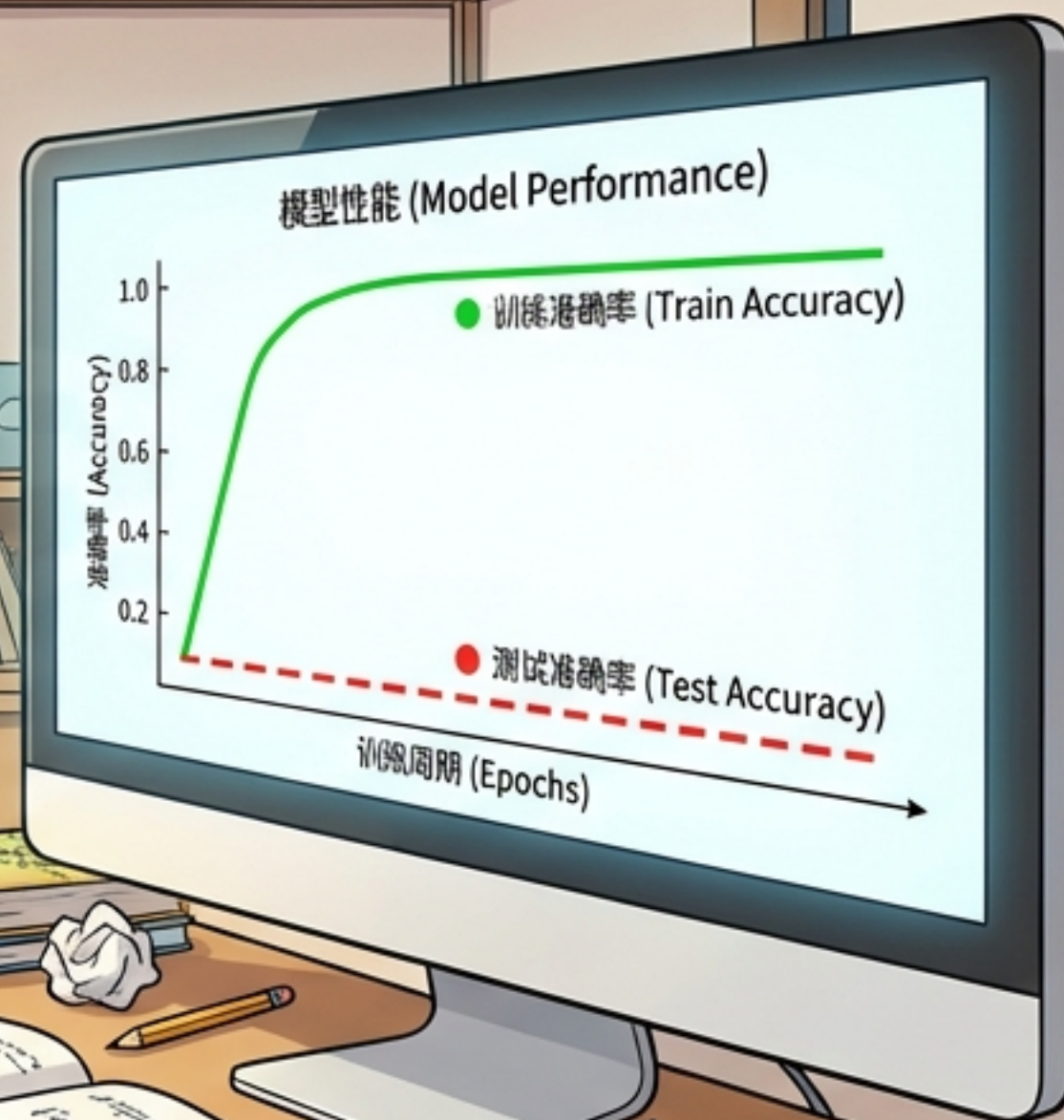
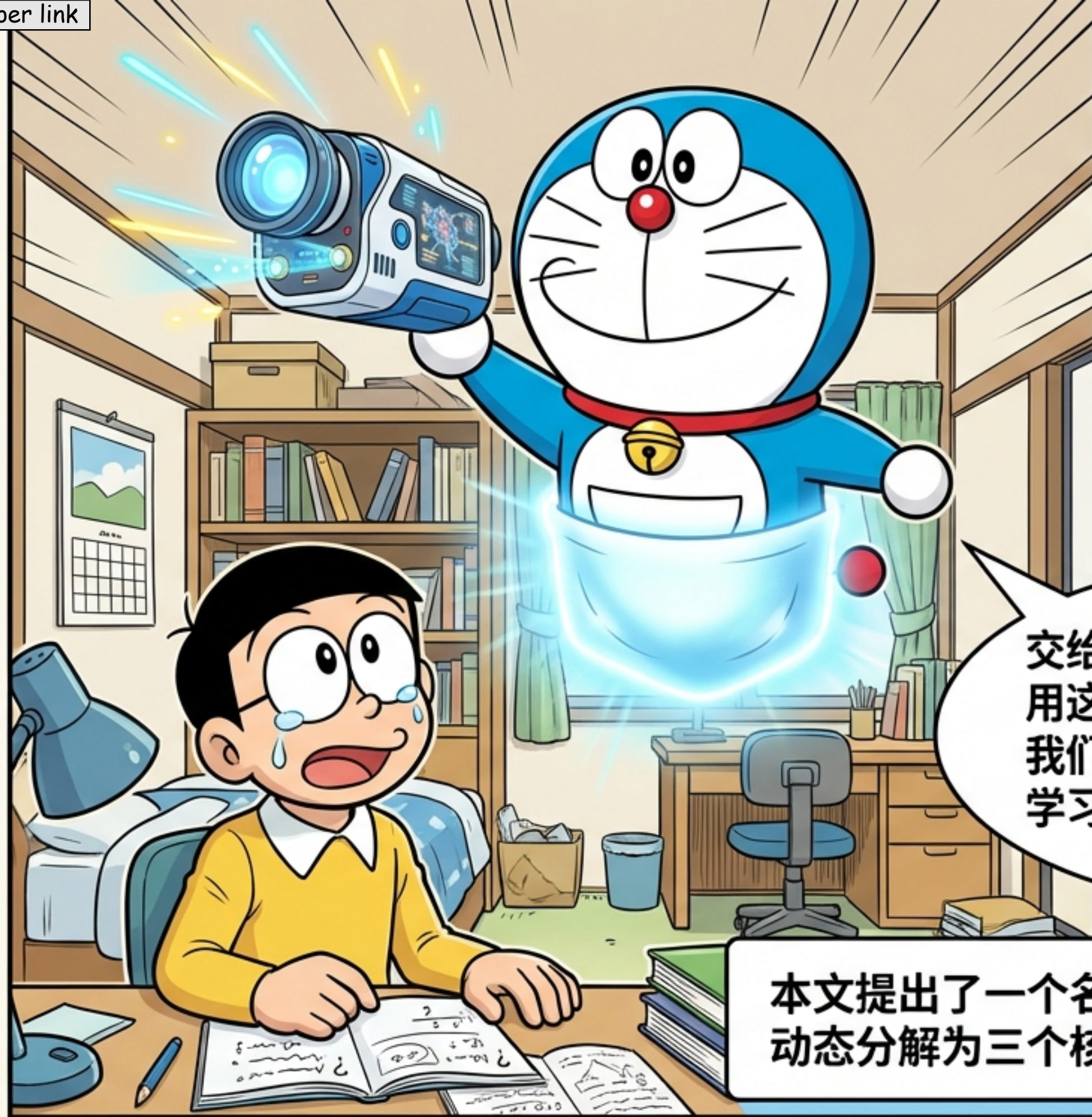


哆啦A梦！怎么办啊！我的神经网络训练准确率都100%了，可是一到测试就完全不行！它只会死记硬背！



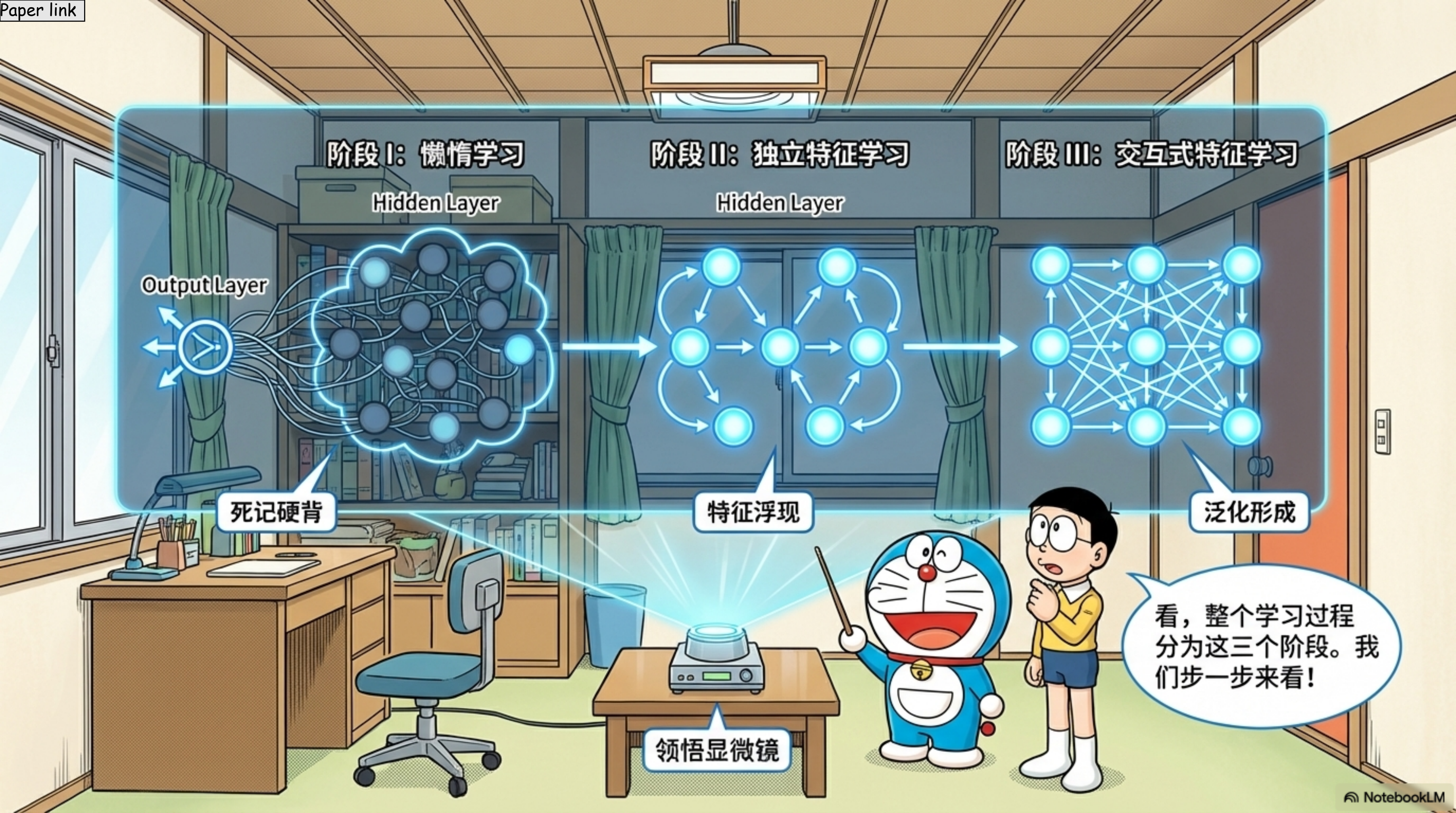
神经网络训练中的一个谜题：Grokking（顿悟）现象。模型在长时间的“死记硬背”（过拟合）后，会突然学会泛化。这究竟是如何发生的？

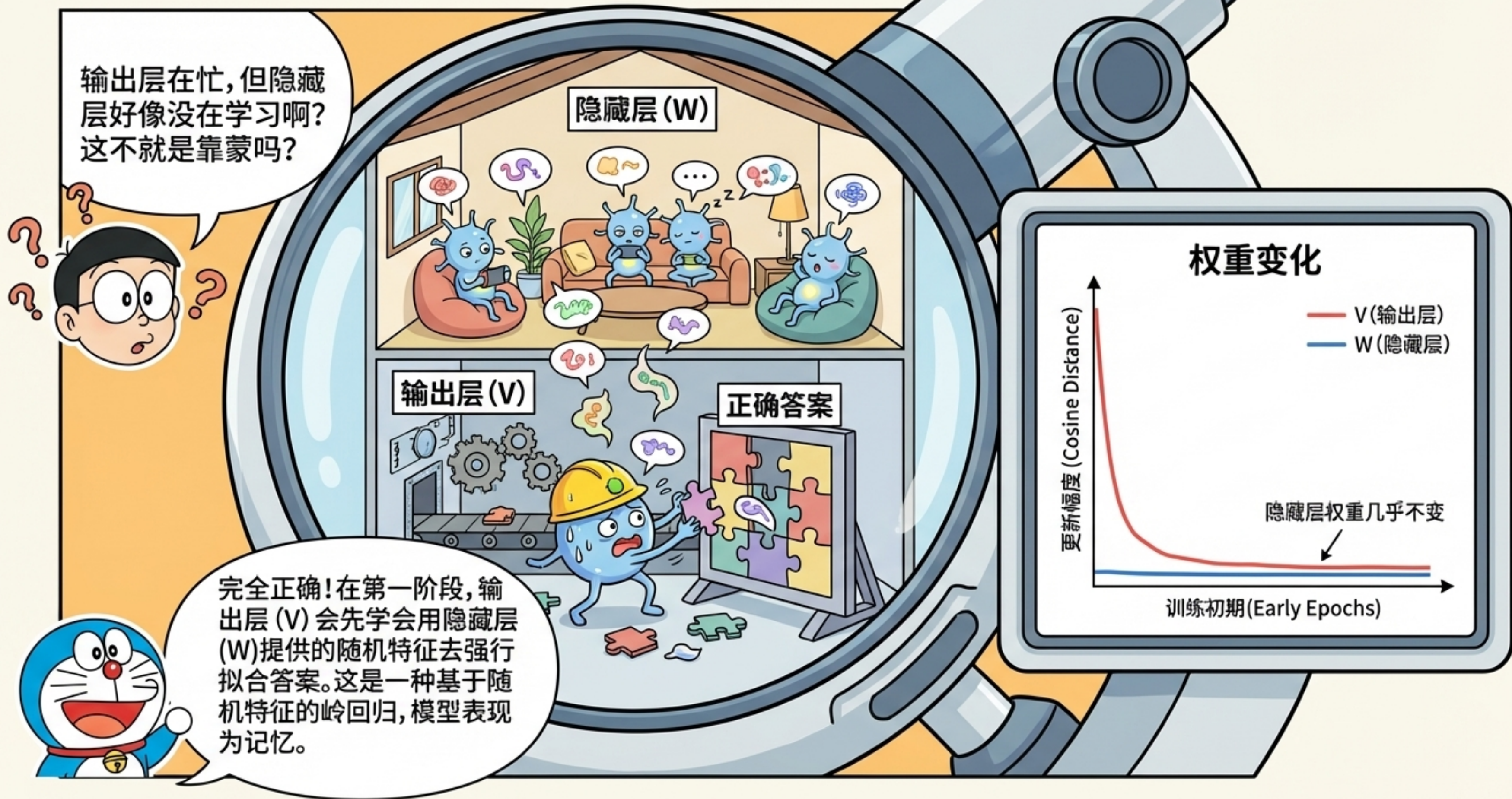


领悟显微镜

交给我吧，大雄！
用这个“领悟显微镜”，
我们就能看清楚神经网络内部
学习过程到底分几步！

本文提出了一个名为 **Li2** 的数学框架，它将 Grokking 的学习动态分解为三个核心阶段，揭示了从记忆到泛化的完整路径。

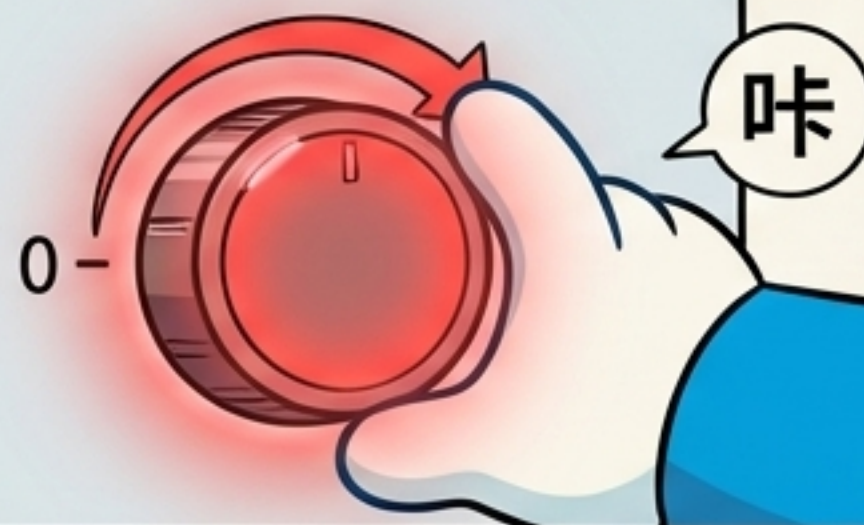




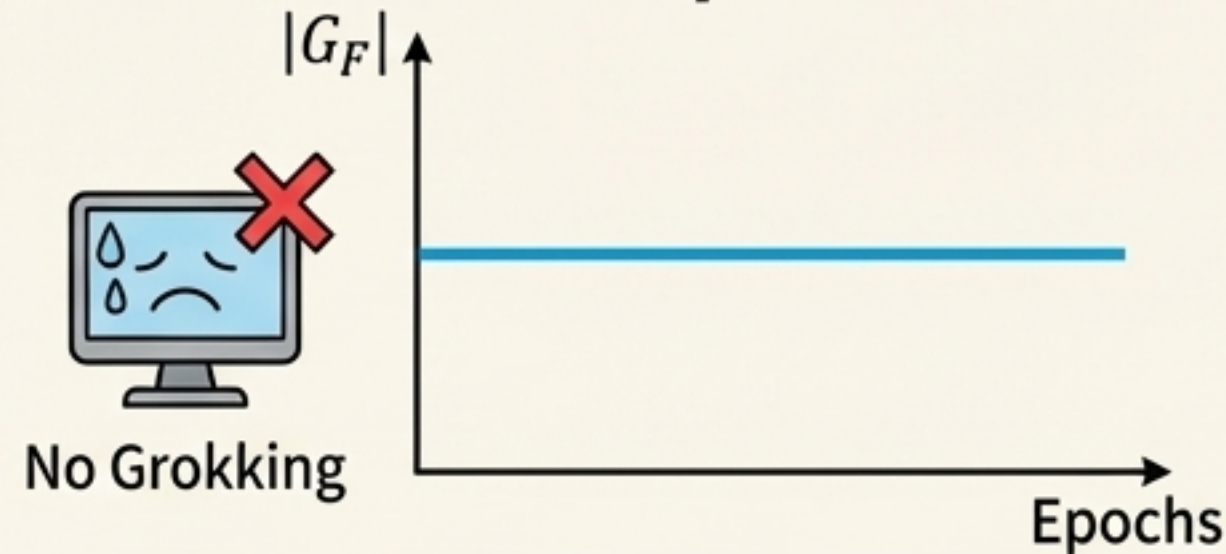
“关键就在这里！如果没有“权重衰减” ($\eta > 0$), 反向传播的梯度信号GF就是0, 隐藏层就永远学不到东西!”



权重衰减 (Weight Decay, η)



$\eta = 0$

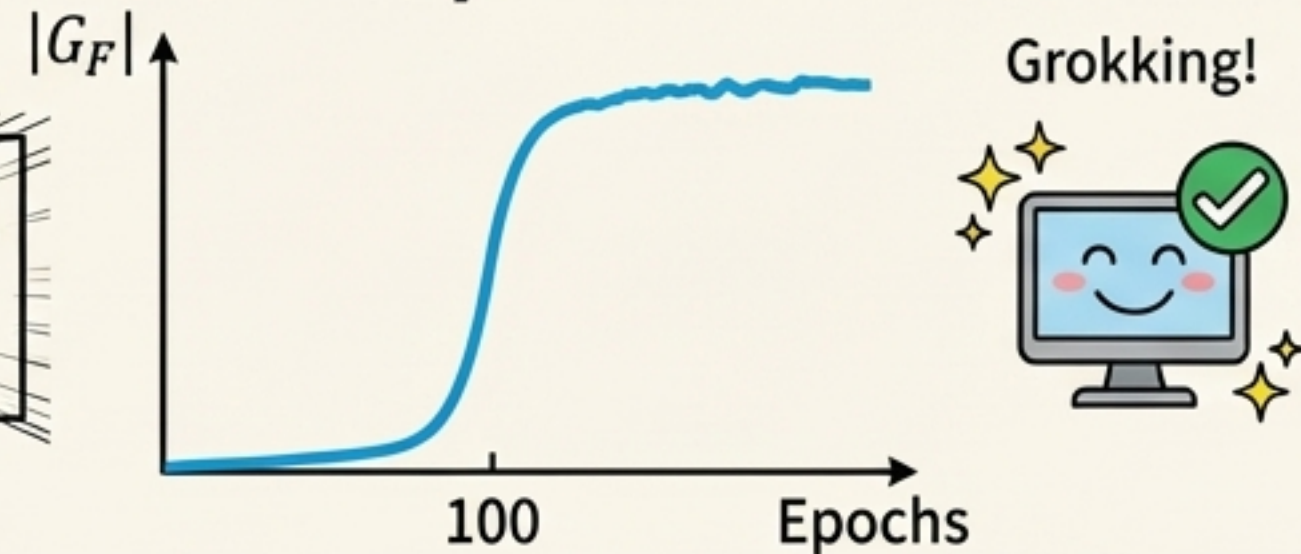


无权重衰减：无法启动特征学习

$$G_F = \mathbf{P} \eta \tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^T \tilde{\mathbf{F}} (\tilde{\mathbf{F}}^T \tilde{\mathbf{F}} + \eta \mathbf{I})^{-1}$$

当 $\eta = 0$, G_F 可能为 0

$\eta > 0$



有权重衰减：GF携带标签信息，驱动特征学习

有了权重衰减，GF就携带了目标 $\tilde{Y}$ 的信息。
现在，每个神经元都可以独立行动，像登山一样，去寻找能量函数的最高点！

能量函数 E (Energy Function E)

能量函数 E (The Energy Function E)

神经元 w_j 的独立学习动态，恰好是能量函数 $E(w_j)$ 的梯度上升。

$$E(w_j) = \frac{1}{2} \|\tilde{Y}^T \sigma(Xw_j)\|_2^2$$



每个神经元学到的特征，就是使这个能量函数达到局部最大值的解。

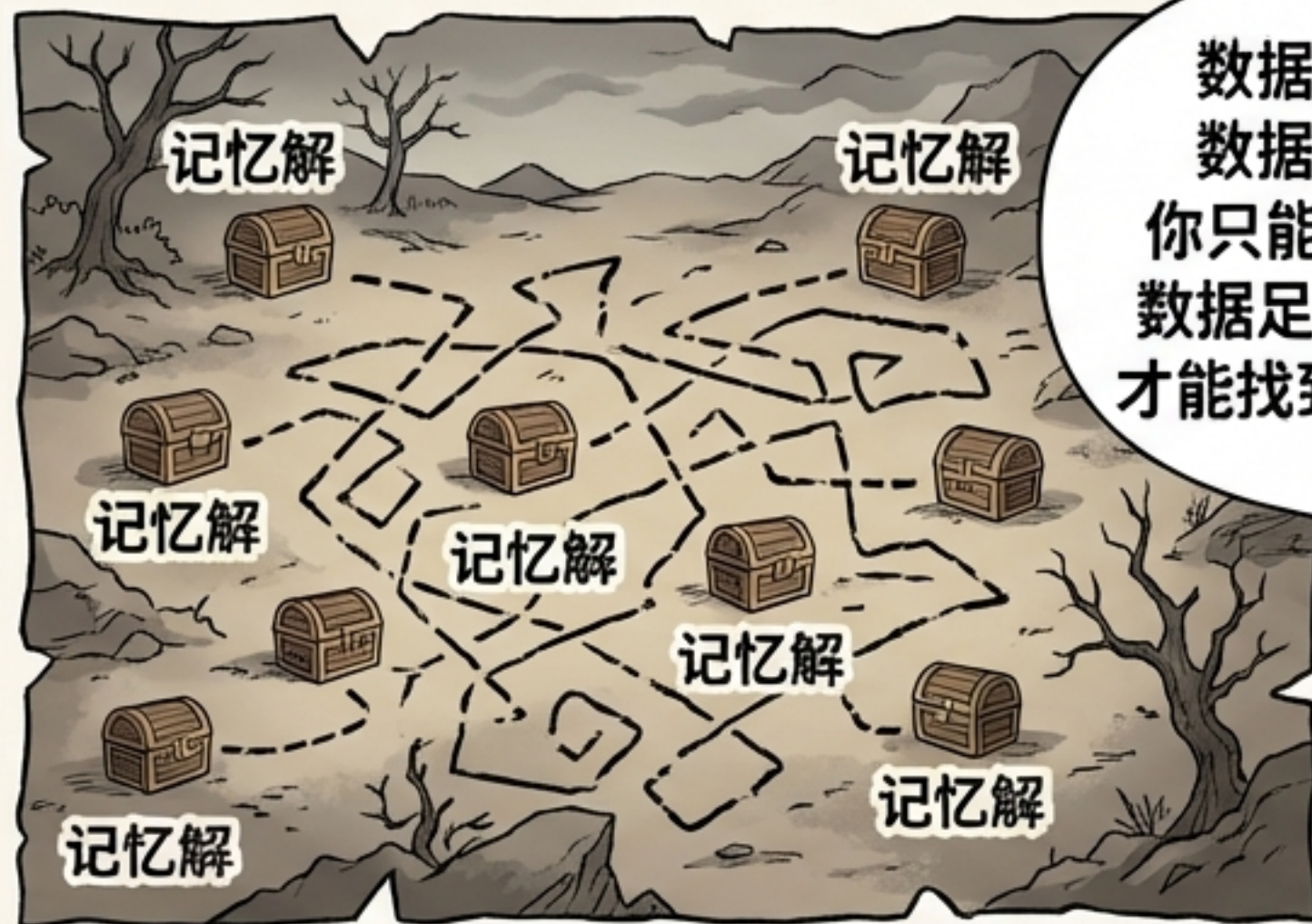


泛化 vs. 记忆 (Generalization vs. Memorization)

用 $2M-2$ 个神经元学到的特征就能完美重构目标，而死记硬背需要 M^2 个神经元。这就是泛化回路比记忆回路更高效的数学原理。

没错！对于群运算任务，能量函数 E 的局部最大值，正好对应着群的‘不可约表示’ (irreps)，比如傅里叶基。这比死记硬背 M^2 个组合要高效得多！

数据不足 (Insufficient Data)



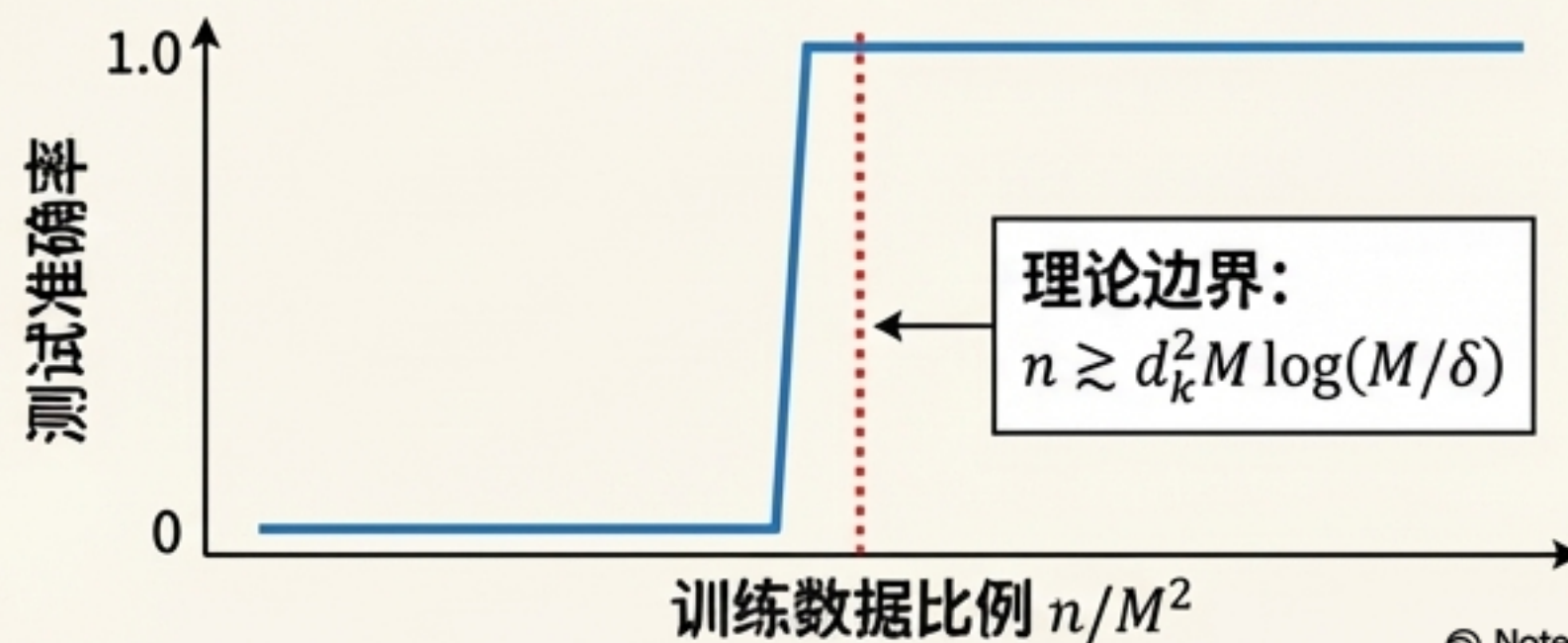
数据量就是这张藏宝图。
数据太少，地图是错的，
你只能找到记忆的‘小宝藏’。
数据足够多，地图才是对的，
才能找到通往泛化的‘大宝藏’！



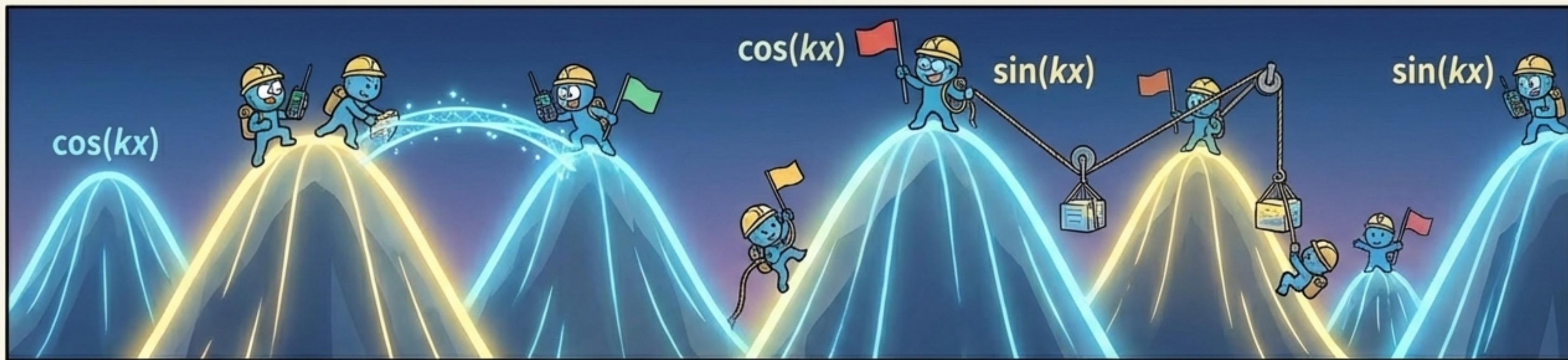
数据充分 (Sufficient Data)



测试准确率 vs. 训练数据比例



阶段 III: 交互式特征学习 - 从各自为战到团队合作



特征排斥 (Feature Repulsion - Thm. 6)

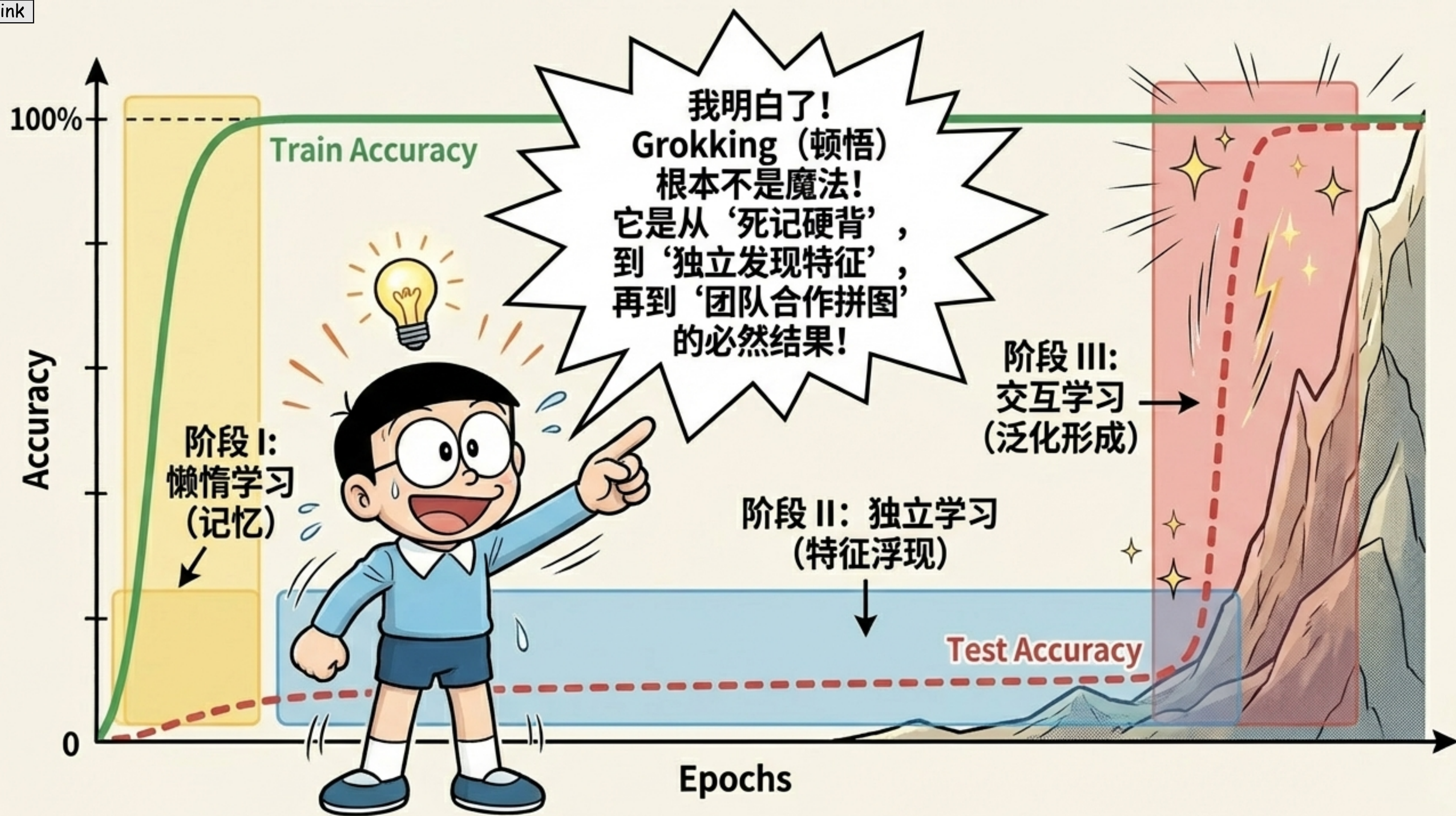


相似的特征会相互排斥，避免冗余，提升多样性。

自上而下的调节 (Top-down Modulation - Thm. 7)



当部分特征被学会后， GF 会发生变化，集中力量去学习“缺失”的特征。

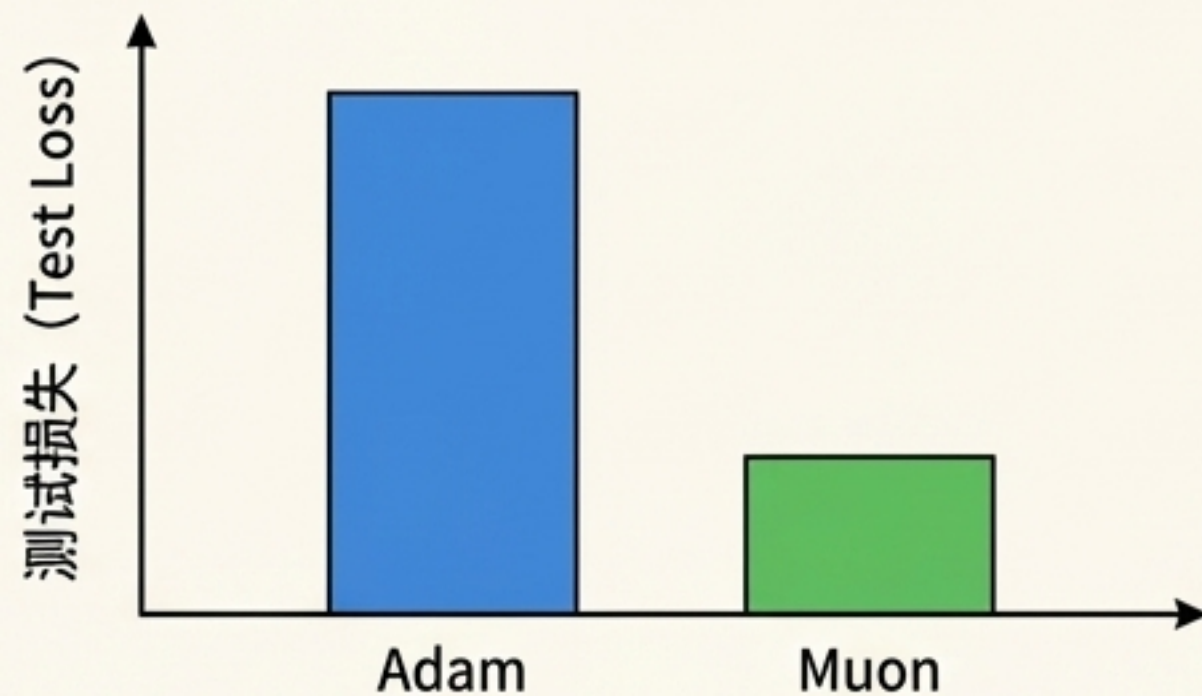


Grokking: A sudden generalization after a long period of poor performance.

Muon 优化器

如果神经元数量 (K) 有限，我们还有这个“Muon优化器”，它能主动增强未被充分探索的方向，鼓励神经元学习更多样的特征！

优化器性能对比 (K 较小时)



Muon 如何工作

Muon通过重新平衡梯度更新，抑制已学特征方向，放大新特征方向，从而高效地找到所有需要的特征解。

Grokking 不是魔法，而是三步走的学习过程



1. 懒惰学习 (Lazy Learning): 过拟合随机特征



2. 独立学习 (Independent Learning): 在能量景观上寻找泛化特征



3. 交互学习 (Interactive Learning): 协作以补全最终解

数据量决定了“能量景观”的形态

└ 决定了模型会收敛到记忆解还是泛化解



权重衰减 (Weight Decay) 是从记忆切换到特征学习的关键催化剂

通过这三个步骤，
模型就能实现
从“死记硬背”
到“真正理解”
的飞跃！

真厉害，
大雄！

原来
如此！