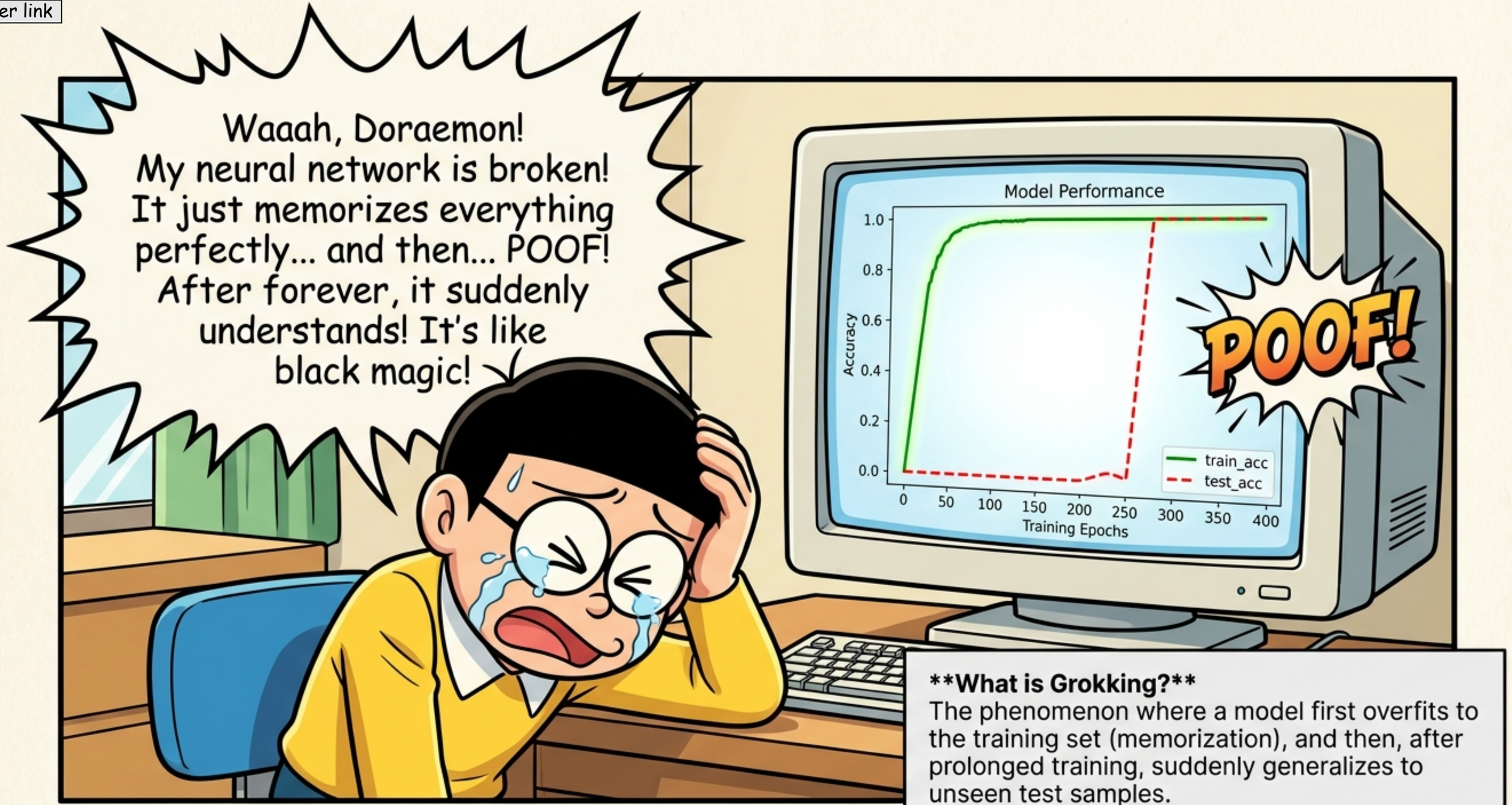
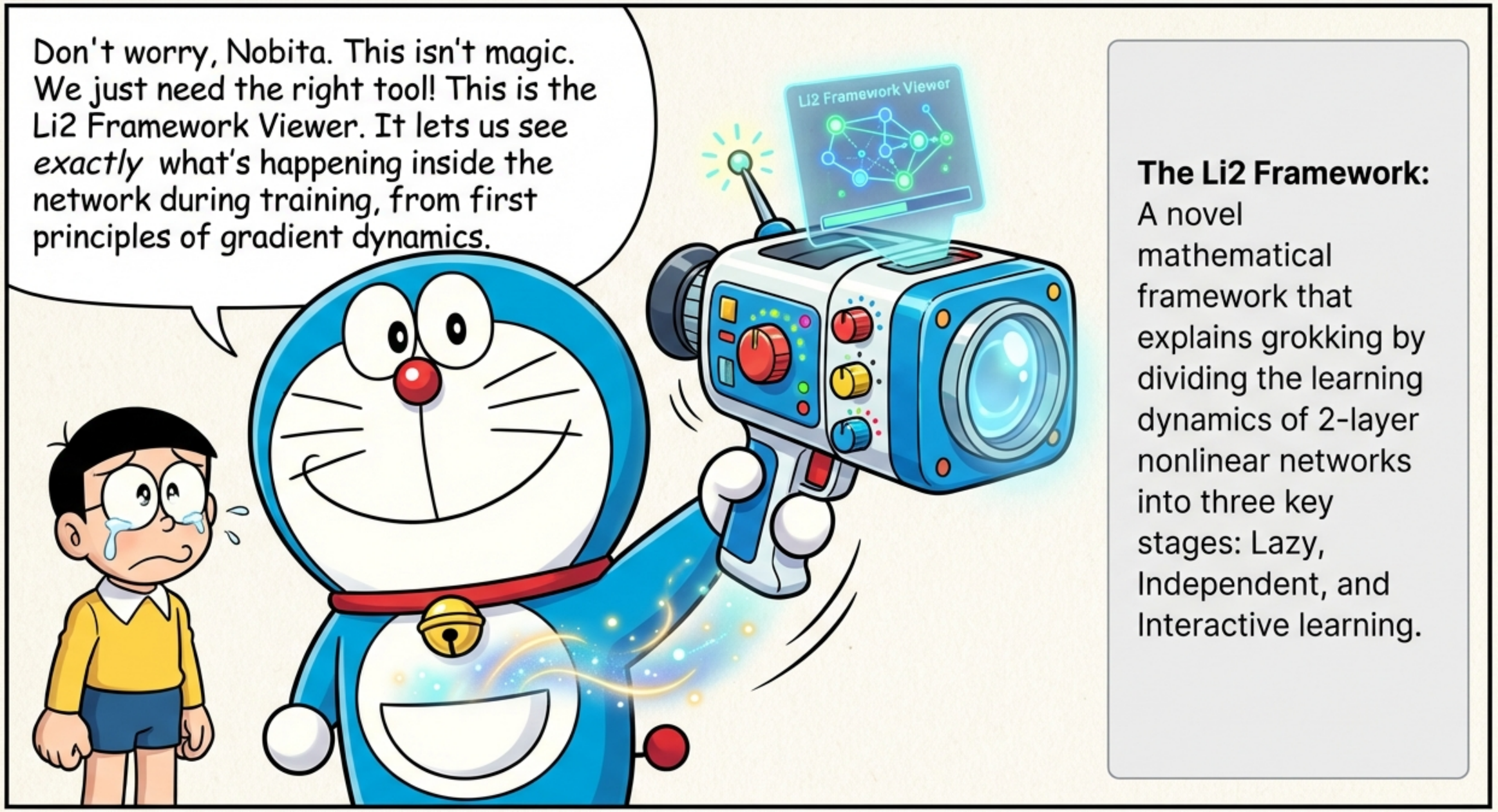
Paper link



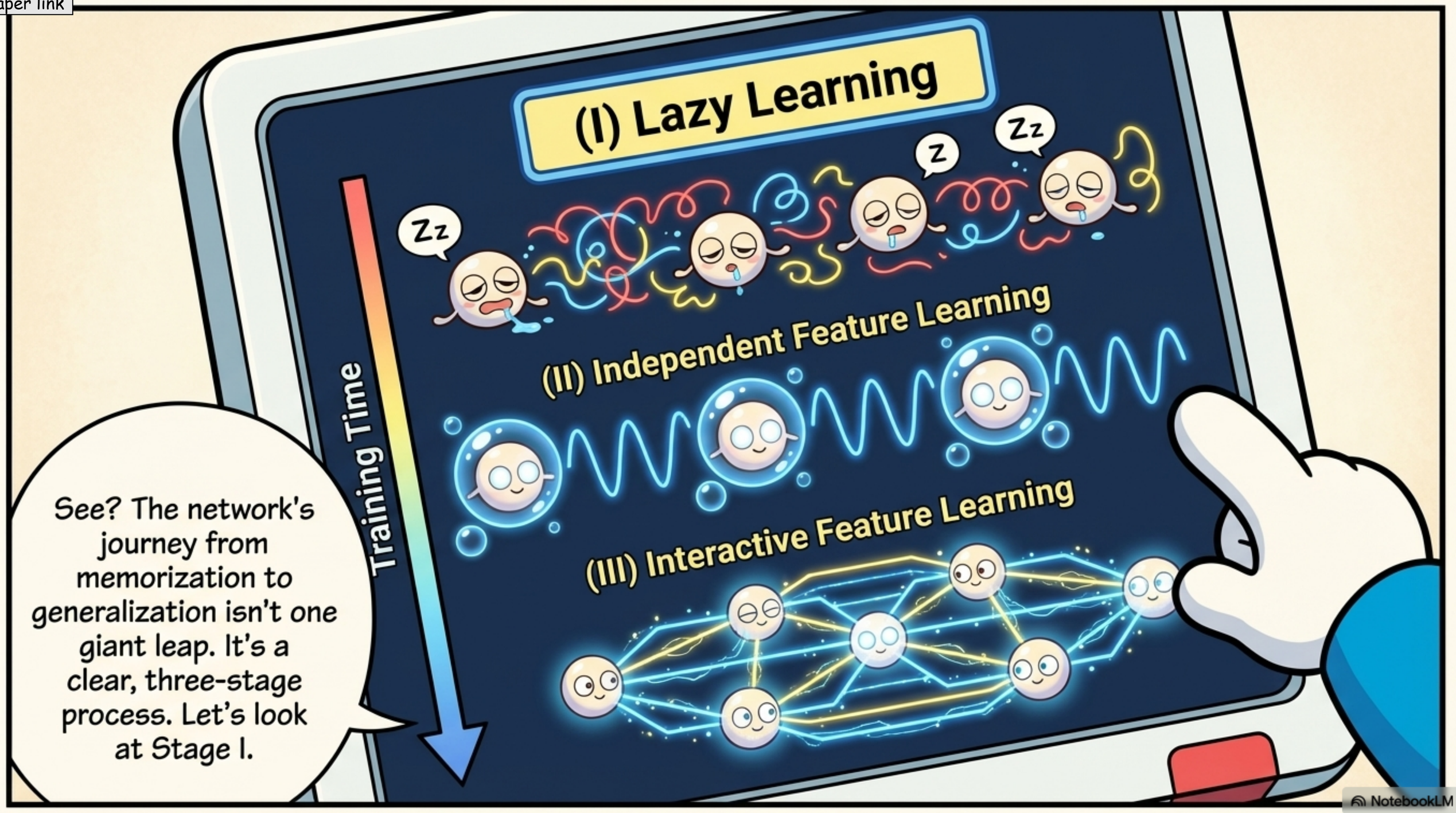
What is Grokking?
The phenomenon where a model first overfits to the training set (memorization), and then, after prolonged training, suddenly generalizes to unseen test samples.
(Source: Power et al., 2022)

Paper link



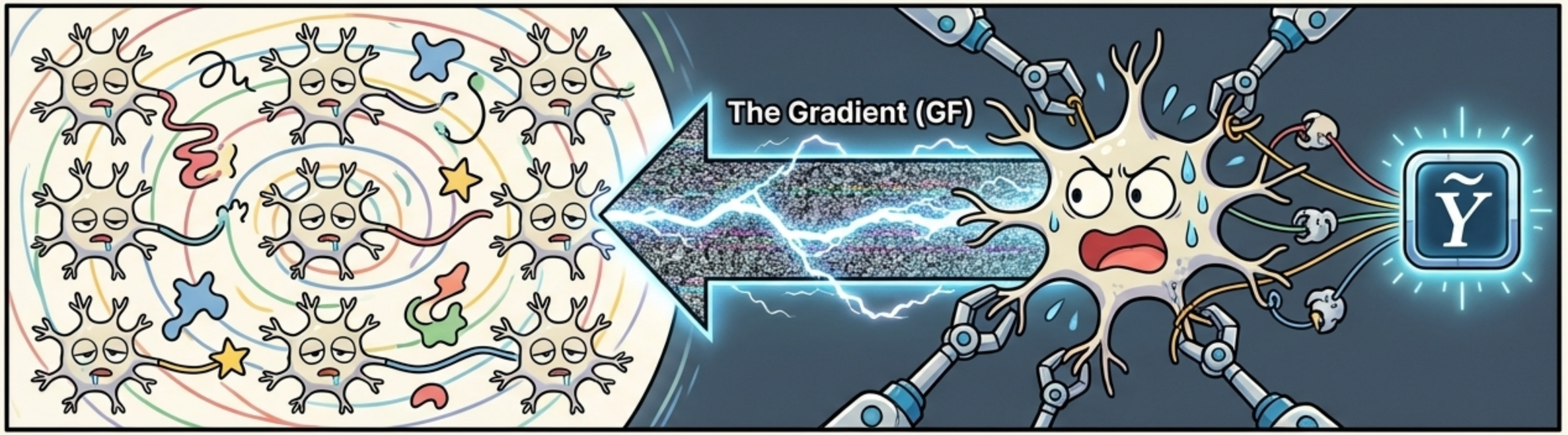
The Li2 Framework: A novel mathematical framework that explains grokking by dividing the learning dynamics of 2-layer nonlinear networks into three key stages: Lazy, Independent, and Interactive learning.

Paper link



Paper link

Stage I: Lazy Learning

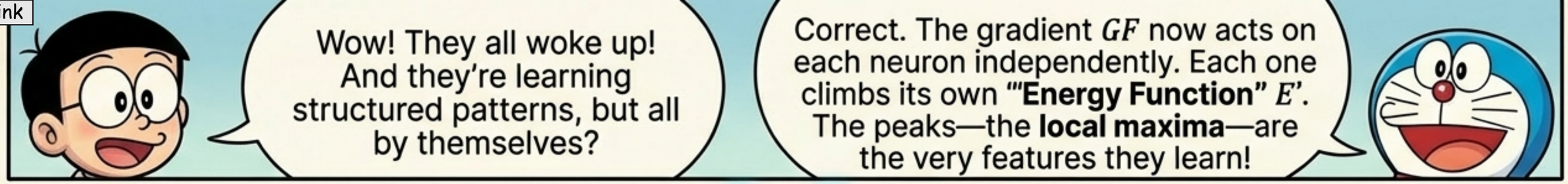


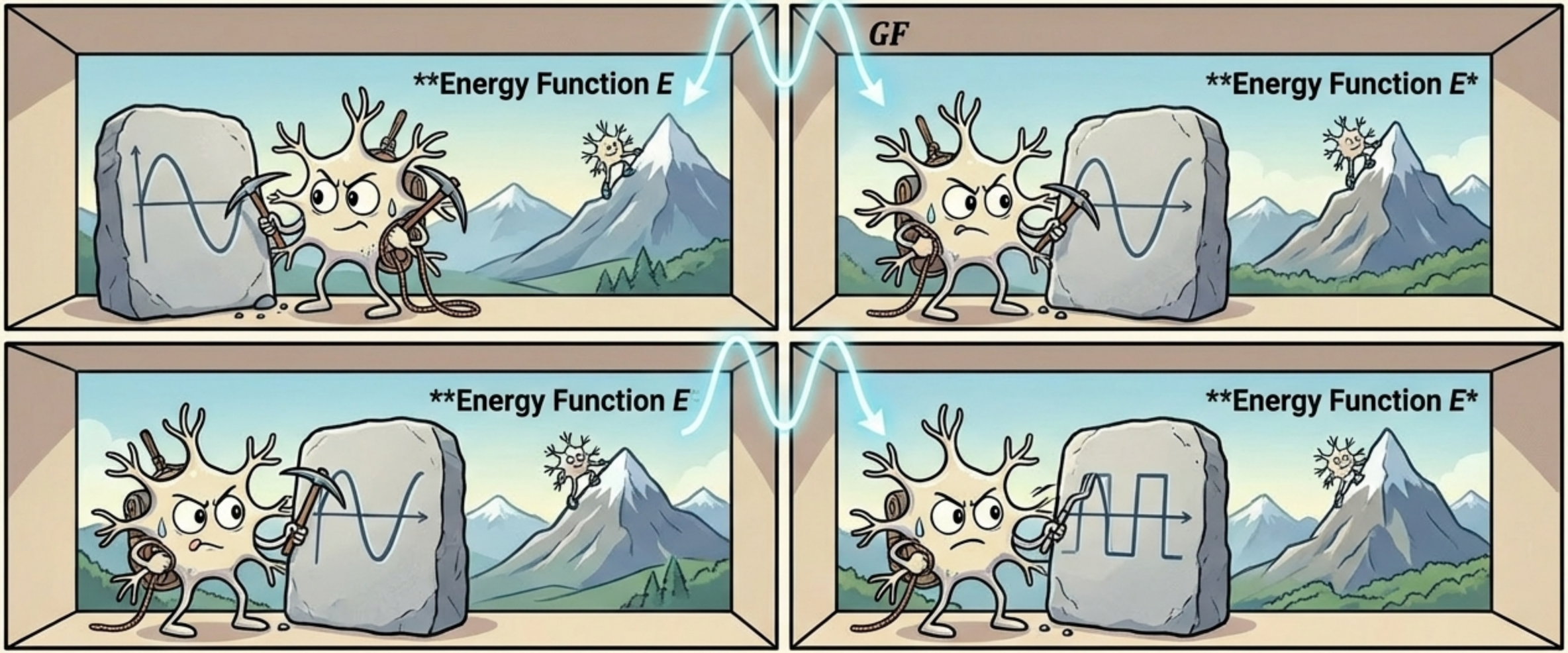
Technical Caption: The backpropagated gradient $GF = P^{\perp 1}(Y - FV)V^T$ is initially noise. Without weight decay ($\eta = 0$), GF' eventually becomes zero, and feature learning stops entirely.

Paper link

**Technical Caption*: With $\eta > 0$, the backpropagated gradient becomes $GF \propto \eta \tilde{Y}\tilde{Y}^T F$. This signal now carries information about the target label, triggering the start of feature learning in the hidden layer. (Source: Lemma 1, Eqn. 6)

Paper link





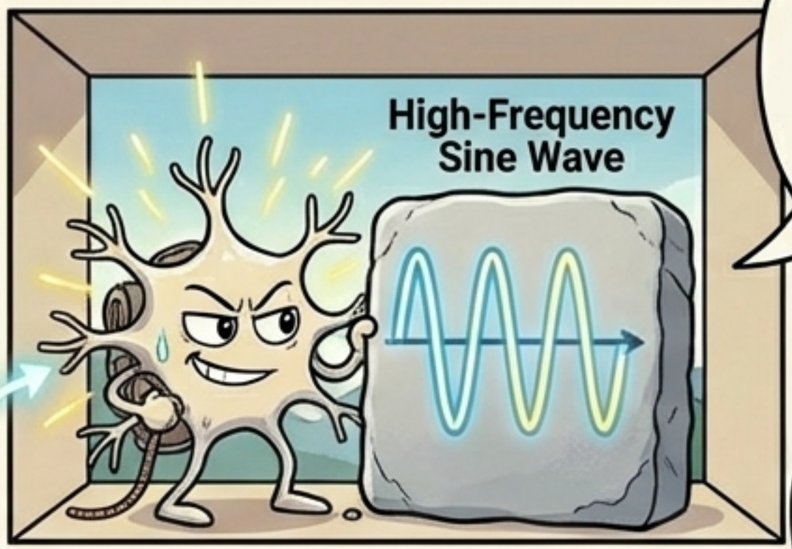
Technical Caption* : The dynamics of each hidden node w_j is exactly the gradient ascent of an energy function: $E(w_j) = \frac{1}{2}\|\tilde{Y}^T\sigma(Xw_j)\|^2$. The emerging features are the local maxima of E. (Source: Theorem 1)

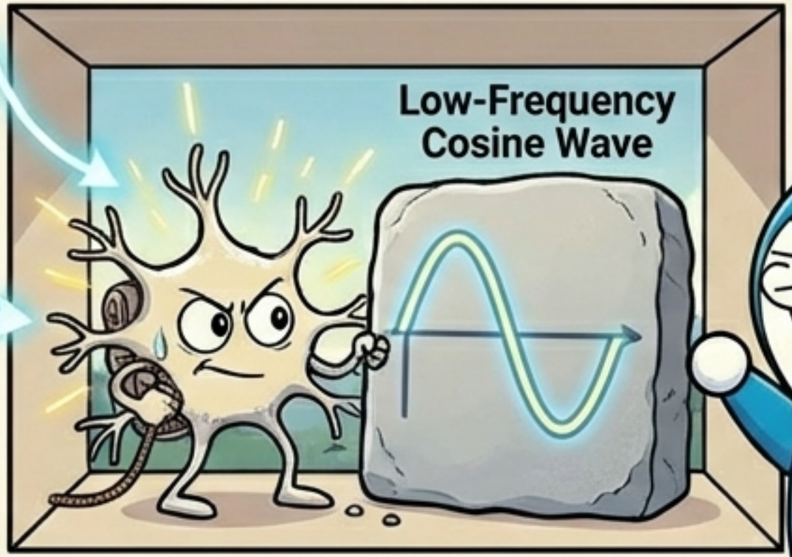
Paper link

Example: Modular Addition

$8 + 5 = 13 \equiv 1 \pmod{12}$

Learned Features: Fourier Bases



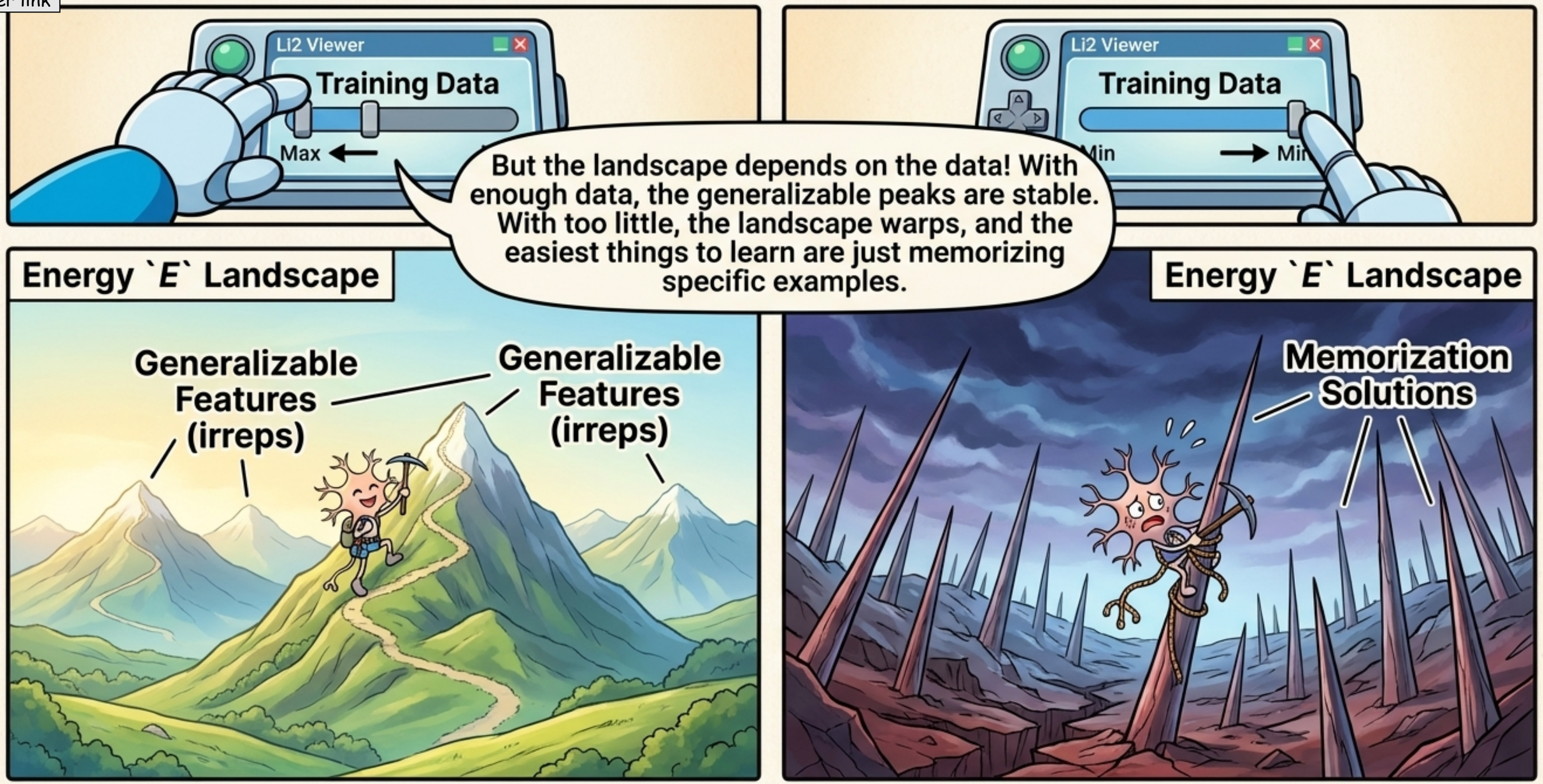


For a task like modular addition, we can prove exactly what these features are. They're the ***irreducible representations*** (irreps) of the task's algebraic structure—in this case, Fourier bases!

So they're learning the fundamental math of the problem, which is way more efficient than just memorizing answers!

****Technical Caption*:** For group arithmetic tasks, the local maxima of E are characterizable. For the Abelian group of modular addition, they are Fourier bases. Reconstructing the target using these $2M$-2 features is far more efficient than a memorization solution requiring M^2 nodes. (Source: Corollary 2, Theorem 3)

Paper link

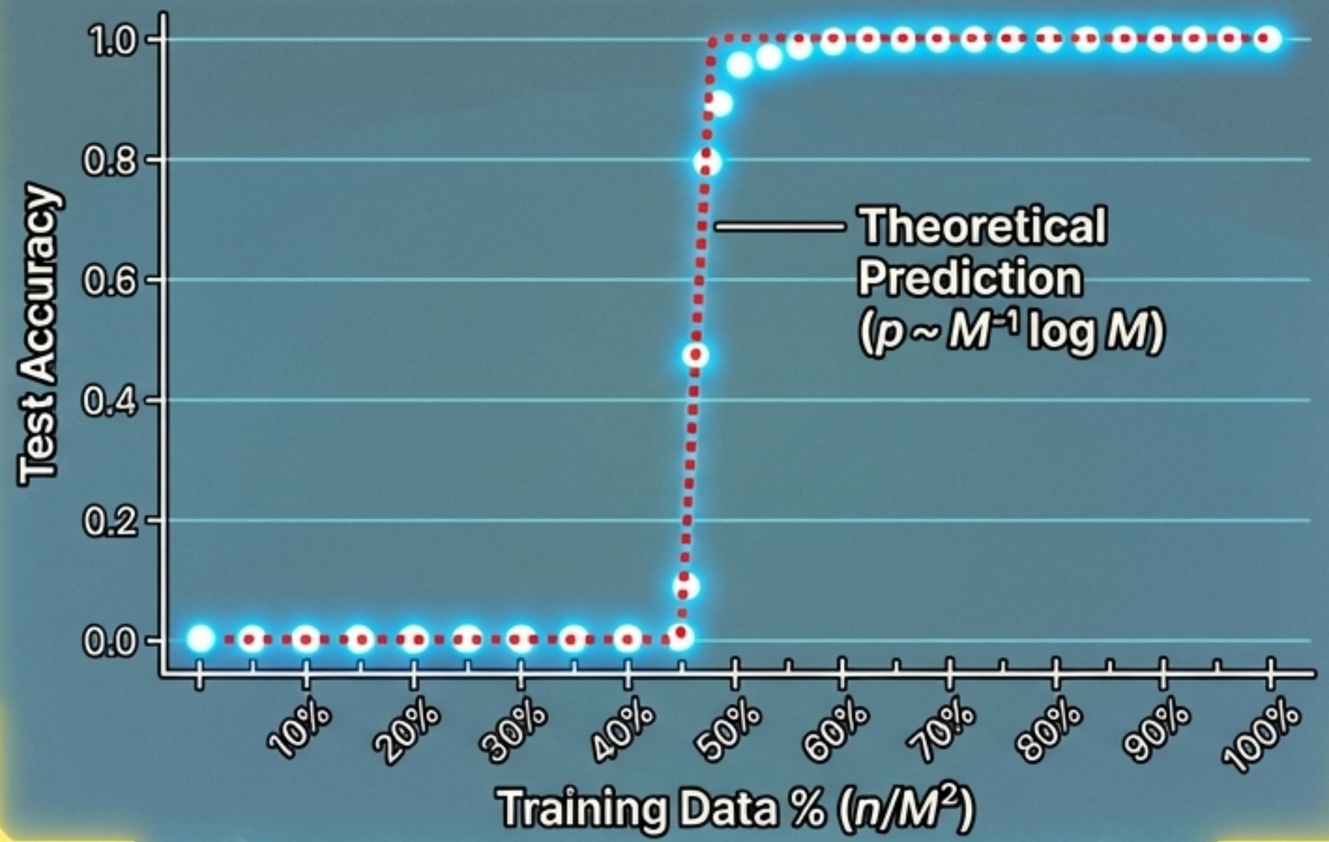


Technical Caption: Data governs the landscape of `E`. Sufficient training data maintains the shape of generalizable local maxima. Insufficient data leads to the emergence of non-generalizable, memorization-based local maxima. (Source: Theorems 4 & 5)

Paper link

This allows us to derive a provable scaling law! We can predict the precise amount of data needed for the network to 'grok'. The theory matches the experiment perfectly.

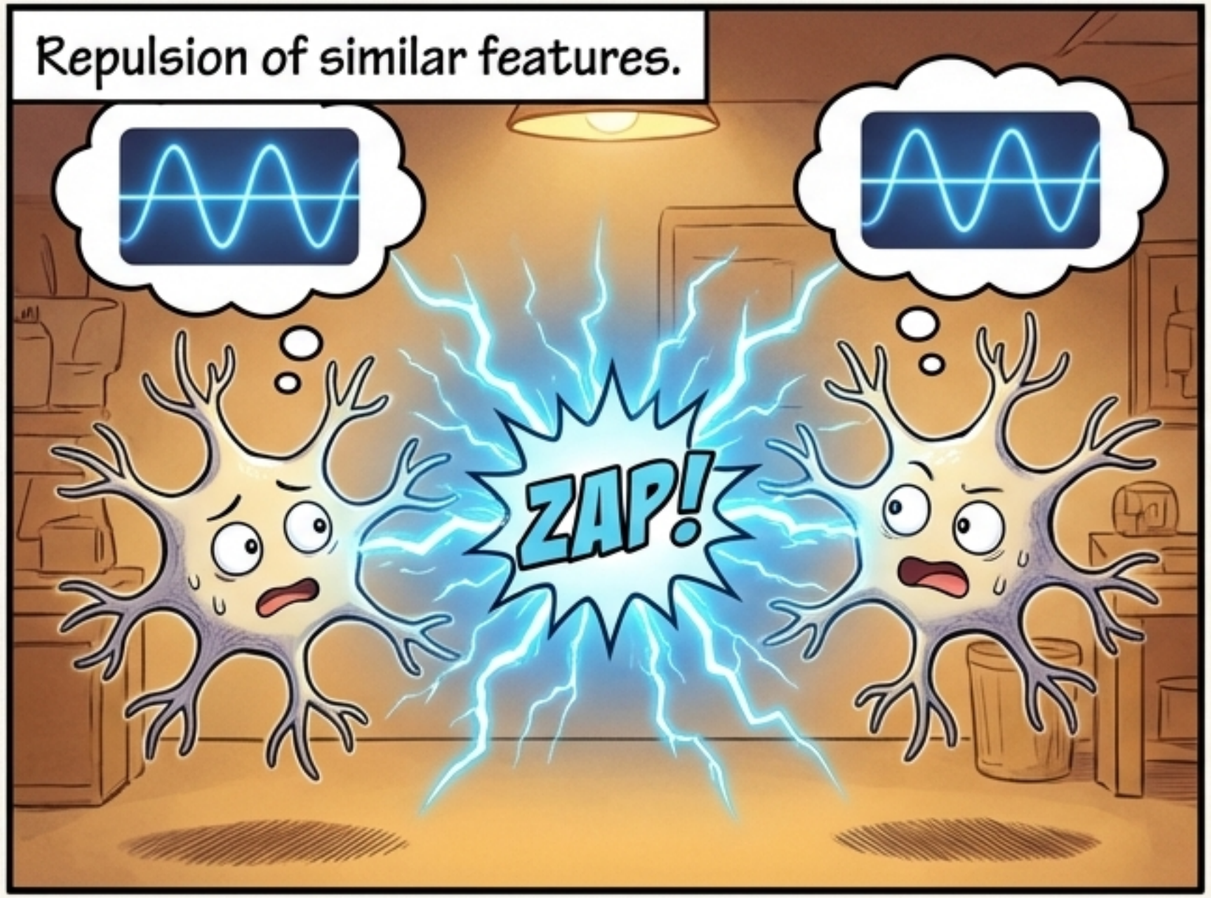
Theory Meets Experiment: A Provable Scaling Law

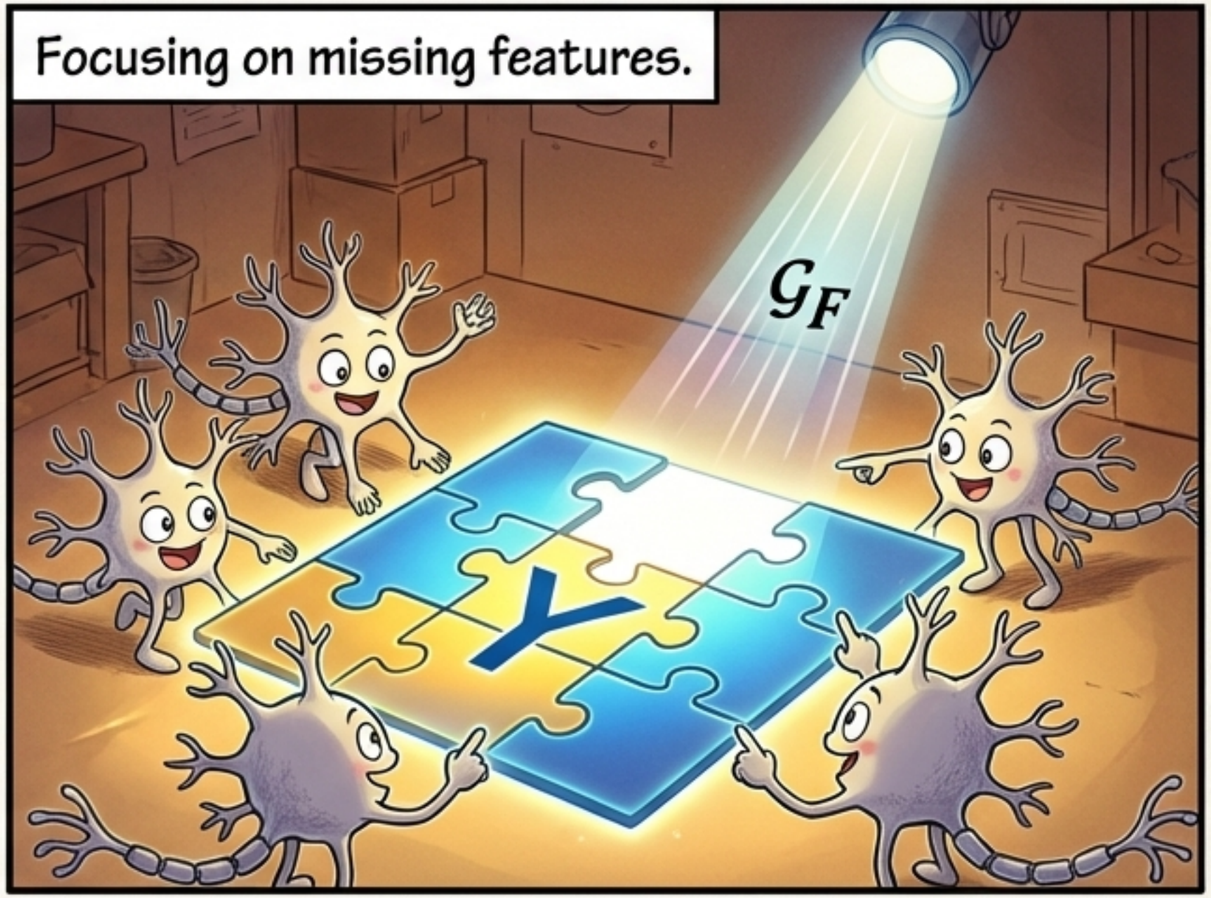


So it's not a mystery at all! There's a mathematical rule for when the big jump in understanding happens!

Paper link

In Stage III, the neurons are no longer independent. They start to interact. Neurons with similar features repel each other to promote diversity. The gradient G_F also changes to guide them, pointing out the parts of the problem they haven't solved yet.

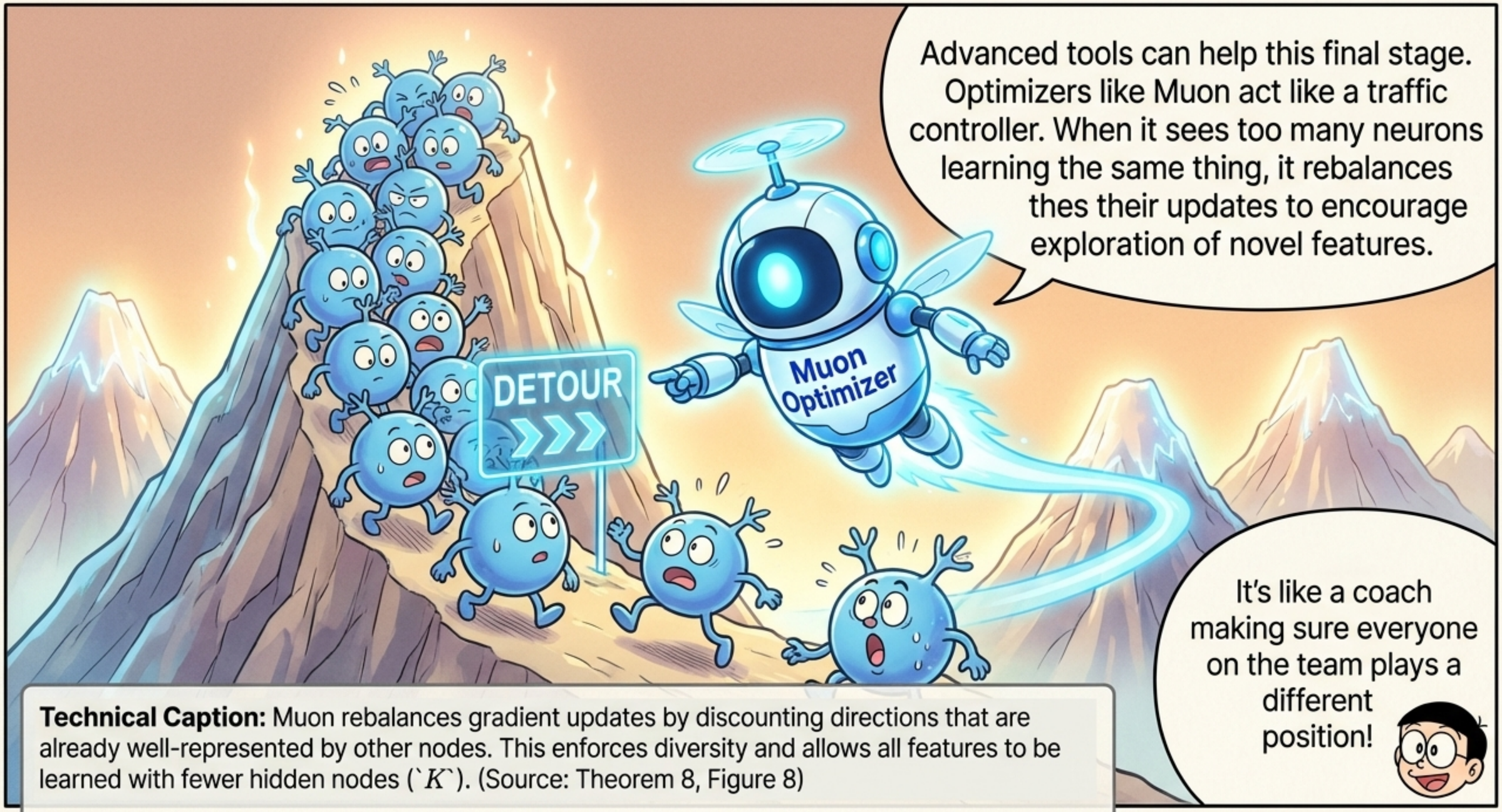




Technical Caption:

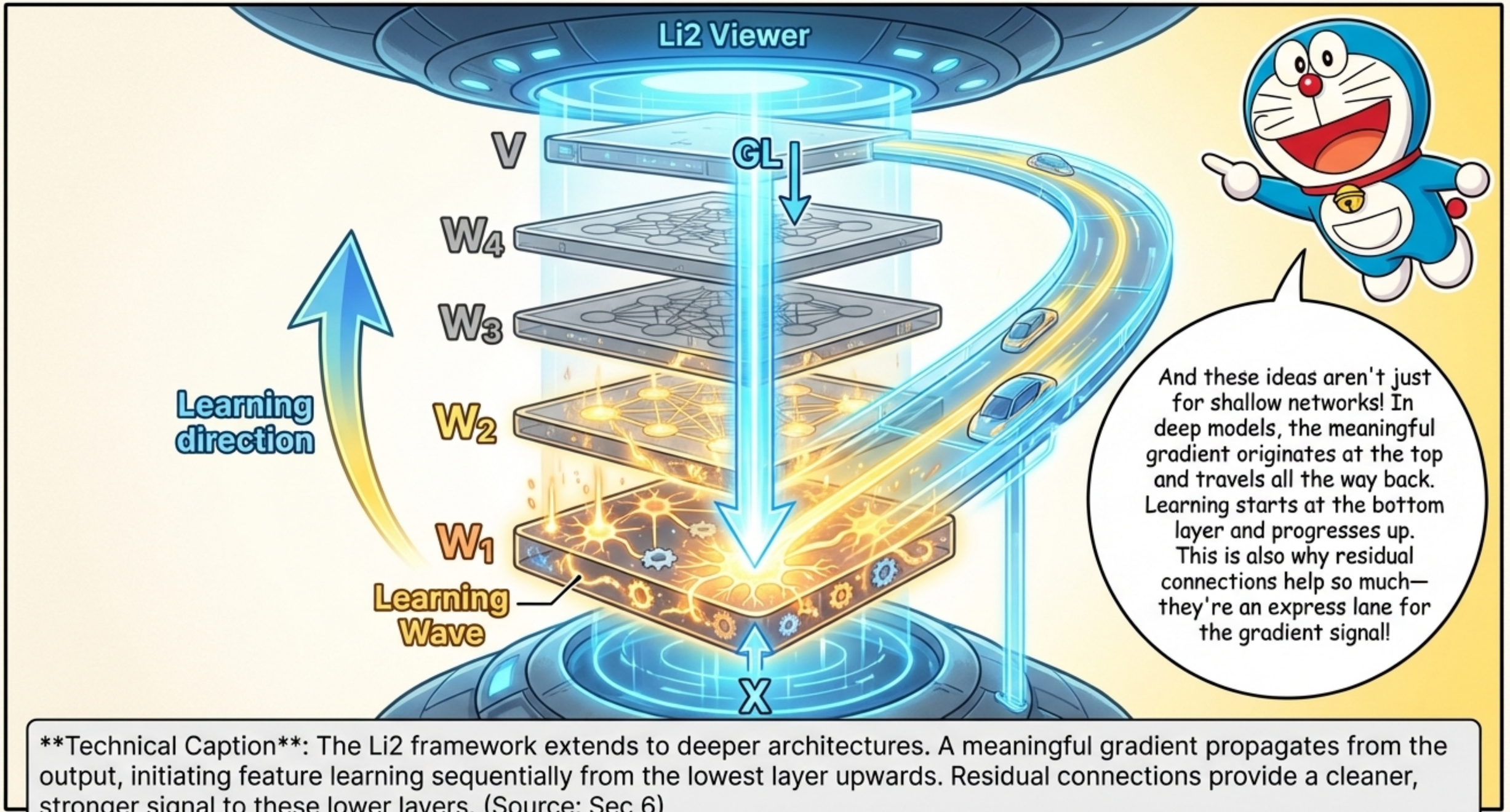
- **Repulsion (Thm. 6):** Off-diagonal terms in the gradient dynamics push nodes with similar activations apart.
- **Top-down Modulation (Thm. 7):** Once a subset of irreps S is learned, GF changes to create a modified energy function E_S with local maxima only on the missing irreps.

Paper link



Technical Caption: Muon rebalances gradient updates by discounting directions that are already well-represented by other nodes. This enforces diversity and allows all features to be learned with fewer hidden nodes (`$K$`). (Source: Theorem 8, Figure 8)

Paper link



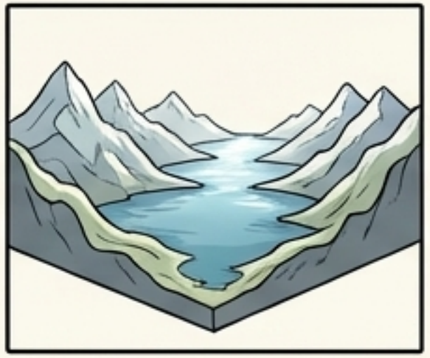
Technical Caption: The Li2 framework extends to deeper architectures. A meaningful gradient propagates from the output, initiating feature learning sequentially from the lowest layer upwards. Residual connections provide a cleaner, stronger signal to these lower layers. (Source: Sec 6)

Paper link

So “sharp” doesn’t just mean "bad," it can mean "early stage overfitting.” And the good, generalizable solutions are “flat” because they represent real, efficient features!

Common Heuristic

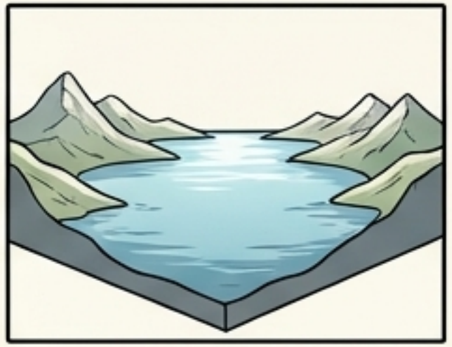
Sharp Optima = Memorization



Flat Optima = Generalization

Li2 Framework View

Stage I: Overfitting on random features



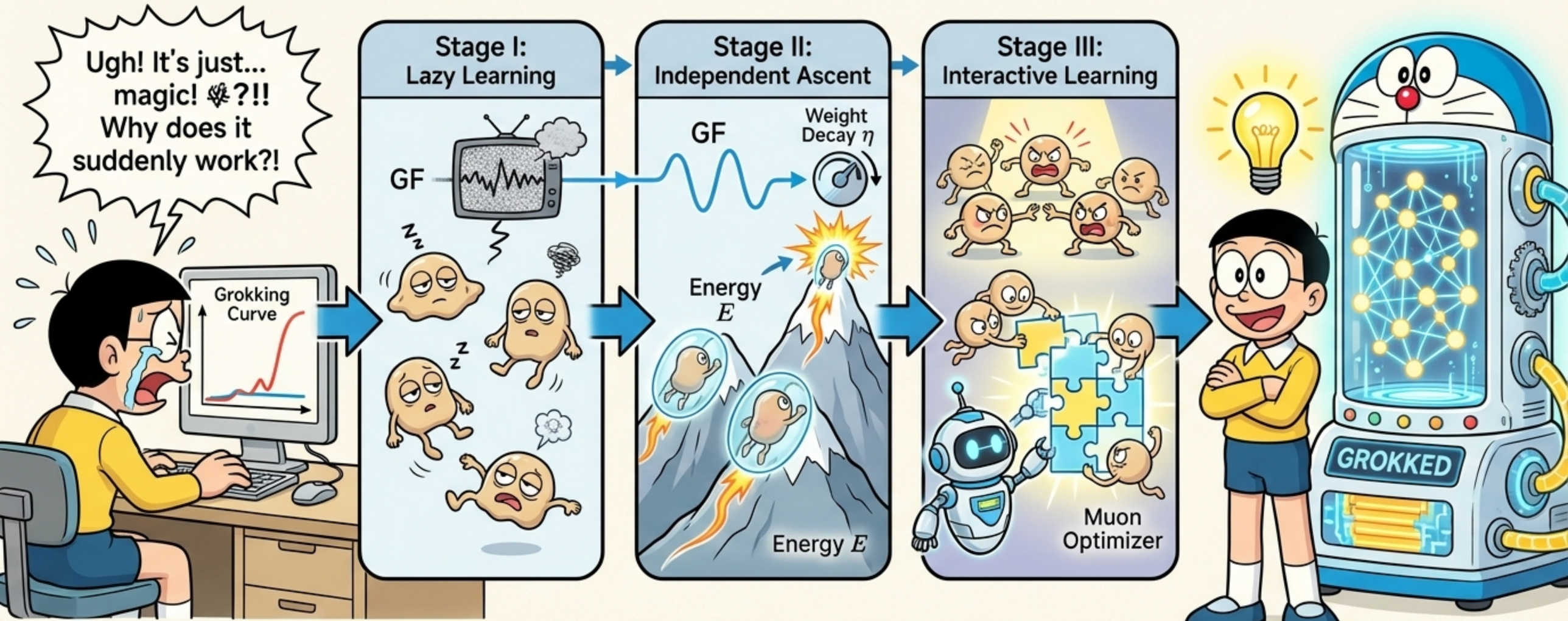
Stage II: Learned features are provably flat local maxima of `E`

Exactly. The initial memortion in grokking corresponds to the sharp optimum of the Stage I ridge solution.

True generalization comes from finding the provably flat local maxima of the energy function E. (Source: Sec 7, Corollary 1)

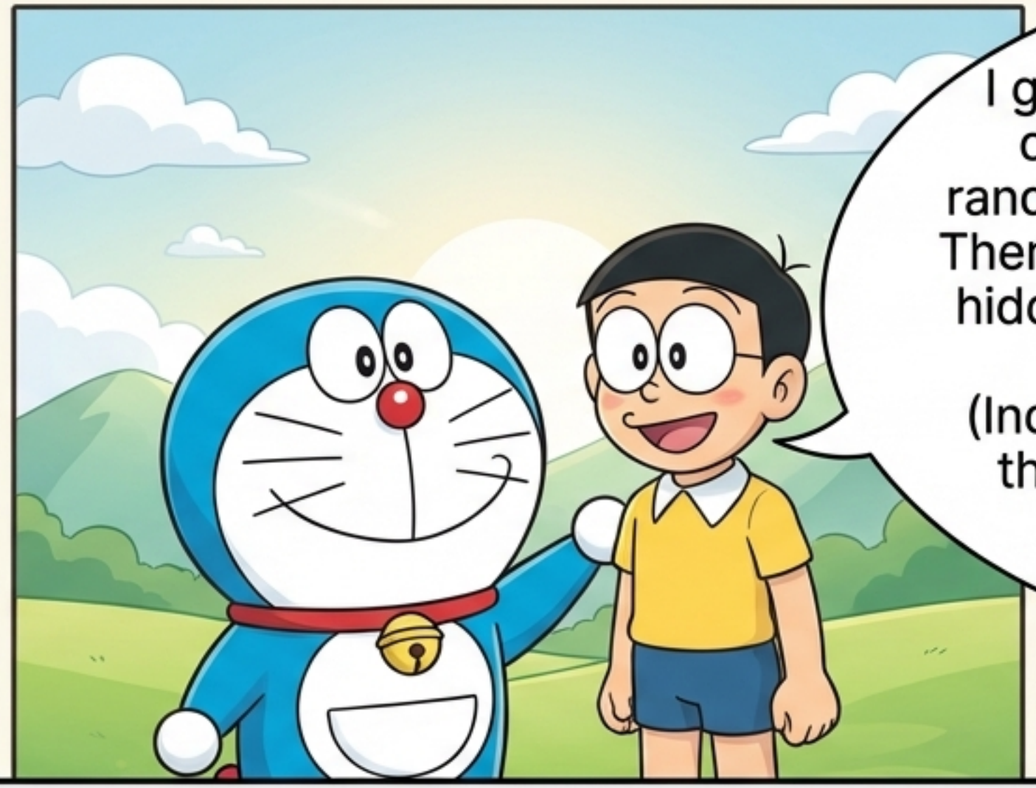
Paper link

Grokking: From Magic to Mechanism.



The journey is not a switch from memorization to generalization, but a structured, three-stage process of feature emergence driven by gradient dynamics.

Paper link



****Doraemon's Concluding Thought****

The Li2 framework demystifies grokking by modeling the precise dynamics of how features emerge. It reveals that grokking is **the observable effect of a network transitioning from overfitting on random features** to **learning the true, efficient sfficient structure of the data**, governed by a data-dependent energy landscape.

By understanding the mechanism, we derive provable scaling laws that connect theory directly to practice.