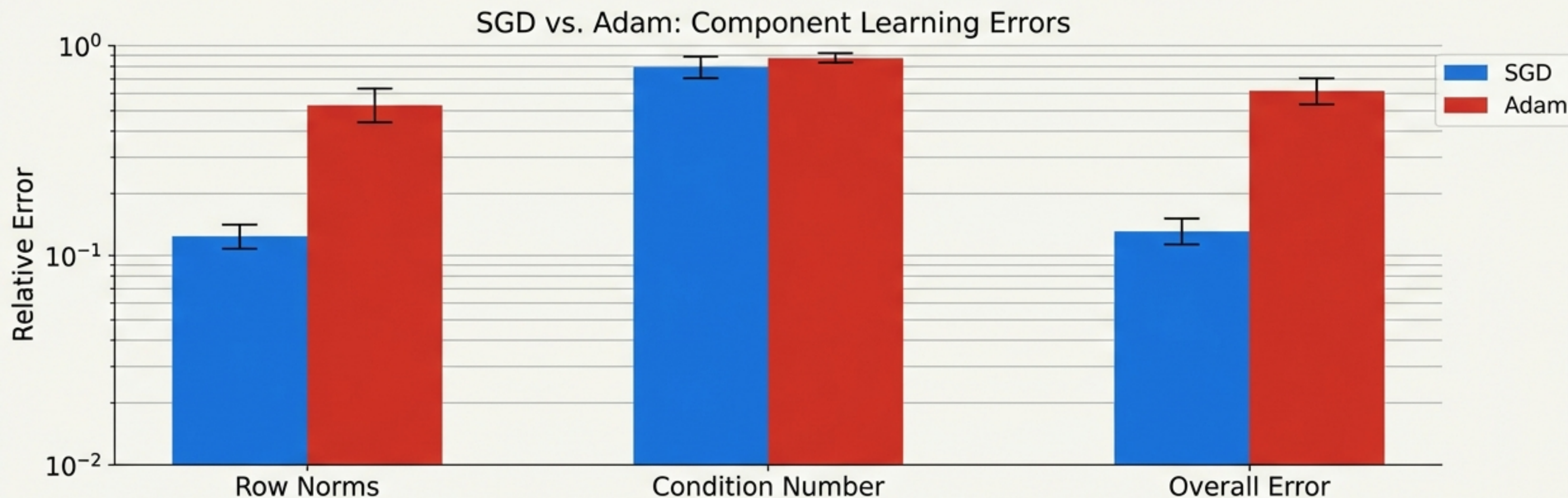
2-Layer: 653x

3-Layer: 13,167x

If a simple 3-layer net has a 13,000x distortion, deep Transformers are likely severely compromised.



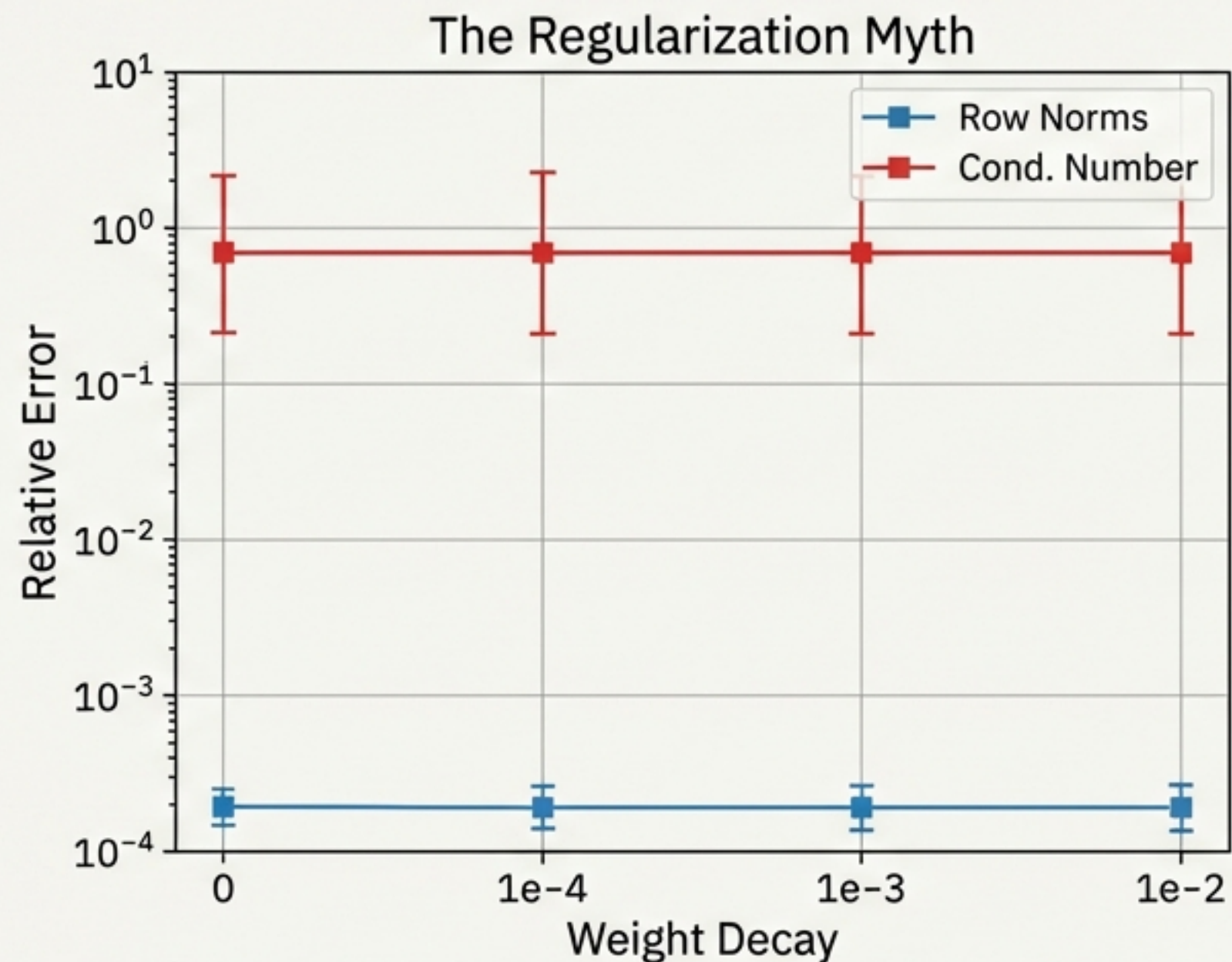
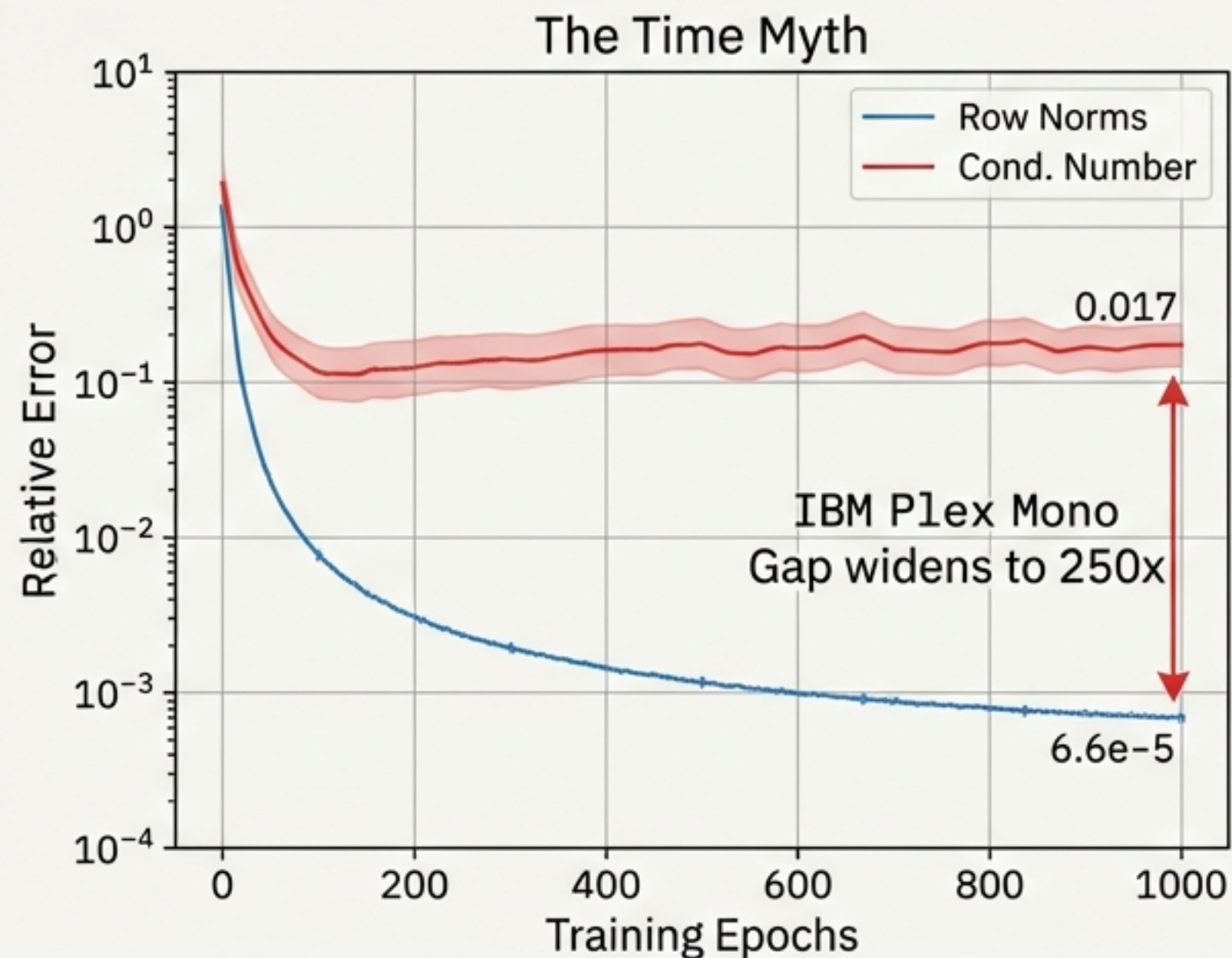
# Standard 'Cures' Fail: The Case Against Adam



- Contrary to popular belief, Adam performs worse than SGD across all structural components.
- Mechanism: Adam's element-wise adaptive rates disrupt the coherent spectral structure.



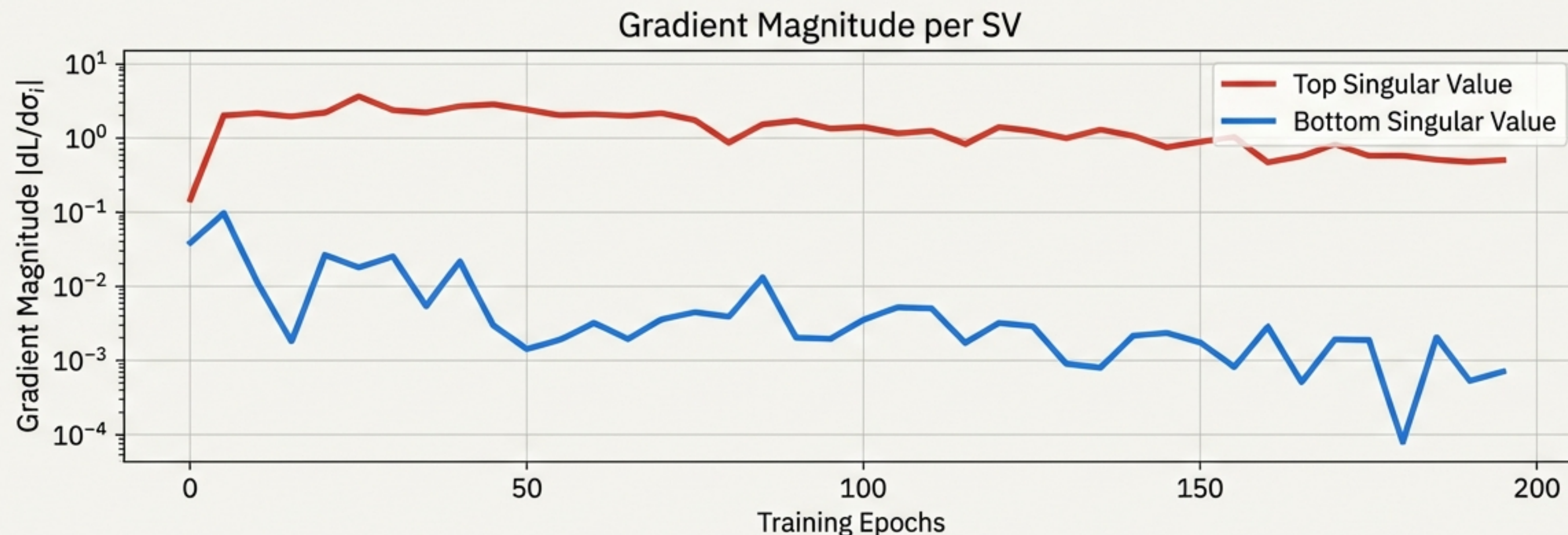
# Chronic Condition: Why Time and Weight Decay Don't Help



Weight Decay has ZERO effect on spectral error.



# The Root Cause: Gradient Signal Imbalance



## Explanation:

The optimizer is 'deaf' to the smallest singular values.

Top singular values receive massive signal. Bottom values receive 10-100x less.

Consequence: Since Condition Number = Top / Bottom, and 'Bottom' never converges, the structure remains broken.



# Clinical Trials: Four Corrective Strategies

<b>Standard SGD</b> The Baseline	<b>Learnable Multipliers</b> The Scale Fix	<b>Spectral Regularization</b> The Shape Fix	<b>SVD Correction</b> The Brute Force
Standard Gradient Descent (LR 0.01)	Per-row/column scaling. $W_{\text{eff}} = \text{diag}(r) W \text{diag}(c)$	Targeted penalty on condition number. $\text{Loss} + \lambda \left( \log \text{Kappa}(W) - \log \text{Kappa}(\text{target}) \right)^2$	Periodic manual adjustment of singular values.



# The Discovery: A Fundamental Norm-Spectral Trade-off

Strategy	Norm Improvement	Condition Number Improvement
Learnable Multipliers	<b>67% (Great)</b>	10% (Fail)
Spectral Regularization	0% (Fail)	<b>77% (Great)</b>
SVD Correction	64%	32%

Distinct flaws require distinct correction mechanisms.  
There is no "Silver Bullet".



# Prognosis for Large Language Models

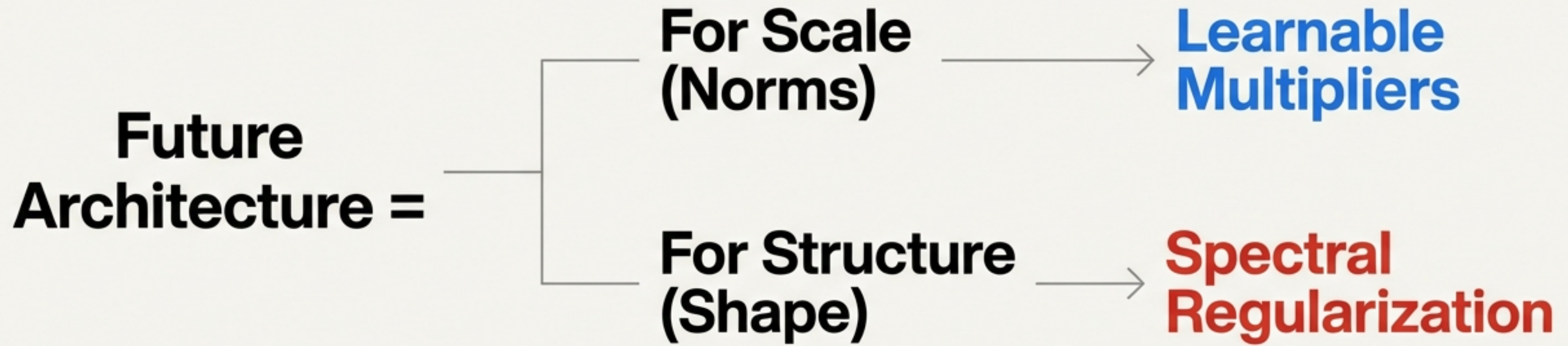


- If a simple 3-layer net distorts structure by 13,000x, deep Transformers are operating with severely distorted spectral vitals.
- Symptoms in the Wild:
  - Training Instabilities
  - Need for Learning Rate Warmup
  - Inexplicable Loss Spikes
- Conclusion: Current LLMs are trained with 'unlearned' internal structures.



# The Prescription: Hybrid Regularization

## Solving the Dual Pathology



To cure the model, we must treat both symptoms: Scale and Shape.