深度流形 神经网络数学

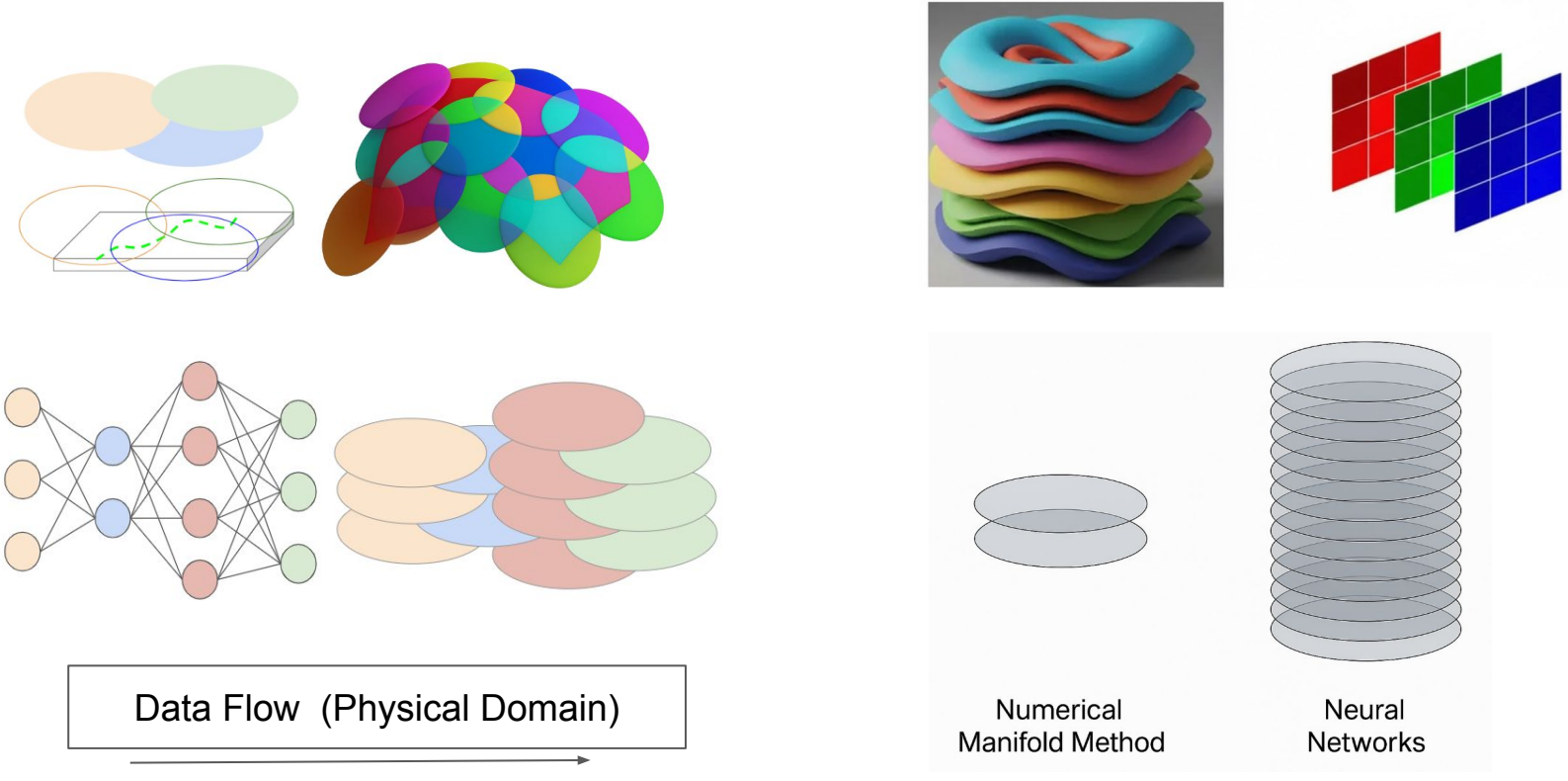
马远, 石根华
deepManifold

Deep Manifold Part 1: Anatomy of Neural Network Manifold, arXiv:2409.17592
Deep Manifold Part 2: Neural Network Mathematics, arXiv:2512.06563

2026.01

深度流形(第一部分):神经网络流形的结构剖析

Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592

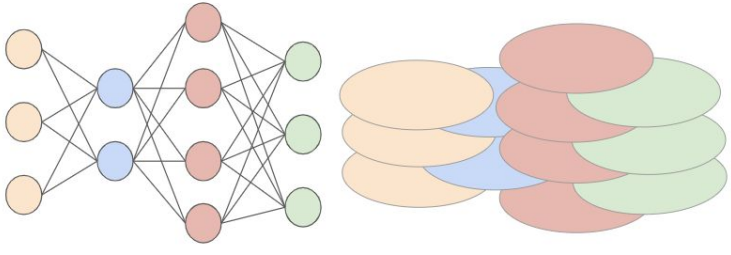


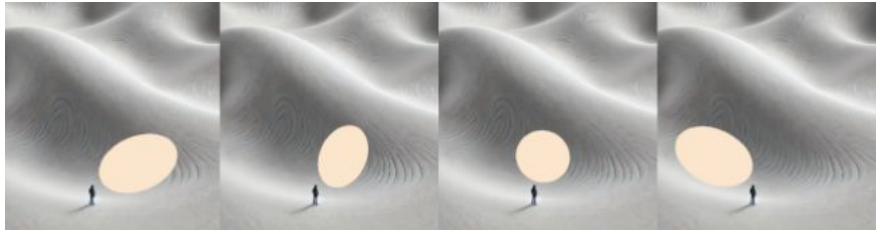
神经网络数学

神经网络几何

几何赐予我们一双从上方洞察一切的眼睛，正是通向自由的 阶梯。

- 相互连通并堆叠的分片光滑流形共同形成了表示空间的几何结构。
- 节点覆盖充当这些分片光滑流形的局部单元，其取向在每一次迭代中都会发生改变。
- 这些分片光滑流形具有可微性和可积性。

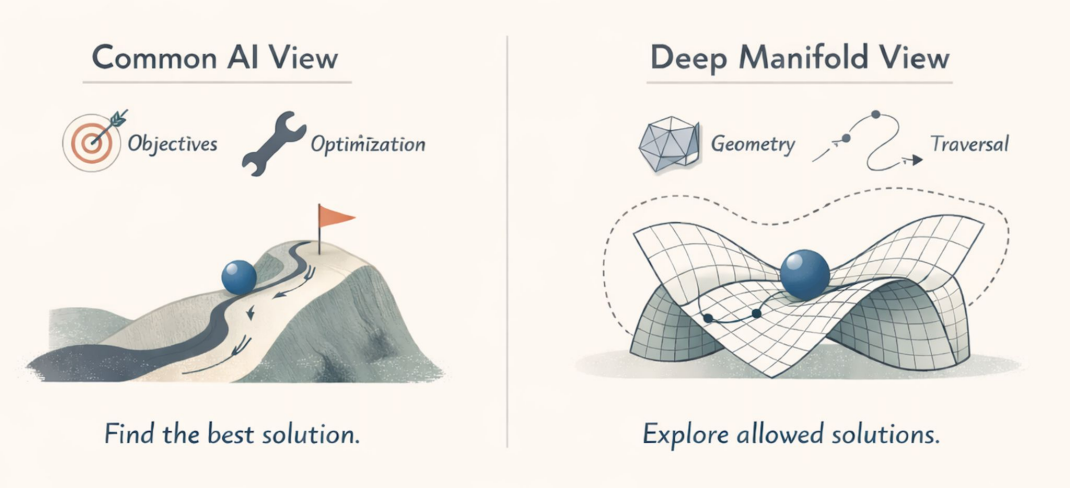




几何主导

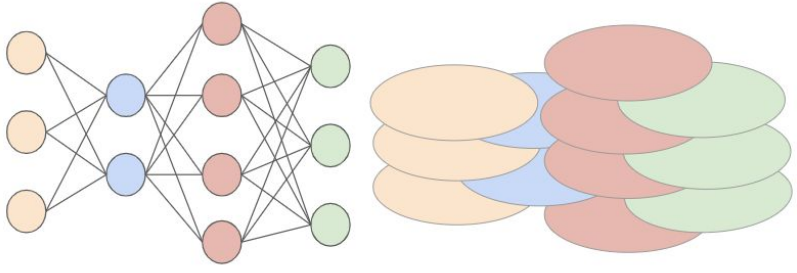
- 主流 AI 观点
  - 目标(损失)+ 优化(参数)决定解
  - 几何结构只是副产物，处于次要地位
- 深度流形观点
  - 几何决定哪些解能够存在
  - 优化只是沿着预先成形的流形进行遍历

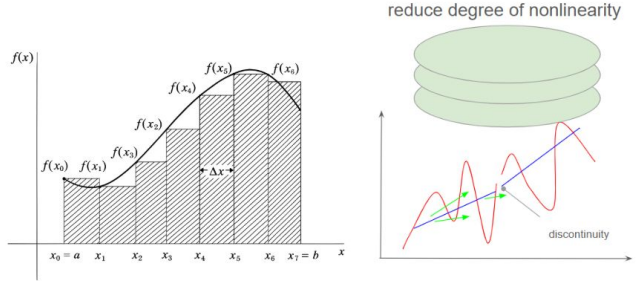
学习是一个反问题且不可辨识，几何是唯一稳定的先验，几何决定推理的内在路径

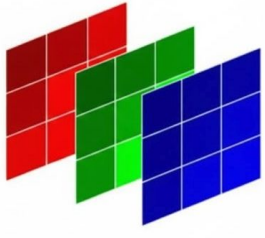


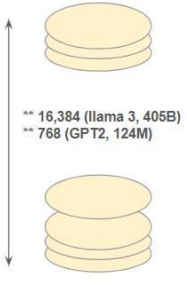
多层分片光滑覆盖(流形)

- 流形包括:一个点、一条直 线、一个圆、一个三角形，以及无限 维的 Banach 流形
- 图像 RGB: 三个堆叠的逐点流
- 神经网络:由连接、拉伸与多层分片光滑覆盖(流形)构成
- 多层分片光滑覆盖(流形)的优势:高阶非线性数据处理





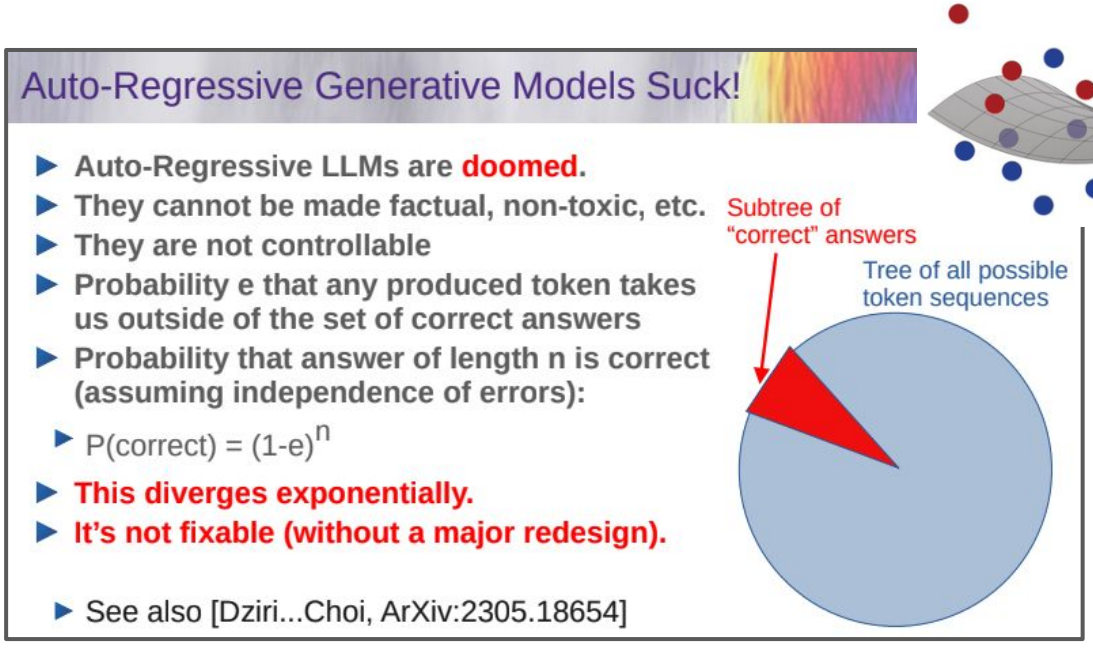


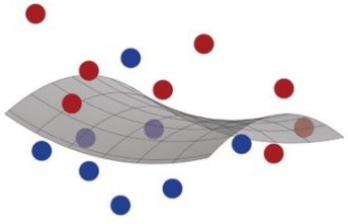


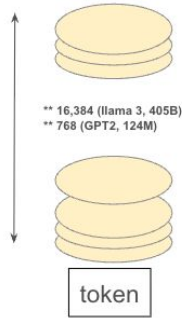
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592;

Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

杨立昆(Y. LeCun) 对于单个流形是正确的，但 为什么 Transformer 表现得如此出色？





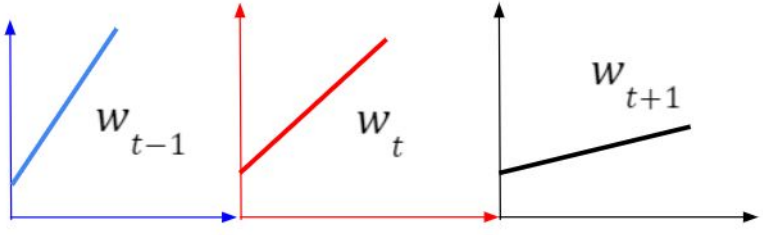


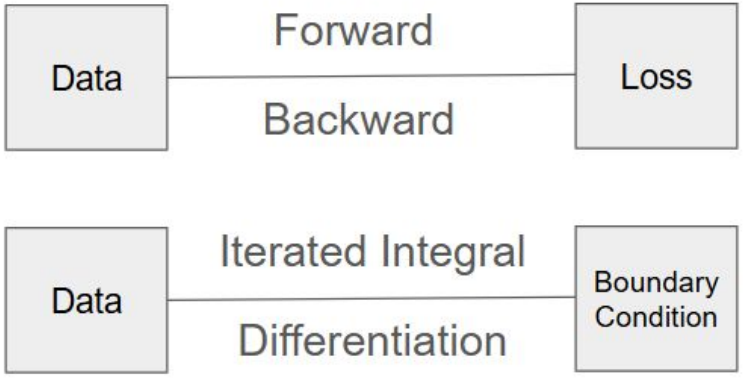
- 指数衰减批判把生成过程视为一条单一流形的轨迹，每一步都存在独立的失效风险.
- Transformer 作用于堆叠的分片流形之上，偏差会被投影至共享的几何子空间
- 错误事件相互重叠，而非逐级累积。生成的稳定性来源于联合界几何（union-bounded geometry），而不是乘性概率的崩塌。

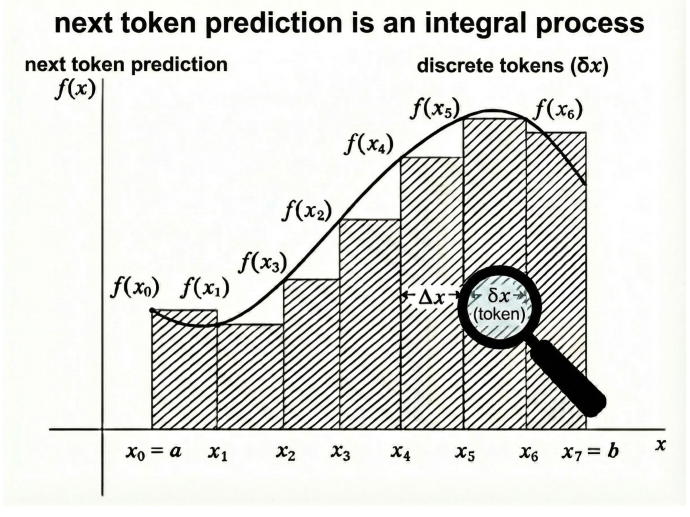
神经网络代数

代数是运算的科学，是所有 变换背后那位沉默的幕后力量

- 坐标系统随每次迭代而演化
- 计数作为最原始的代数单元
- 前向传播的迭代积分结构
- 激活值是无属性的







神经网络方程

方程是实现计算的静默连接器

- 不动点残差作为原始方程

$$x_{\text{text}} = E(\text{“dog”}), \qquad x_{\text{image}} = E(\text{dog pixels})$$

$$f(x) - x = e(x), \qquad \min_{\theta} e(x) \qquad \theta^* = \arg\min_{\theta} \mathbb{E}_x \left\| f_\theta(x) - x \right\|.$$

- 神经网络不动点的拉格朗日公式

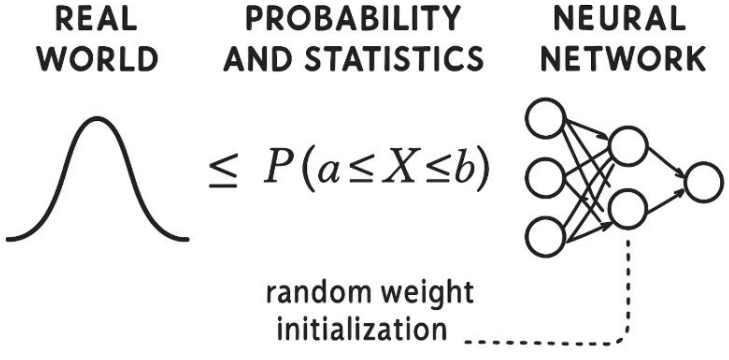
$$L(\theta, \lambda) = \mathbb{E}_x \left\| f_\theta(x) - x \right\|^2 + \lambda \, g(\theta)$$

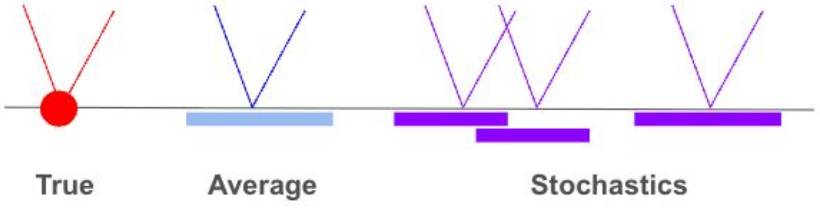
$$\nabla_\theta L(\theta, \lambda) = 0 \implies \text{critical point of } \mathbb{E}_x \| f_\theta(x) - x \|^2$$

神经网络随机性

随机世界是一个不平等的世界，但它才是真 实的。

- 随机性体现为不等式，以及基于加法（求和）的群体统计特性。
- 统计结构赋予神经网络学习固有随机性的真实世界的能力，并使其自然形成随机不动点。
- 它轻而易举

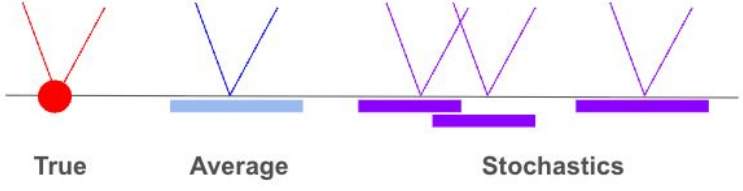


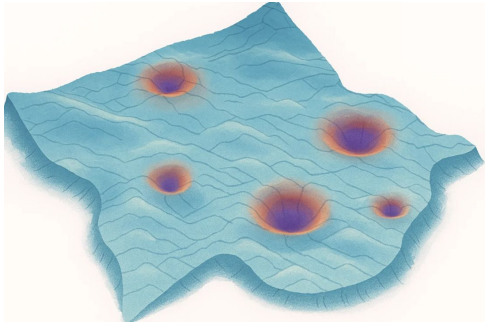


神经网络不动点

不动点理论乃迭代之理论，直至不动

- 迭代过程遍历数十亿个分片光滑流形，从而在基础模型内部催生出无数不动点与收敛路径
- 正因为训练数据自身蕴含高阶非线性、曲率（二阶导数）以及适度的扰动，模型才能够分辨出指向不动点的正确收敛方向

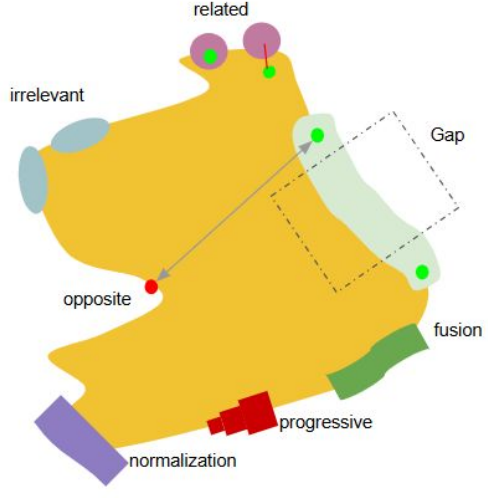


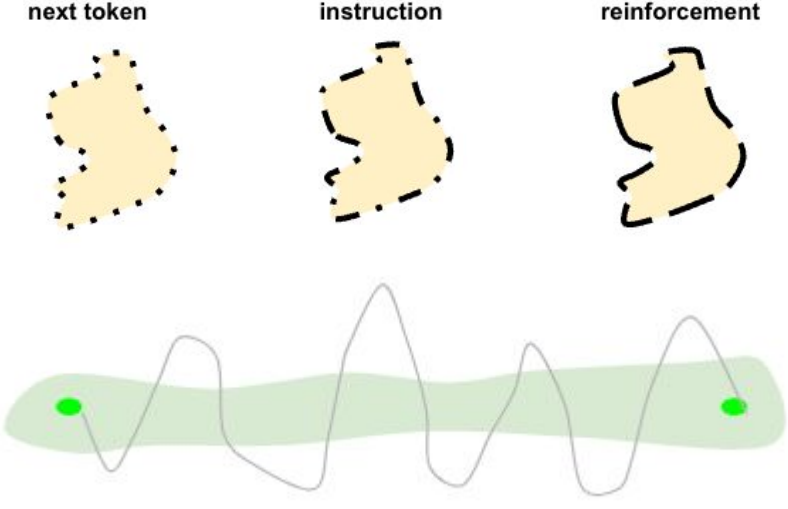


神经网络边界条件

边界条件赋予迭代以目的和方向

- 边界条件是迭代方向的唯一来源，并在训练过程中决定收敛路径
- 当基础模型缺乏静态不动点时，对称的、弱的以及离散的边界条件成为引导高阶非线性系统收敛的必要条件。





可学习数值计算

数学与数值计算

- 数学
  - 原则上具有普适性：寻求在不同领域中成立的解析解
  - 数学擅长于描述：但往往不可直接求解或 计算
  - 结构性限制：解析解的存在并不意味着具有 闭式表达，尤其是在高 阶非线性、不连续或随机系统中
- 数值计算
  - 伽辽金方法：一种数值求解器，在选定的表示空间上，以弱一致性取代精确可解性。
  - 逼近定理：在给定度量下，保证目标函数类可被某一指定函数族以任意精度逼近的理 论结果。离散化、近似、迭代
  - 高度适应性：可引入经验项，采用自适应网格/层结构，以可用收敛性优先于精确性。
  - 实用主义：只要能够收敛，方法并不受限。只要能收 敛 **油盐酱醋**都可以加

数值计算可能具有迷惑性。在缺乏坚实数学基础的情况下，它仍然可以在实践中走得很远. 这在超级计算机时代曾经发生，如今在 AI 中再次出现。算力的扩展并不等同于数学上的扩展。

从不动点到可学习计算

- 不动点定义了什么是学习.

$$x_{\text{text}} = E(\text{“dog”}), \qquad x_{\text{image}} = E(\text{dog pixels})$$

  - 优美的描述，但不是求解器；没有数值计算流程
  - 没有进展的度量方式，也无法处理约束（架构、数据）
- 拉格朗日形式使其变得可求解

$$f(x) - x = e(x), \qquad \min_{\theta} e(x) \qquad L(\theta, \lambda) = \mathbb{E}_x \, \|f_\theta(x) - x\|^2 + \lambda \, g(\theta)$$

  - 拉格朗日平衡态 = 神经网络的不动点
  - ƛ = 边界约束施加器，g(∅) =0 架构 / 数据约束
- 数值迭代使其成为现实
  - 只有在约束条件下对残差进行迭代消减时，数学才会转化为计算
  - 伽辽金方法：方程只有在残差化并通过迭代之后，才在数值上变得可求解
  - 学习并不是由目标函数所定义的，而是由是否存在稳定的数值迭代过程来决定的。

Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

不同模型，不同推理路径

- 神经网络不动点方程:

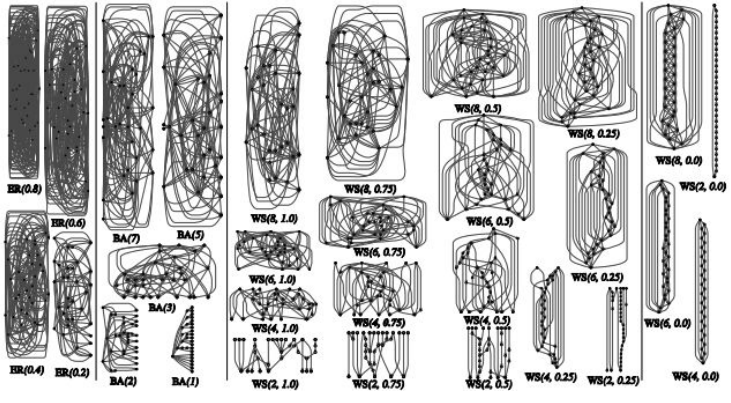
$$f(x) - x = e(x), \min_{\theta} e(x)$$

- 神经网络不动点的拉格朗日形式:

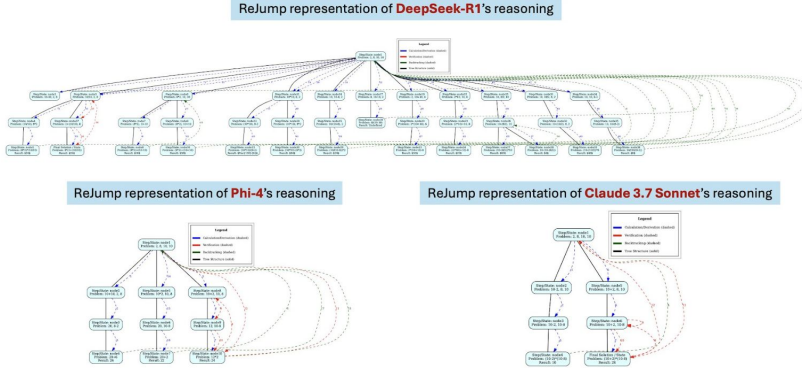
$$L(\theta, \lambda) = \mathbb{E}_x \left\| f_\theta(x) - x \right\|^2 + \lambda\, g(\theta)$$

g(θ): 相同提示，不同推理路径，却得到同样准确的输出

- 将模型的架构约束和/或数据约束写成等式约束形式 g(θ) = 0



Exploring Randomly Wired Neural Networks for Image Recognition, arXiv:1904.01569



ReJump: A Tree-Jump Representation for Analyzing and Improving LLM Reasoning, arXiv:2512.00831v2

基于伽辽金方法的神经网络数值解释

- 神经网络应当从伽辽金型数值形式来理解，而非从传统优化理论来解释
- 传统伽辽金方法
  - 求解算子方程 $\mathcal{N}(u) = 0 \quad \text{on}\Omega$
  - 以弱形式强制满足解 $\int_\Omega \mathcal{N}(u_h) v_h \, d\mu = 0$
    - 有限维试探空间、残差正交性、数值迭代
- 神经网络并非在优化函数，而是在学习到的流形上对残差方程进行数值求解。

伽辽金方法	神经网络（深度流形）	**无固定几何** 的伽辽金方法，其中几何本身是通过**学习得到的**
算子	隐式数据诱导算子	
试探(检验)空间	学习得到的流形 $\mathcal{M}_\theta$	“**求解**”的含义 **不是**：寻找全局最小值 **而是**：在可接受的流形区域内 实现稳定的不动点一致性
基函数	无属性的激活表示	
弱形式	残差能量积分	
数值求积	小批量采样	
组装	堆叠的分片流形	**训练**在数值意义上求解变分**不动点系统**，推理沿着学习得到的几何结构进行遍历。
数值求解器	反向传播	
收敛	不动点稳定性	

作为可学习数值计算的 AI

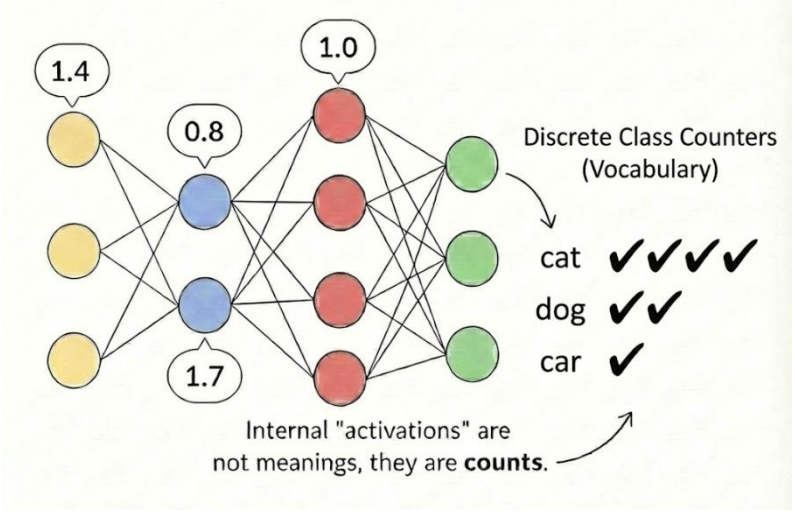
- 神经网络继承了数值计算的世界观:是求解器, 而非定理 证明器.
- 核心张力:如何从领域特定的数 值方法走向数学 层面的普适性
- 在缺乏静态不动点的情况下, 神 经网络不受单一路径约束. 它们在学习到的几何结构中探索、适应并即兴演化
- 数值计算实用主义: 只要能够收敛, 方法并不受限。只要能收 敛, **油盐酱醋**都可以加
- 网络又没有固定的收 敛点, 那就更加**天马行空**, 孙悟空闹天宫, **无法无天**, 八仙过海, **各显神通**

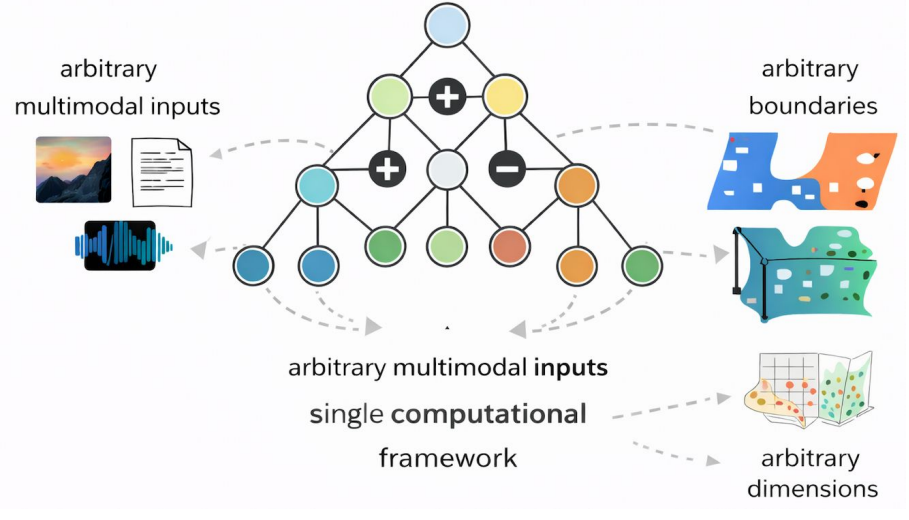
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

神经网络无属性

无属性与计数

- 激活值可以视为隐藏表示或潜在变量，但它们缺乏**可断言**的定义和**内在属性**。
- **分类**的任务在于询问：哪个类别获得了比其他类别**更多**的支持。
- **无属性**：单个激活并不编码语义属性
- **离散**决策兼容性：从连续累积到离散类别选择（argmax）的**计数**过渡是干净的。





OpenAI 并没有错

大型语言模型在根本上是在其所学 习到的表示空间上进行分类。

Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI
Ofir Nachum OpenAI
Santosh S. Vempala† Georgia Tech
Edwin Zhang OpenAI

September 4, 2025

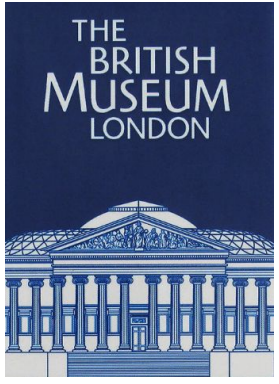
Abstract

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

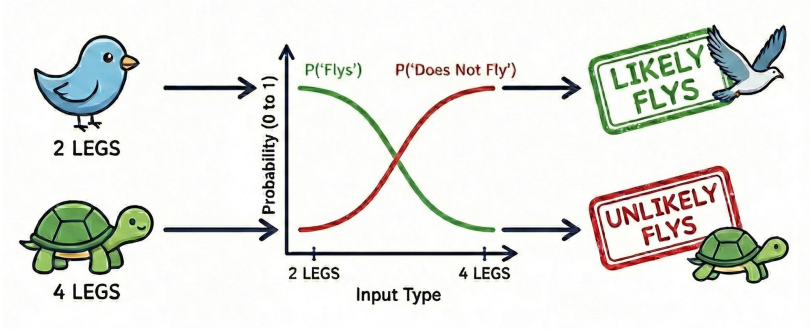
分类

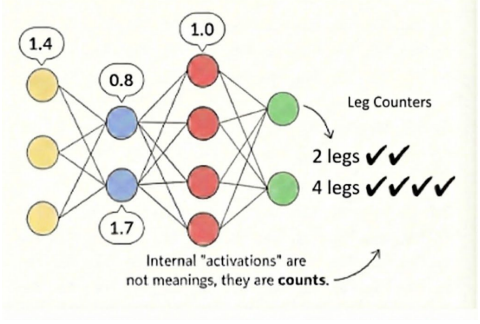
分类是将物品放置在书架上的系统；**编目**是对每件物品的记录，以及在已**分类**物品中定位特定物品的方式

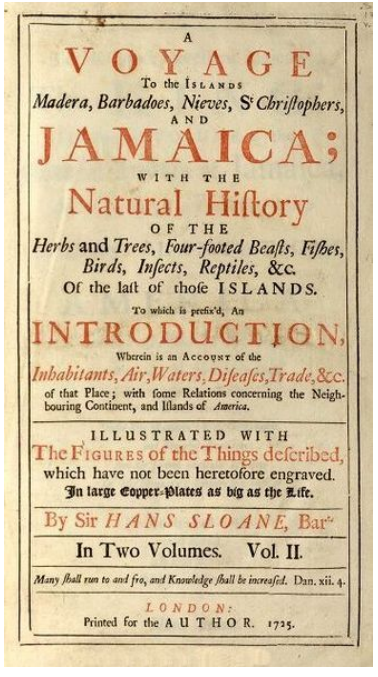
大英博物馆启蒙厅



- 汉斯·斯隆的《牙买加岛野生植物**目录**》(1696年)
- 在启蒙时代，**编目和分类**通过将知识归纳为可**计数**且可比较的特征，进而开启了理性推理的可能性。
- **两条**腿的动物通常会飞，**四条**腿的动物一般不会飞。
- 神经网络遵循类似的原理：通过聚合激活**计数**来进行分类，仅此而已。






A
VOYAGE
To the ISLANDS
Madera, Barbadoes, Nieves, S^t Chriſtophers,
AND
JAMAICA;
WITH THE
Natural Hiſtory
OF THE
Herbs and Trees, Four-footed Beaſts, Fiſhes, Birds, Inſects, Reptiles, &c.
Of the laſt of thoſe ISLANDS.
To which is prefix'd, An
INTRODUCTION,
Wherein is an Account of the
Inhabitants, Air, Waters, Diſeaſes, Trade, &c.
of that Place; with ſome Relations concerning the Neighbouring Continent, and Iſlands of America.
ILLUSTRATED WITH
The FIGURES of the Things deſcribed,
which have not been heretofore engraved.
In large Copper-Plates as big as the Life.
By Sir HANS SLOANE, Bar^t
In Two Volumes. Vol. II.
Many ſhall run to and fro, and Knowledge ſhall be increaſed. Dan. xii. 4.
LONDON:
Printed for the AUTHOR. 1725.

无属性与计数

- 如果我们仔细考察 Transformer 模型的，就会发现它们的运作本质上是分类性的
- 词汇表（vocabulary）界定了离散的类别空间，每一个激活则代表计数过程的一个局部阶段
- 每一层都作为一个积分算子（integral operator），其作用是聚合证据并将其向前传递

$$h_\ell(\xi) = \int_{M^{(\ell-1)}} K_\ell(\xi, \eta)\, \sigma\big(W_\ell h_{\ell-1}(\eta)\big)\, d\mu_{\ell-1}(\eta)$$

- 对这些积分算子进行堆叠，便可实现完整的 L 层流形积分:

$$h_L(\xi_L) = \int_{M^{(0)}} \cdots \int_{M^{(L-1)}} \left(\prod_{\ell=1}^{L} K_\ell(\xi_\ell, \xi_{\ell-1}) \right) \Phi(p)\, d\mu_{L-1}(\xi_{L-1}) \cdots d\mu_0(\xi_0)$$

- 下一标记(next token)的类别决策，则通过将最终流形表征积分穿过类别特定的计数场来获得:

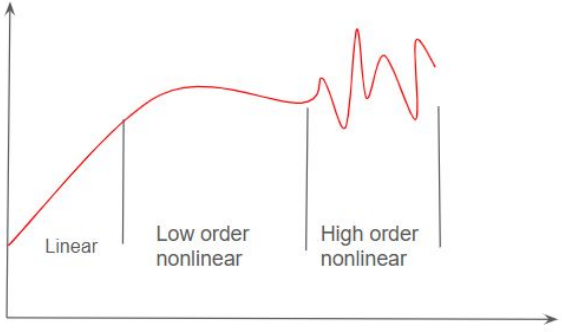
$$z_c = b_c + \int_{M^{(L)}} \theta_c(\xi)\, h_L(\xi)\, d\mu_L(\xi), \qquad p(c \mid p) = \frac{e^{z_c}}{\sum_{k \in V} e^{z_k}}$$

Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

数据

高阶非线性数据

美国国旗：条纹之间（白/红）的颜色跳变，以及白色星星与蓝色背景之间的跳变，都属于锐利的颜色突变。这种 abrupt changes 可以被视为高阶非线性（high-order nonlinearities）。



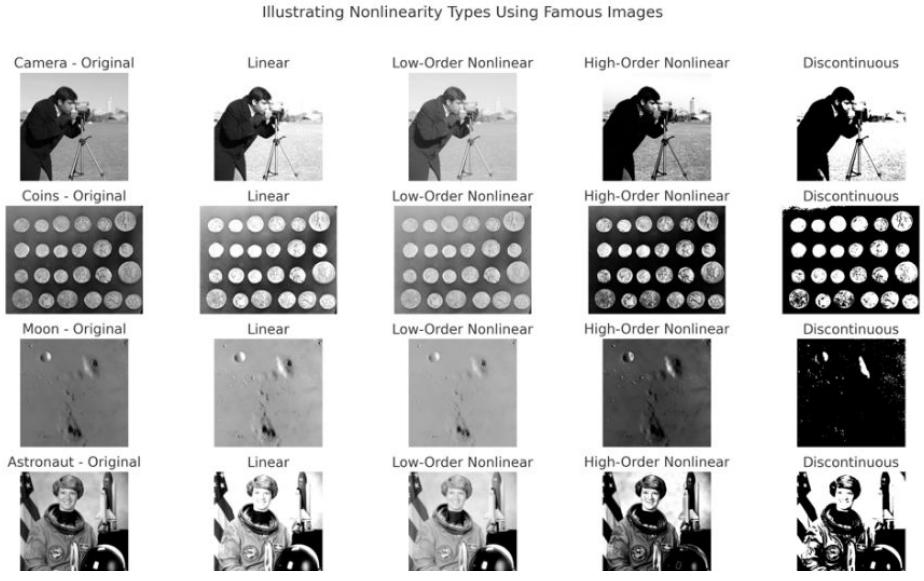
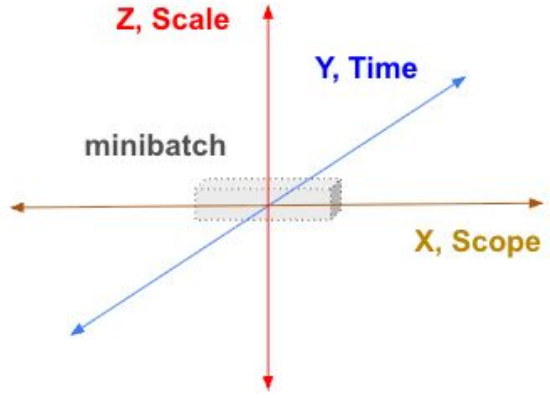


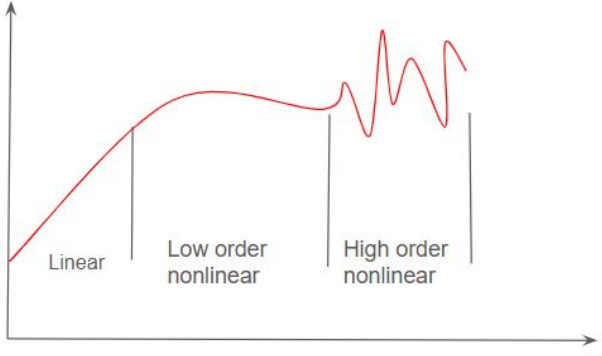
Table 3: NLP Nonlinearity

Type	Example Sentence	Relationship Description
Linear	"More sugar makes it sweeter."	Direct, proportional relationship between input and output.
Low Order Nonlinearity	"A little wine relaxes, too much ruins the night."	Smooth, curved effect — like a quadratic or saturating response.
High Order Nonlinearity	"I never said she stole the money."	Meaning changes based on multi-token interaction or emphasis.
Discontinuous	"Not bad" means "good."	Small token change causes sudden semantic shift or inversion.

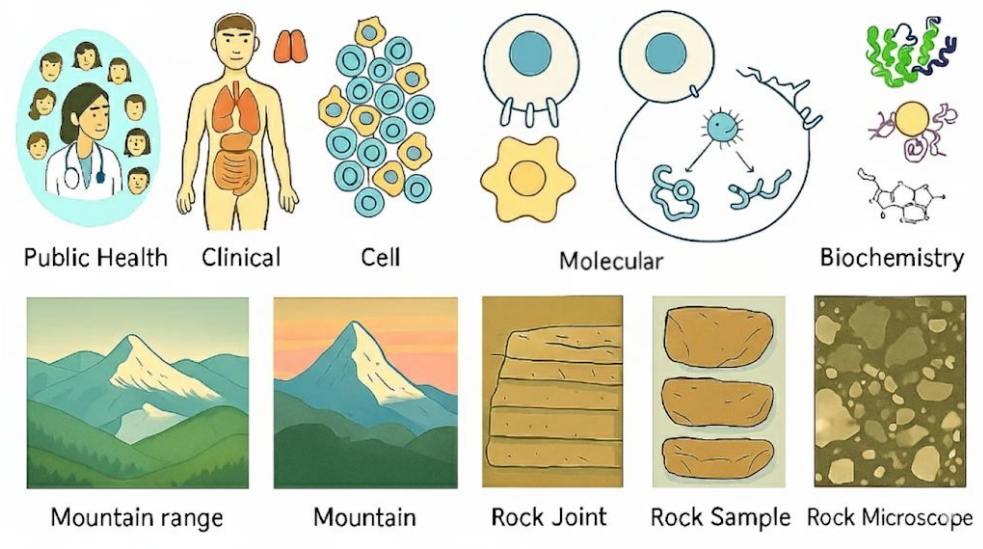
数据复杂性

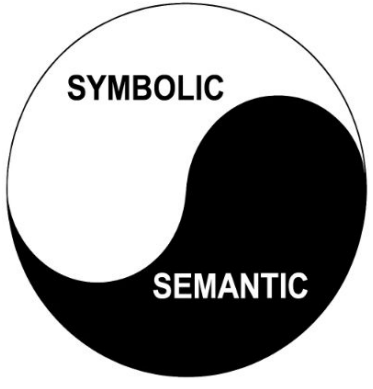


小批量训练是神经网络计算中定点数收敛的一个主要挑战



数据中的高阶非线性是神经可塑性的主要来源





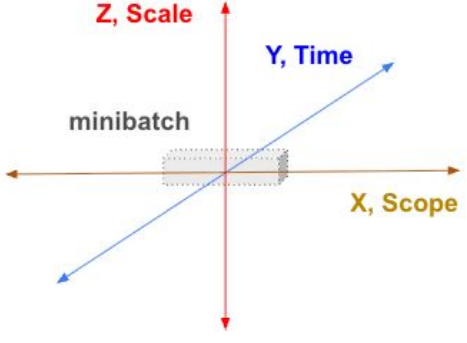
符号 vs. 语义是一个尺度问题

Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

作为不适定反问题的时空失配

- 神经网络是正向与逆向过程的结合体，这使其能够从过去(的数据)中进行学习
- 负时间是反问题的核心属性.
- 大语言模型中的时空失配源于时间的缺失。由于缺乏 显式的代码/时间戳边界条件，模型推理会在不同 时代之间产生混叠
- 神经网络的“无属性”特征非但没有解决 时间歧义，反而将其进一步放大
- 迁移路径
  - 训练数据中的显式时间戳。这使数据流形恢复了方向感
  - 提示词层面的时间边界条件

'How many World Cups has Argentina won?', the correct answer is three, as Argentina won in 1978, 1986, and 2022. However, if the majority of the training data for an LLM predates 2022, the model might incorrectly answer with two. The explicitly timestamp was missing from the training data, or weight much less in the model.



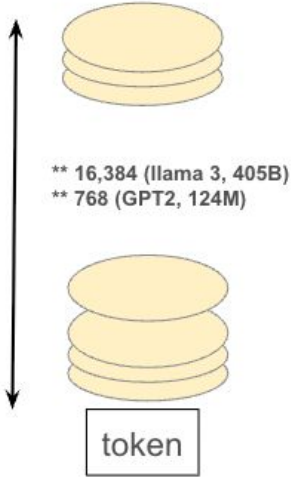
Trusting Your Evidence: Hallucinate Less with Context-aware Decoding, W. Shi et al. 2023

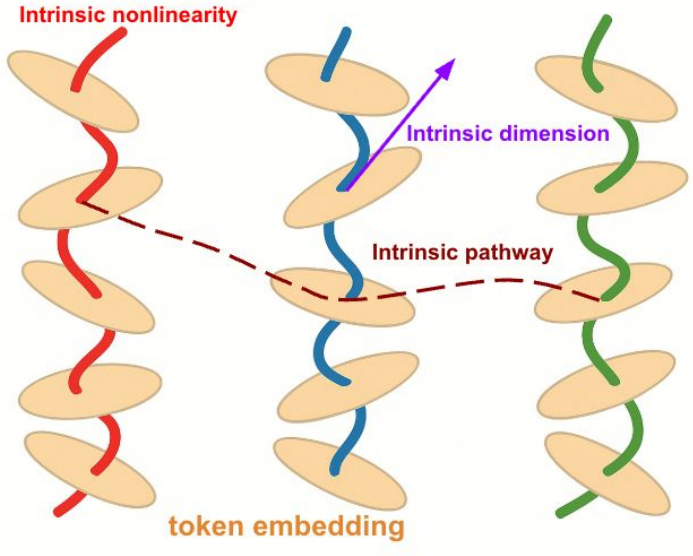
时空失配并非泛化 误差，而是正向 -逆向学 习系统中缺失 显式时间边界条件的 结果。从“深度流形”视角来看，神 经网络并非泛化失效，而是收敛到了由数据决定的随机不 动点。尽管网 络的前向-逆向特性使其能整合 历史信息，但由于 训练中缺乏显式时间戳，时间未被编码为约束边界，从而 导致了系统性的时空失配

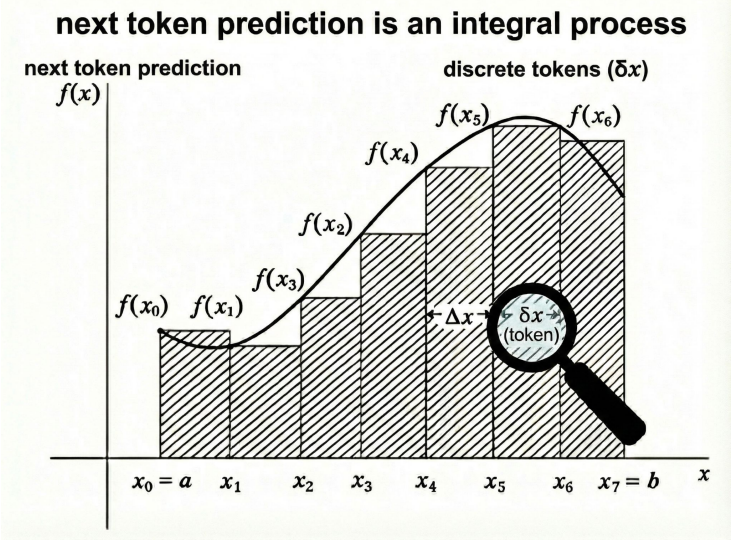
推理

下一标记预测是一个积分过程

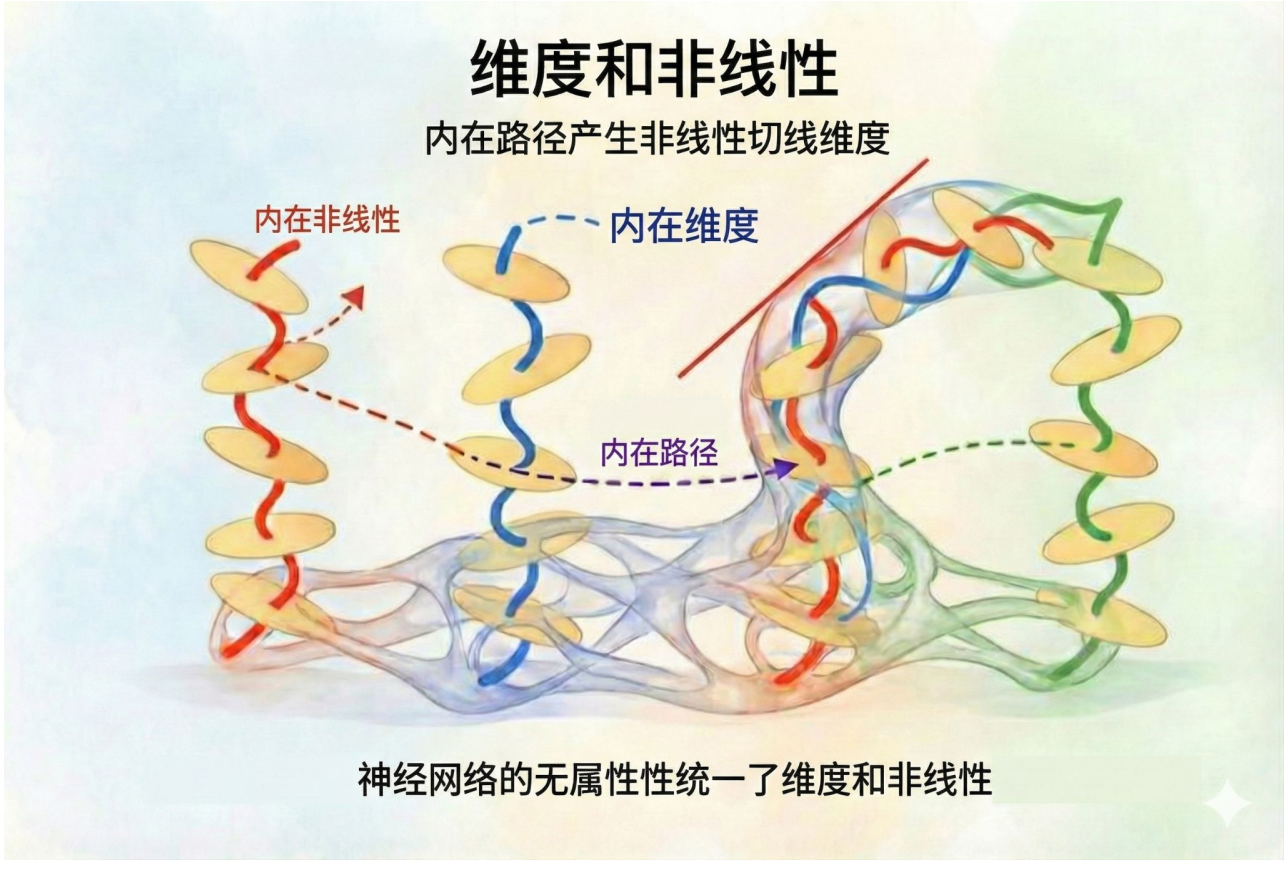
- 令牌嵌入的每个元素定义了一个局部分段流形.
- 一个令牌嵌入是由多个分段流形堆叠而成，从而定义了其固有非线性
- 固有维度与嵌入元素之间固有非线性诱导的切空间相对应，正如维度理论（代数）中所定义
- 固有路径是该场的一条积分曲线: 数百万、甚至潜在数十亿条这样的路径同时共存。
- 神经网络的无属性性使得上述一切成为可能。







在没有 积分的情况下取 导数，导数的意 义何在？

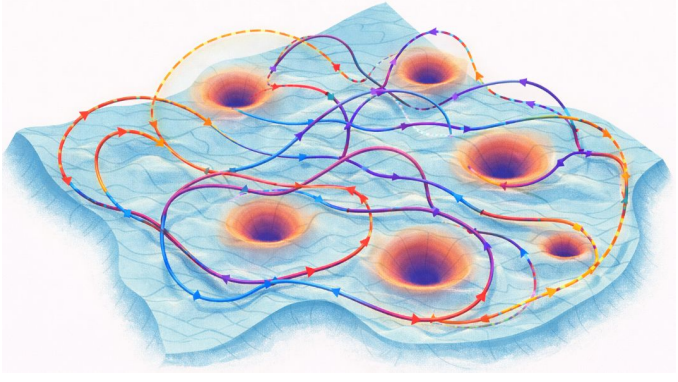


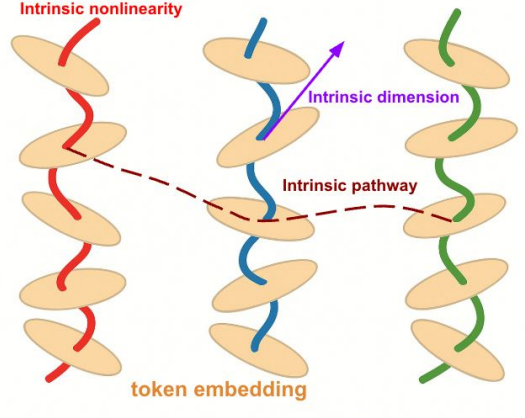
- **非线性**：由每一个 token 嵌入的元素所定义，体现为分段流形结构
- **维度性**：非线性的切空间（维度定理的代数表述）

看似的维度压缩，实则是沿着内在非线性进行的内在维度提取。

推理复杂性: 动态不动点与积分路径

- 推理是对庞大积分路径族的同时遍历
  - 提示词 / 指令定义了积分的初始边界条件
- 学习得到的几何结构中共存着大量不动点
  - 每一次 token 预测对应一次局部收敛
- 推理过程中的不动点是动态的
  - 上下文更新 ⇒ 边界条件变化 ⇒ 不动点漂移
- 推理过程可由积分场的傅里叶展开级数近似表示
  - 在不改变已学习几何结构的前提下, 实现显著加速与成本降低

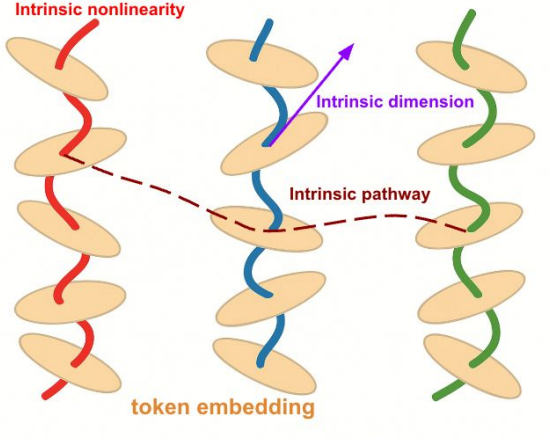


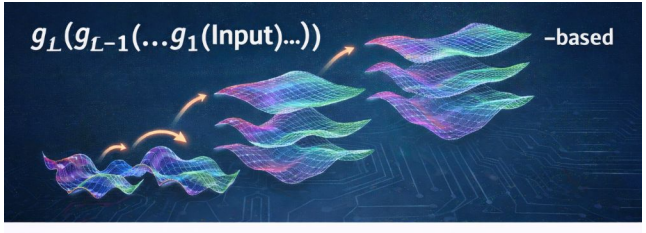


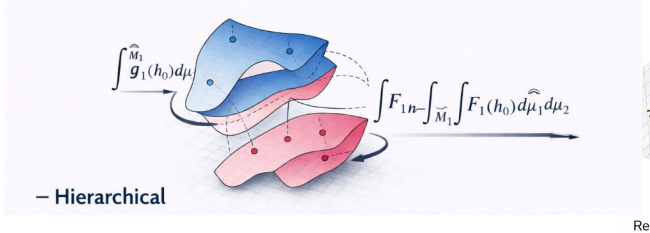
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

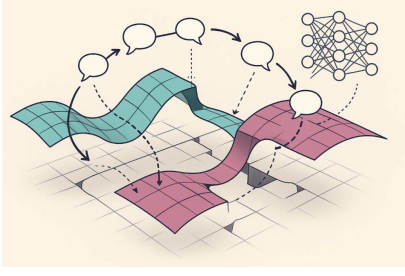
多轮、层级化与递归式

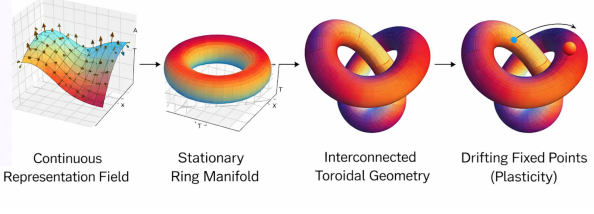
- **多轮（Multi-Turn）**：神经网络内部可涌现出数量极其庞大的内在路径，其规模可能高达万亿级。
- **层级化（Hierarchical）**：神经网络以函数复合为基础，在堆叠流形之上执行嵌套且反复迭代的积分运算。
- **递归式（Recursive）**：神经网络会发展出相互连接的环状、类环面（Toroidal）的几何结构。











Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

大语言模型中的因果推断

1. 大语言模型已经表现出因果行为

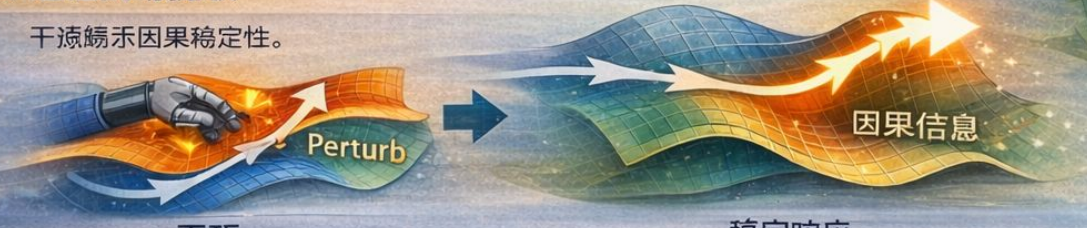
从数据中学习模式，而非符号化因果图。

2. 并非发生在有序空间中

因果关系存在于分布式流形之中。

3. 通过扰动提取

干预揭示因果稳定性。



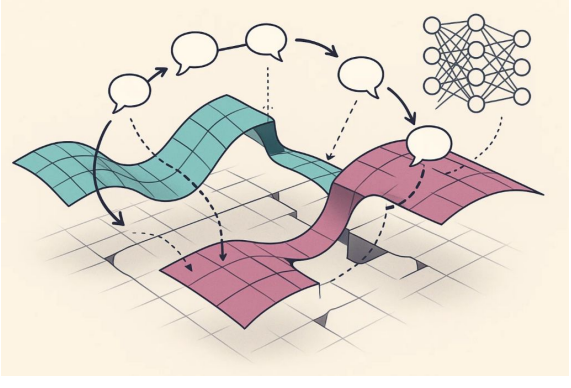
因果性源自在扰动下保持不变的结构

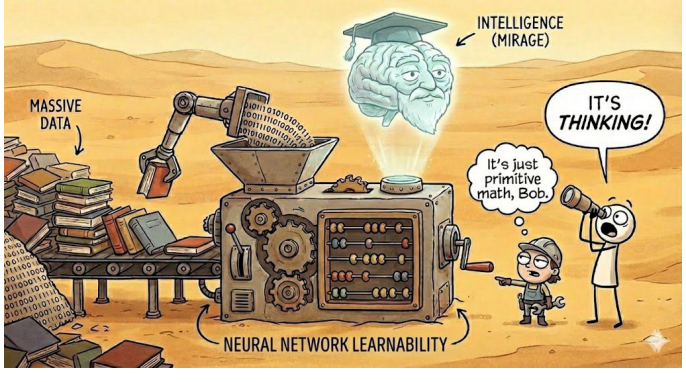
深度流形 · 第二部分. 神经网络数学 arXiv:2512.06563

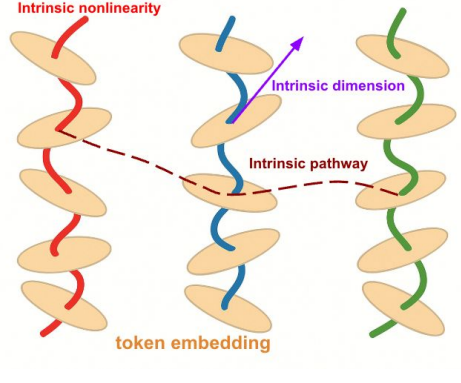
神经网络：一种强大的学习网络

无属性性 神经网络的计算可还原为最原始的计数操作（加减. 正是这种**无属性性**，将一切统一到同一个计算框架之中： • 任意多模态输入得以统一； • 任意边界条件得以统一； • 插值与外推得以统一； • 维度与阶数得以统一。	**堆叠的分片流形** 能够降低数据中的高阶非线性，并生成多条收敛路径。 **坐标变换** 最简单的基函数形式，并通过自身的不断变化来学习数据（即"数据拟合"） **前向，反向联合迭代** 在每一次迭代中实现坐标变换，从而提升数据的学习效率

- 神经网络的可学习性：神经网络本身并不具备内在的推理或认知能力；这些能力完全是通过模型的可学习性，从数据中学习而来的。
- 规模定律在很大程度上仍停留在经验层面，缺乏统一的理论基础；尤其是缺乏对学习效率的原则性刻画，并在很大程度上忽略了学习复杂性。



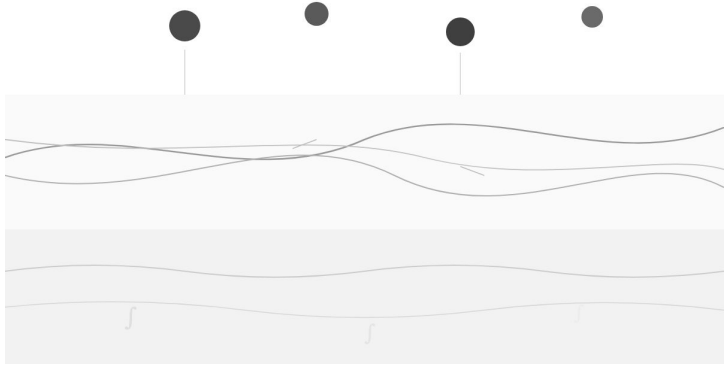




Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

涌现与复杂性

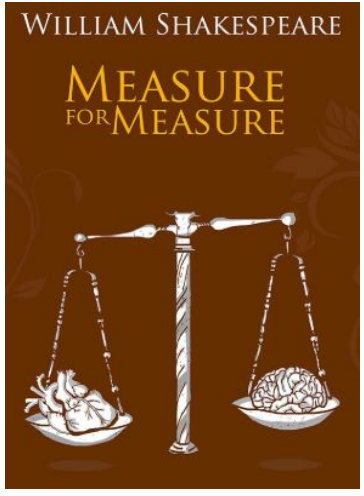
- 涌现与复杂性是不可分割的
  - 复杂性是涌现的基底(如土壤)，涌现是其可观察的结果(如果实)。复杂性的提升会扩展并孕育出更丰富、更优越的涌现行为。
- 深度流形将其归因于：
  - 将涌现与复杂性归因于“无属性”(property-lessness)
  - 这是神经网络如此强大的根本原因
- 从分类到涌现：以无属性为基础
  - 分类先于解释，其根基在于数学的无属性。
  - 大英博物馆的“启蒙展厅”将分类视为科学的起点。



Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

《一报还一报》

有巨人的力量固然好，但是像巨人那样使用这种力量便是残暴 (莎士比亚)



神经网络是一种强大的可学习数值计算模型。其最显著的优势之一在于‘**无属性**’。 这种**无属性**赋予了神经网络从多模态混合数据中学习几乎任意内容的能力，并实现跨学科的迁移。然而，它并**未锚定于**任何具体的物理规律、已知的**科学原理、物理 维度与时间维度**， **道德框架或常 识**。其忠实度与可信度，仅仅取决于训练数据的品质。水能载舟，亦能覆舟

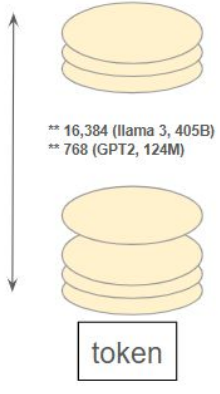
神经网络：一种强大的学习网络

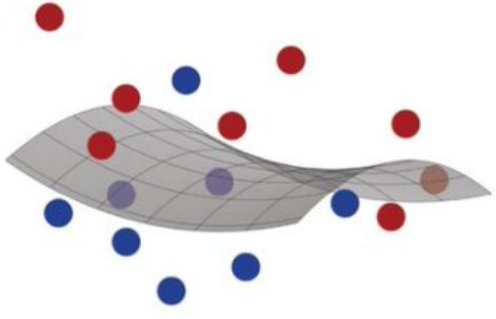
无属性性 神经网络的计算可还原为最原始的计数操作（加减. 正是这种**无属性性**，将一切统一到同一个计算框架之中： • 任意多模态输入得以统一； • 任意边界条件得以统一； • 插值与外推得以统一； • 维度与阶数得以统一。	**堆叠的分片流形** 能够降低数据中的高阶非线性，并生成多条收敛路径。
	坐标变换 最简单的基函数形式，并通过自身的不断变化来学习数据（即“数据拟合”）
	前向，反向联合迭代 在每一次迭代中实现坐标变换，从而提升数据的学习效率

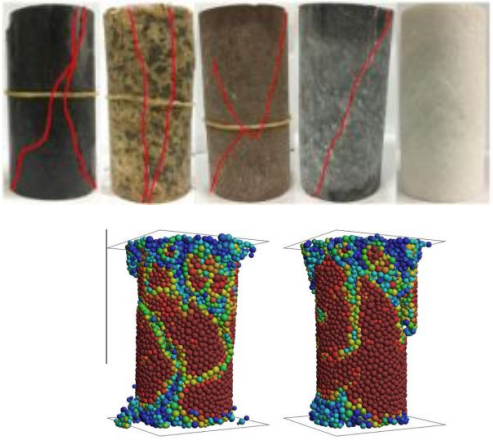
训练进程

Transformer 位置嵌入

- 这是 Transformer 的一项设计败笔/缺陷
  - Transformer 天然地运行在叠加的、分片流形之上。由于嵌入 维度极高，token 表征在实际计算中趋近于点状存在，如同一粒粒沙子。这使模型对数据中的高阶非线性具有极强的适应能力，同时在整体上促进了流形的平滑性（局部 连续、全局跃迁）与计算稳定性。但这种特性也会导致计算预算的迅速膨胀，并最终形成缩放上限，其本质与我们在离散元方法（DEM）中观察到的有问题的绑定机制在精神上是相通的。
- Transformer 有两个根本性的方面，长期以来并未被 AI 社区明确认识到



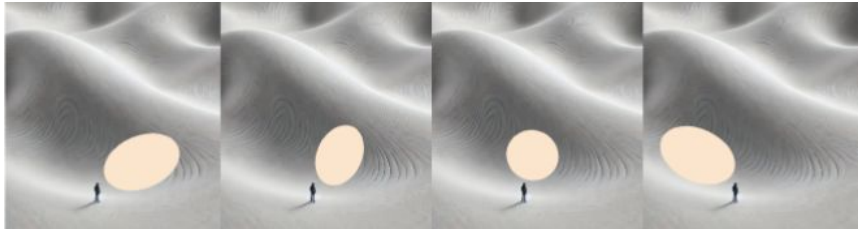


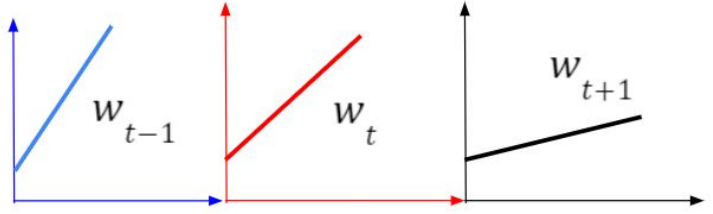


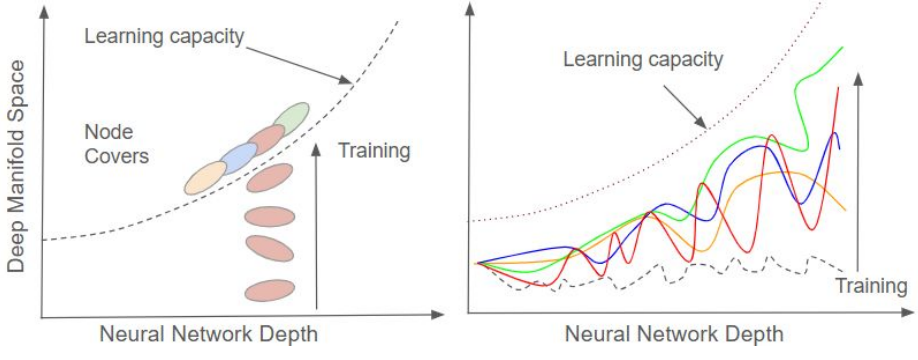
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

持续学习

- **为什么这是可能的**：神经网络的节点坐标可以在迭代过程中无限制地变化——这是任何经典数值方法或计算框架从未具备过的能力。
- **是什么在制约它**：神经可塑性以及动态漂移的不动点限制了系统稳定性，使得长时间尺度的持续学习在本质上变得困难。







Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

强化学习(RL)与不动点扰动

- 如果缺乏不动点，迭代过程将无法依赖内部几何结构，只能完全由边界条件来引导。
- 在缺乏对称边界的情况下，迭代过程只能依赖于弱的、软性的、离散的边界条件。
- “弱与软”：数值层面的扰动极小，但在整体尺度上系统仍保持高度的可适应性。
- 离散性：即便边界条件是离散的、稀疏的或非严格的，网络仍然能够重构其内在的积分轨迹，并定位新的不动点。

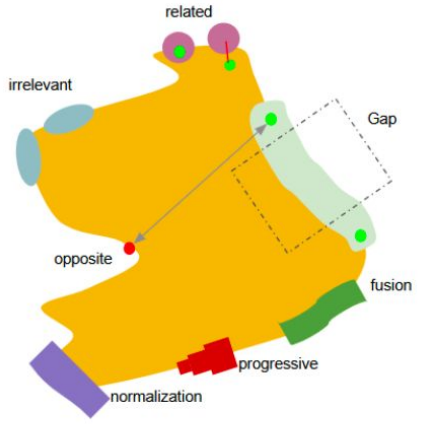


Fig. 14: Symmetric Boundary Condition

Then all three backward-pass regimes unify under a single fixed-point objective:

$$\theta^* \in \arg\min_{\theta} \mathbb{E}_p \Big[\alpha\, KL(p_{\text{data}} \| \, q_\theta) + \beta\, C(\Phi_\theta(p)) + (1-\alpha-\beta)\, \ell(q_\theta, y) \Big] \qquad (71)$$

Fig. 15: Weak and Discrete Boundary Condition

Table 2: Three stages of backpropagation Fixed Point Iteration.

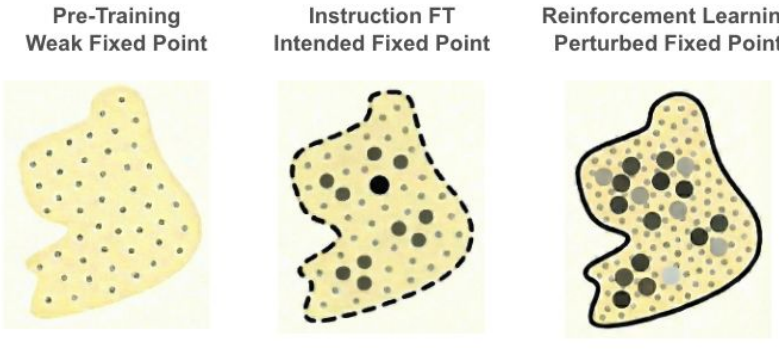
#	Stage Name	Boundary Type	Iteration Type
0	Pre-Training	Implicit boundary	Weak fixed-point iteration
1	SFT	Smi-Structured boundary	Intended fixed-point iteration
2	RL	Explicit boundary	Perturbed fixed-point iteration

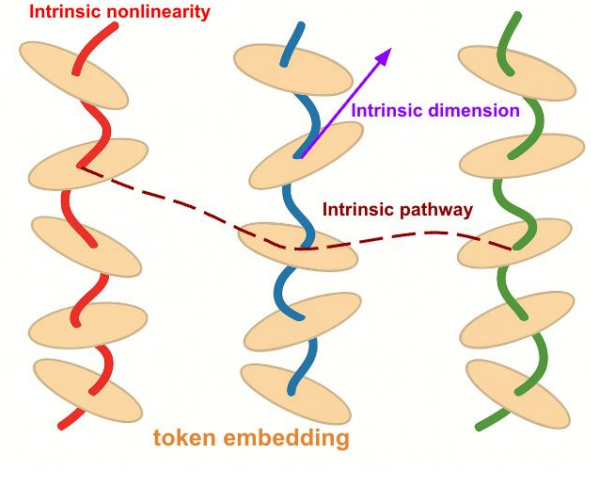
Fig. 13: Foundation Model Boundary Conditions

Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

固定点演进

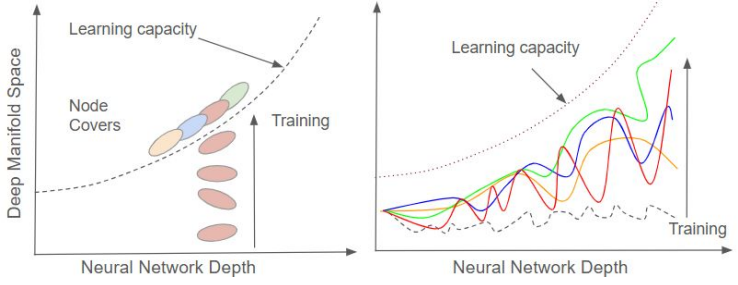
- 在实践中，一个固定点对应于我们所称的解、答案或决策
- 弱固定点并不意味着该点本身是弱的；相反，其固有收敛路径是弱的且弥散的。
- 意图固定点（Intended Fixed Point）充当锚点，通过显式边界约束来塑造收敛盆地。
- 扰动固定点（Perturbed Fixed Point）往往是最有效的，因为弱的且随机的边界条件能够激活更丰富的收敛路径
- 在高阶非线性条件下，扰动对于固定点迭代的收敛变得至关重要。
- 强化学习的有效性源于对称、弱且离散的边界条件，这些条件延迟了神经可塑性。

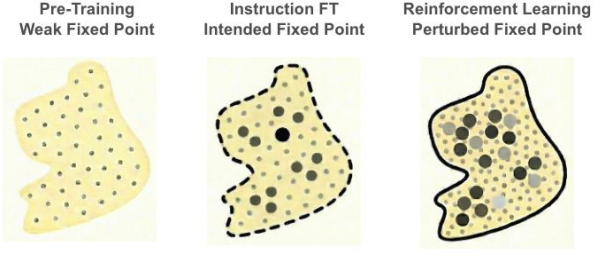




可练性、可学性与神经塑性

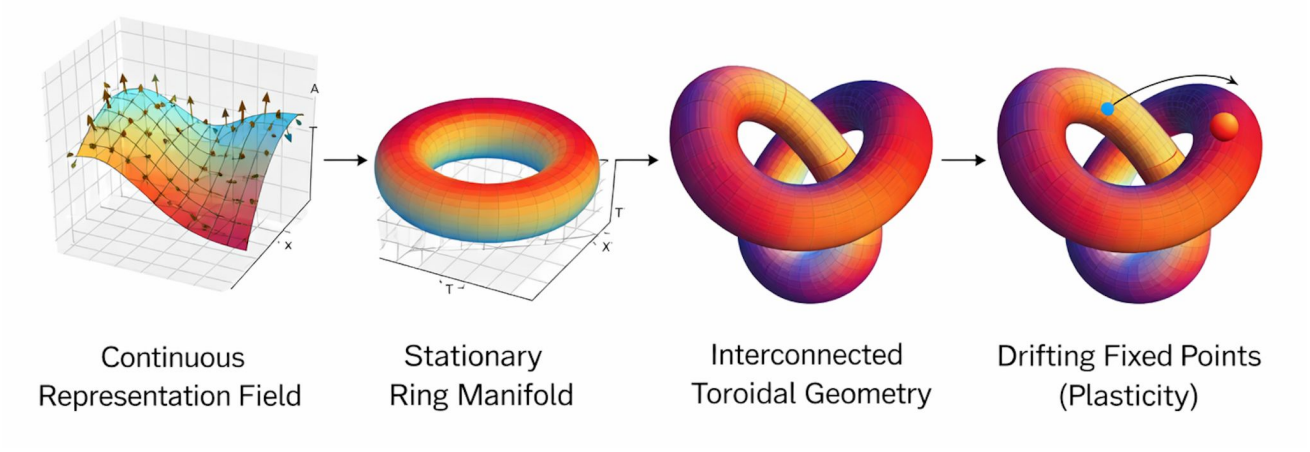
- 通过坐标无限制变化实现无界可训练性
  - 因为节点坐标(流形覆盖)在每次迭代中都会 发生变化，神经网络可以被无限期训练
- 可练性≠ **可学性**
  - 持续优化可能无法产生新的学习，反而会导致与先前学习到的结构发生错位。
- 神经塑性是瓶颈
  - 延迟神经可塑性是架构设计与训练策略中的一个关键目标
- **学习复杂性**
  - 数据诱导的复杂性:高阶非线性与近乎无限的数据范围
  - 架构诱导的复杂性:刚性神经网络架构
  - 边界诱导的复杂性:不对称的、强制的或点式边界条件
  - 优化诱导的复杂性:批量结构与学习率调度





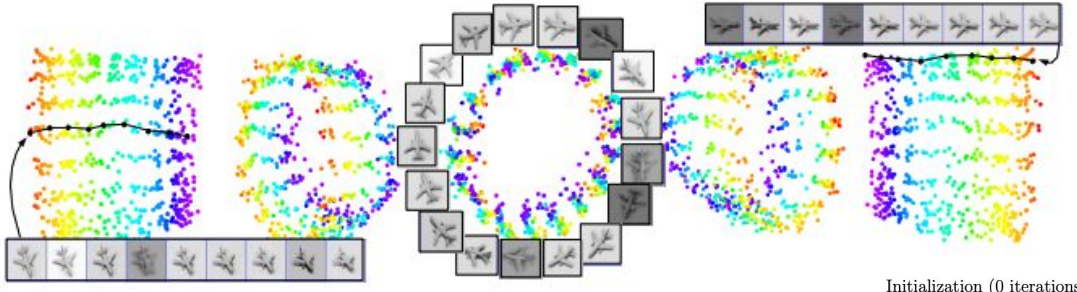
定常性、对称性与神经塑性

- **定常性（stationarity）**产生于真实世界数据施加了大量互不相容的约束；在随机训练下，这些约束在统计上相互抵消，从而产生退化方向（degenerate directions），沿这些方向变分梯度（variational gradient）消失.
  - 深度流形中的定常性正是定义神经固定点的变分条件：神经固定点是在弱的数据诱导边界条件下，隐式拉格朗日量的定常解
- **对称性（环状）**在如下情况下涌现：当这些退化定常方向允许连续变换而保持泛函不变时，导致定常解集合组织成闭合的水平集流形，表现为环状或壳状结构。
  - 互联环面几何结构产生于多个环状定常流形共享重叠数据约束的情形（例如，“bank”具有多重含义），从而将原本独立的环耦合成立体缠绕的环面，并实现语义叠加以及神经可塑性。
- **神经塑性**源于互联环面结构，因为这些结构创造了扩展的、耦合的平坦方向，在这些方向上，微小的边界扰动会使解沿着定常流形移动，而不是将其恢复到唯一的配置。

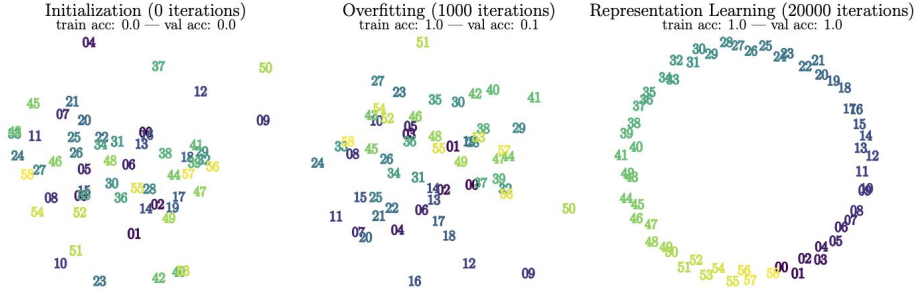


环状是否是最稳定的定常流形？

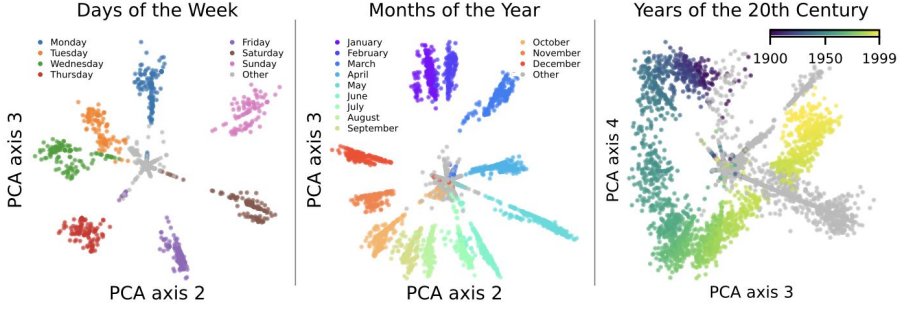
局部上**是**，整体上**否**。环状结构在各向同性约束下局部是最稳定的定常流形，但当多个不变性耦合形成更高阶几何时，全局稳定性会破缺。



Dimensionality Reduction by Learning an Invariant Mapping, R. Hadsell… Yann **LeCun**, 2005



Towards Understanding Grokking: An Effective Theory of Representation Learning. (M. Liu… Max **Tegmark**, 2022)

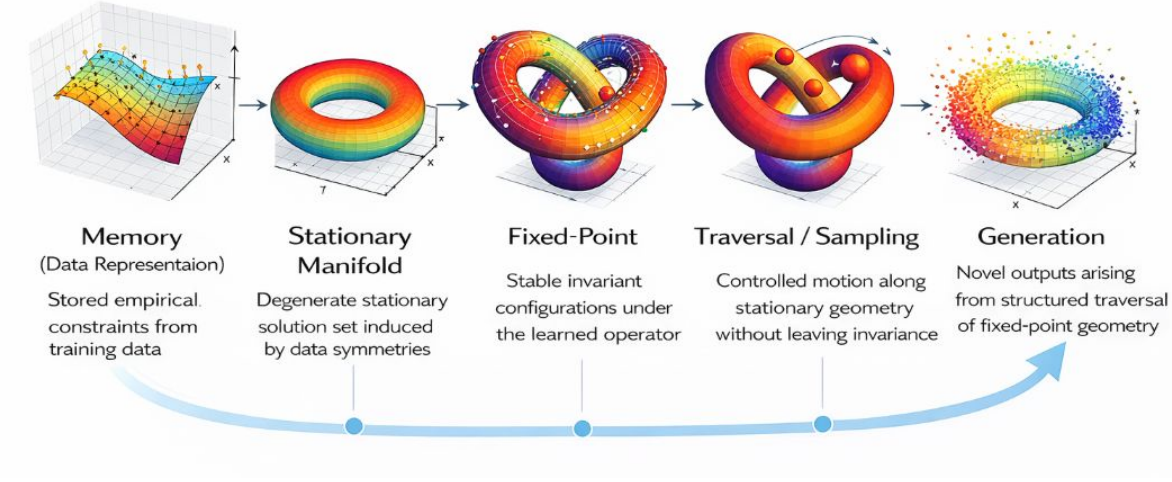


Not All Language Model Features Are One-Dimensionally Linear, (J. Engels… Max **Tegmark**, 2024)

训练复杂度:从样本记忆到结构泛化

深度学习中的几何结构与不动点

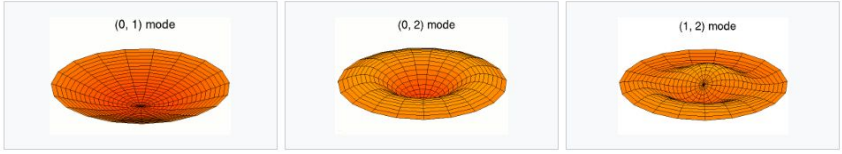
- 训练是几何结构的演化，而非参数优化。
  - 学习过程本质上是重塑可行解的几何空间，而不是最小化某个预先给定的函数。
- 数据施加约束，几何吸收约束
  - 经验数据逐步塑造表示流形，使其演化为满足约束的驻定结构。
- 驻定性定义了“学到什么”。
  - 学习到的知识对应于在数据诱导算子下形成的驻定流形及其不动点集合。
- 泛化是几何上的遍历。
  - 生成源于在不变几何结构内的有序运动，而非对训练样本的简单记忆。

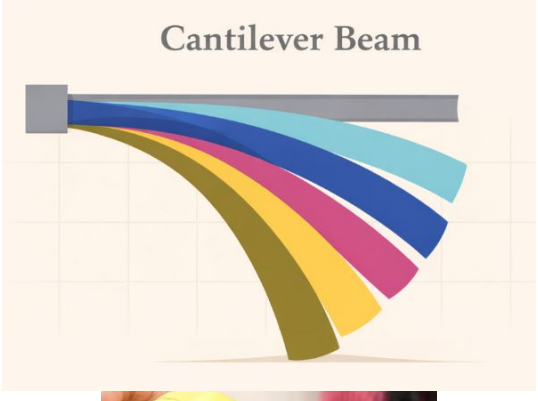


Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

更大的模型并非终极解决方案

- 为什么扩展规模有效
  - 更大的模型增加自由度
  - 支持高阶非线性
  - 吸收近乎无限的数据范 围、规模与多样性
- 为什么单纯扩展规模会失败: Babuška's 悖论
  - 纯粹扩展会引入累积的数值误差
  - 大规模节点交互放大不稳定性
  - 当几何结构漂移时，更多参数 ≠ 更优解
- Transformer：仍然是目前最好的选择
  - 局部刚性强，结构上较脆弱
  - MoE：自由度更高、隐式联邦化，但脆弱性依然存在
- 扩散模型自有其美
  - 全局平滑、鲁棒
  - 计算代价高
- Play-doh 的弹性模型
  - 柔软、大可变形的几何结构
  - 在高阶非线性下保持稳定
  - 随数据复 杂度扩展，而非仅随参数数量扩展

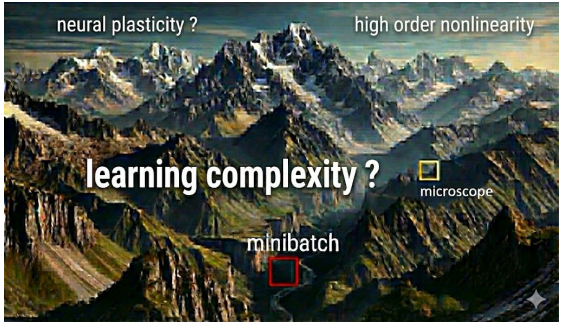




模型 CAP 定理

单次前向推理：在 {覆盖性、准确性、性能} 三者中，最多只能同时优化其中两项

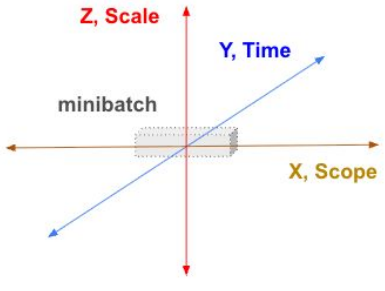
- 覆盖性（Coverage）指模型在语义、符号、时间、模态以及非线性等多个轴向上，试图表征真实世界流形的范围与程度。
- 准确性（Accuracy）指局部几何保真度：包括曲率 对齐、不动点稳定性，以及各个流形切片内的残差正确性。
- 性能（Performance）指推理过程中的数值计算效率。



Concept	Complexity Theory View
Time	How long an algorithm takes (time complexity)
Dimension	Higher input/state space dimension → more computation
Nonlinearity	Nonlinear systems are often harder to analyze and solve

Table 5: A Classification of Function Complexity

Class	Definition	Complexity index $k(f)$	Scaling with input size n
L (Linear)	$Ax + b$	0	$O(n)$
P (Low-order polynomial)	degree k	k	$O(n^k)$
H (High-order nonlinear)	infinite expansion	∞	grows faster than any fixed $O(n^k)$
D (Discontinuous)	piecewise / jumps	$\perp$	exponential partitions possible

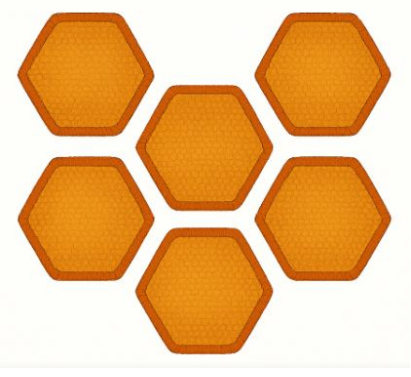


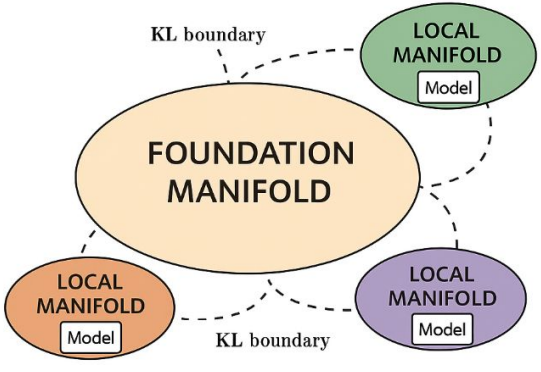
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

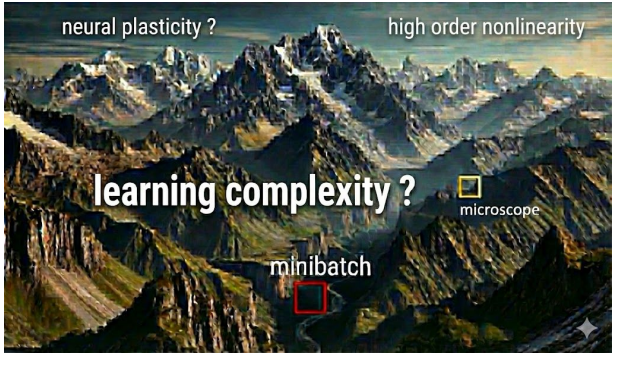
流形联邦:现实世界与世界模型

未来是”迷你”小型 AGI联邦

- 数据复杂性诱导学习复杂性
- 学习复杂性源于数据的高阶非线性
- 联邦本身就是一个流形概念
  - 每个模型对应一个独特的流形
  - 就像机器学习集成(ensemble)一样，每个模型捕捉高阶非线性的不同方面
- 局部联邦以小型弹性模型的马赛克形式运作
- 全球联邦构成了深度流形联邦学习





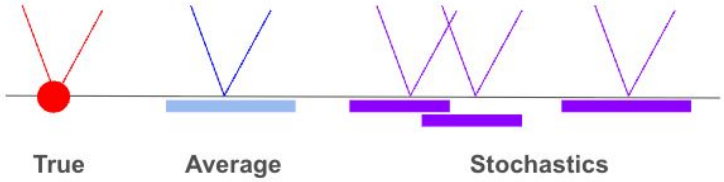


Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

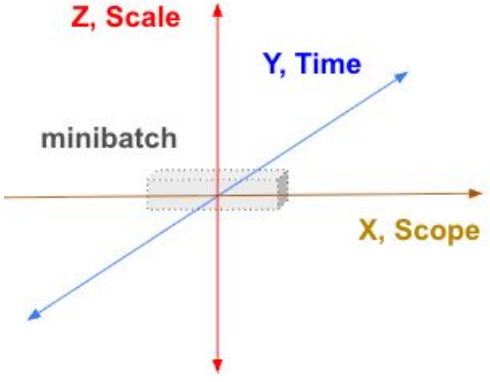
科学与工程中的人工智能

科学与工程中的人工智能

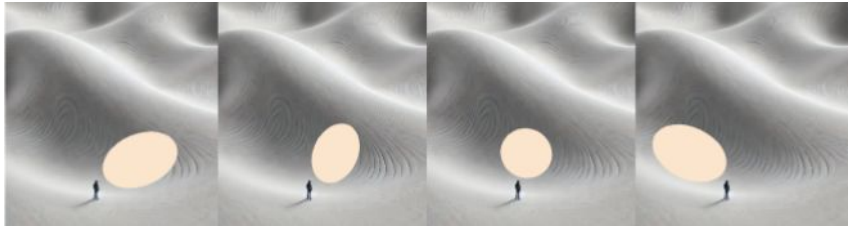
神经网络的“无属性”特性是其通用学习能力的基础



1. 神经网络的极致学习能力
   a. 前向/反向**统一迭代**
   b. **无属性**：适用于任意数据类型与任意边界条件
   c. 损失函数：不依赖任何**物理方程或数学**
   d. **无界学习空间**：**凭借统一的数值表示**：**无属性**（token embedding），在分辨率、领域、数据规模以及多模态上均表现出色
   e. **随机不动点**：比人类主观解释更贴近真实世界

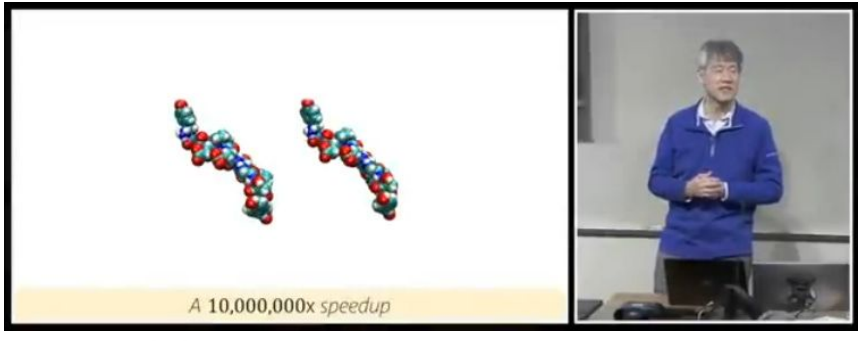


2. 何以实现
   a. **逆问题**求解变得自然且轻松无碍
   b. 施加**任意边界条件**可实现高达一千万倍的加速
   c. **可发现性**：**插值与外推**在流形轨迹上实现融合
   d. **本构建模**：支配关系仅通过观测直接从数据中自然浮现
3. 物理信息神经网络（PINN）与神经算子
   a. 物理信息神经网络与传统数值计算互为补充，二者不可相互替代
   b. 在高度复杂的问题上，神经算子表现有限



AI4SE:加速

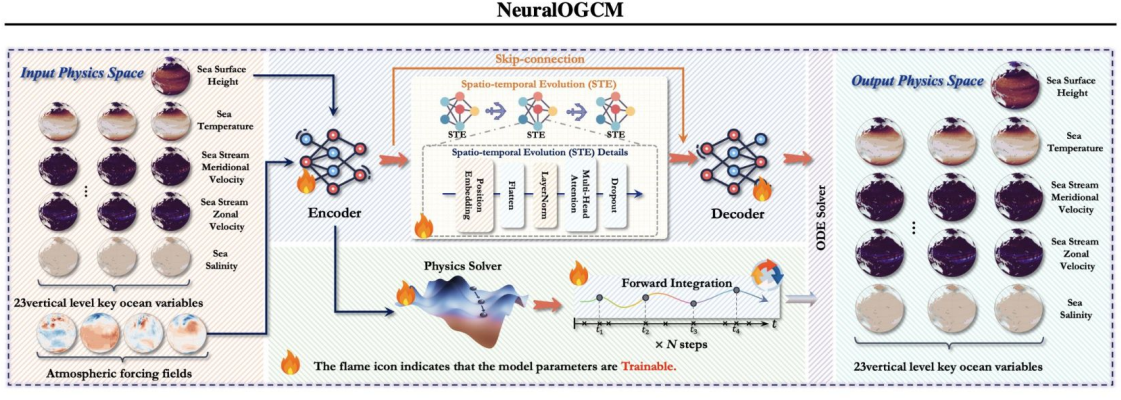
- 任意边界条件: 高达一千万倍
- 显式耦合与隐式耦合



50+ parameters

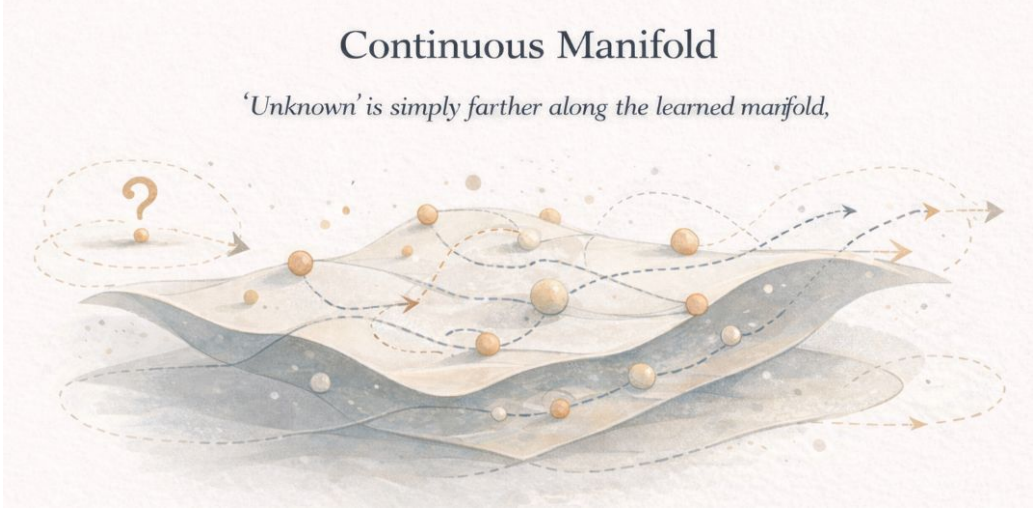
Famous Equations in Physics

1. Schrödinger Equation:
$$i\hbar\frac{\partial\psi}{\partial t} = \hat{H}\psi$$
2. Dirac Equation:
$$(i\gamma^\mu\partial_\mu - m)\psi = 0$$
3. Einstein's Field Equations:
$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}$$
4. Navier-Stokes Equation:
$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) = -\nabla p + \mu\nabla^2\mathbf{u} + \mathbf{f}$$
5. Maxwell-Boltzmann Distribution:
$$f(v) = \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-\frac{mv^2}{2kT}}$$
6. Planck's Law:
$$E = \frac{h\nu}{e^{\frac{h\nu}{kT}} - 1}$$
7. Newton's Second Law :
$$F = \frac{dp}{dt}$$
8. Klein-Gordon Equation:
$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2 + \frac{m^2c^2}{\hbar^2}\right)\phi = 0$$
9. Black-Scholes Equation:
$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$
10. Vlasov Equation:
$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f + \frac{\mathbf{F}}{m}\cdot\nabla_v f = 0$$
11. Fokker-Planck Equation:
$$\frac{\partial f}{\partial t} = -\nabla\cdot(\mathbf{F}f) + D\nabla^2 f$$
12. Landau-Lifshitz-Gilbert Equation:
$$\frac{d\mathbf{M}}{dt} = -\gamma\mathbf{M}\times\mathbf{H}_{\text{eff}} + \frac{\alpha}{M_s}\mathbf{M}\times\frac{d\mathbf{M}}{dt}$$
13. Einstein's Mass-Energy Equivalence:
$$E = mc^2$$
14. Boltzmann Transport Equation:
$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f + \mathbf{F}\cdot\nabla_v f = \left(\frac{\partial f}{\partial t}\right)_{\text{collision}}$$
15. Van der Waals Equation of State:
$$\left(P + \frac{a}{V^2}\right)(V - b) = RT$$
16. Raychaudhuri Equation:
$$\frac{d\theta}{d\tau} = -\frac{1}{3}\theta^2 - \sigma_{\mu\nu}\sigma^{\mu\nu} + \omega_{\mu\nu}\omega^{\mu\nu} - R_{\mu\nu}u^\mu u^\nu$$
17. Langevin Equation:
$$m\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} = \eta(t)$$



可发现性：插值与外推

- 神经网络既没有显式的知识边界，也没有隐式的知识边界
- 插值与外推坍缩为一个统一的过程：流形遍历
- 发现是如何涌现的：
  - 近无限的数据覆盖范围在持续扩展
  - 无属性表示使得坐标重塑不受任何限制
  - 受边界条件约束的迭代引导运动沿固有路径进行
- 神经网络不会跨越知识边界，因为根本不存在知识边界
- 外推不过是稳定的流形遍历，直至约束失效
- 发现是几何必然性，而非符号推理



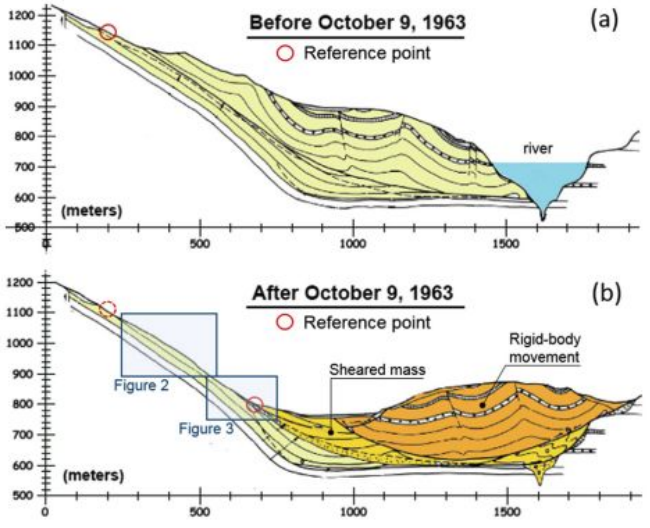
数据驱动的本构模型构建

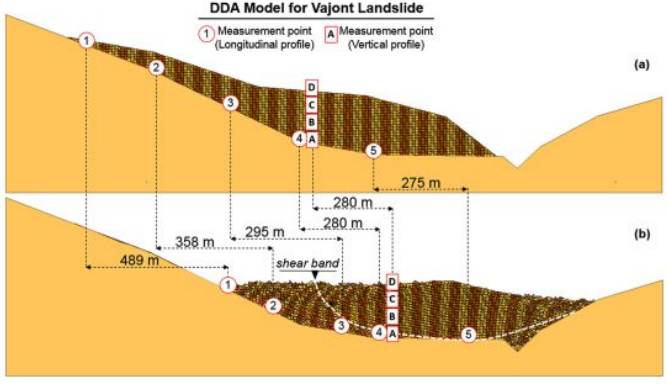
- 本构定律 = 本构流形
  - 物理系统遵循固有关系(如应力–应变关系)
  - 于高阶非线性情形中，不存在闭合解析形式的本构定律
  - 神经网络将这种关系学习为一个本构流形
- 从方程形式到几何描述
  - 定律转变为通过学习获得的几何结构，而非预先规定的公式
  - $\sigma = f(\varepsilon)\,(\varepsilon,\sigma)\in\mathcal{M}_\theta,\ \mathcal{M}_\theta=\{(\varepsilon,f_\theta(\varepsilon))\}$
- 固定点解读
  - 本构更新被视为一个固定点问题
  - 训练以最小化残差为目标，进而恢复内在本构定律
  - 学习得到的行为以数值固定点的形式趋于稳定
- 投影至学习得到的流形
  - 观测点被投影到 $\mathcal{M}_\theta$
  - 本构更新通过选取最近的可容许本构状态实现
  - 对噪声、可塑性以及不完整观测具有稳健性

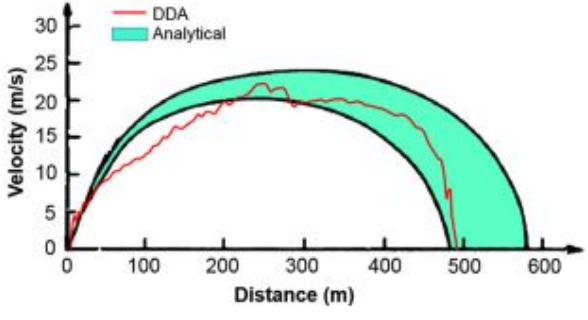
数值计算的局限性

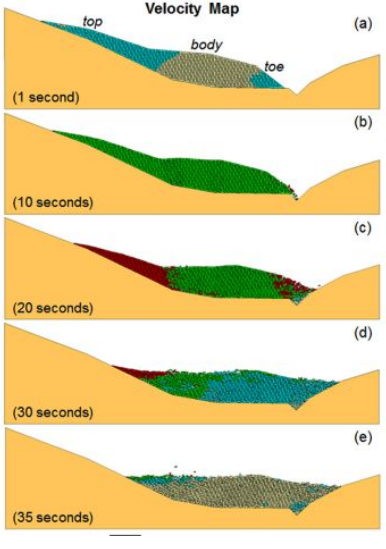
- 方程是被假定的;现实世界并非如此
  - 经典求解器依赖已知的控制算子;而真实系统往往是部分未知、持续漂移或尚未建模的。
- 几何是固定的，而非可学习的
  - 网格 / 基函数 / 坐标系的选择来自外部设定。当现实需要新的表示结构时，求解器无法生成新的表示几何。
- 不确定性被视为噪声，而非边界条件
  - 随机性通常以附加项的形式引入（误差条、湍流模型），而不是作为塑造可解空间的一等约束加以整合
- 复杂性被外包给“闭合项”
  - 当物理机制缺失时，往往引入摩擦因子、本构关系、子网格模型或手工补丁，但这些做法难以扩展
- 缺乏可扩展的多工况（多区间）转变机制
  - 不连续性、相变、损伤与分岔迫使求解器切换、重新划分网格或重写模型。

Vajont Landslide, 1963





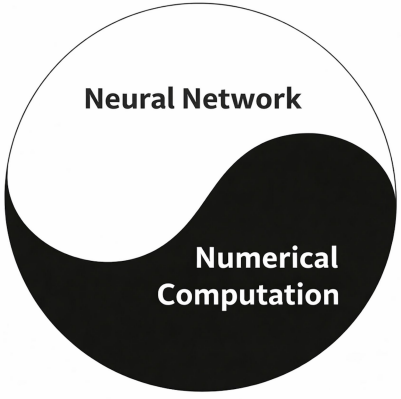




- Rapid sliding and friction degradation: Lessons from the catastrophic Vajont Landslide, J. Ibañeza,, Y. Hatzorb, 2018, Engineering Geology
- Discontinuous Deformation Analysis (DDA)

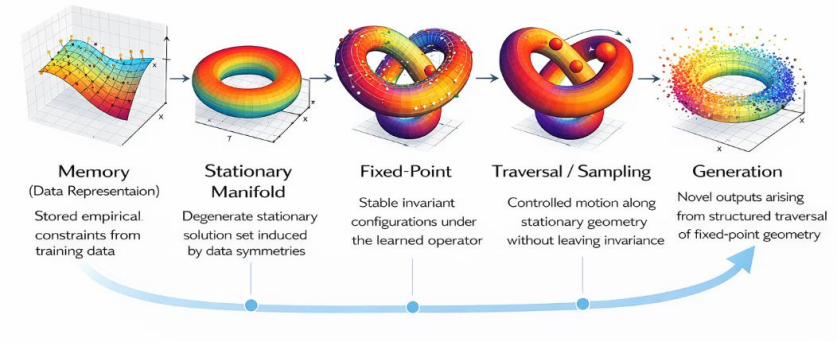
耦合方法

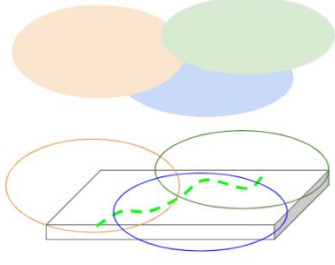
把全局学习得到的几何结构耦合为局部场地上的数值真解



- AI 擅长全局泛化
  - 从所有历史数据与模拟案例中学习
- AI 是无属性的（propertyless）
  - 能够处理异构数据（图像、文本、信号、日志等）
- 以物理原理为基础的求解器，如 DDA, FEM 与 NMM
  - 天然支持反问题、不连续性以及强非线性问题

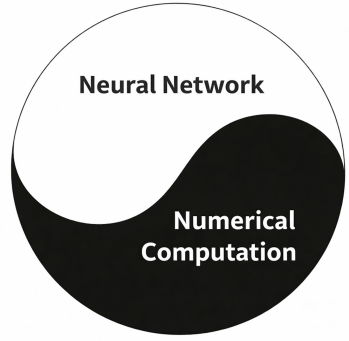
作为可学习数值计算的神经网络，与以物理原理为基础的数值计算彼此互补，而非相互替代。

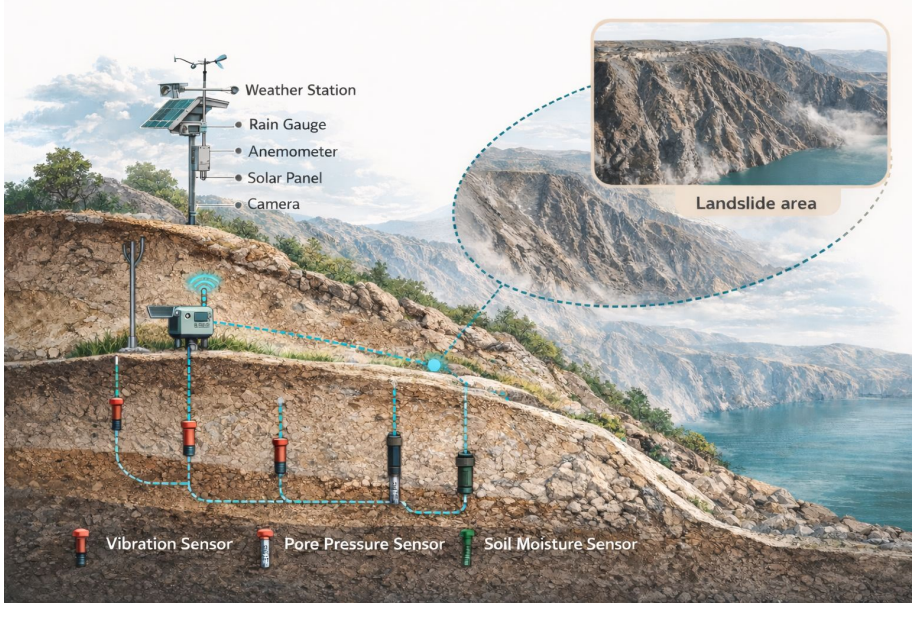


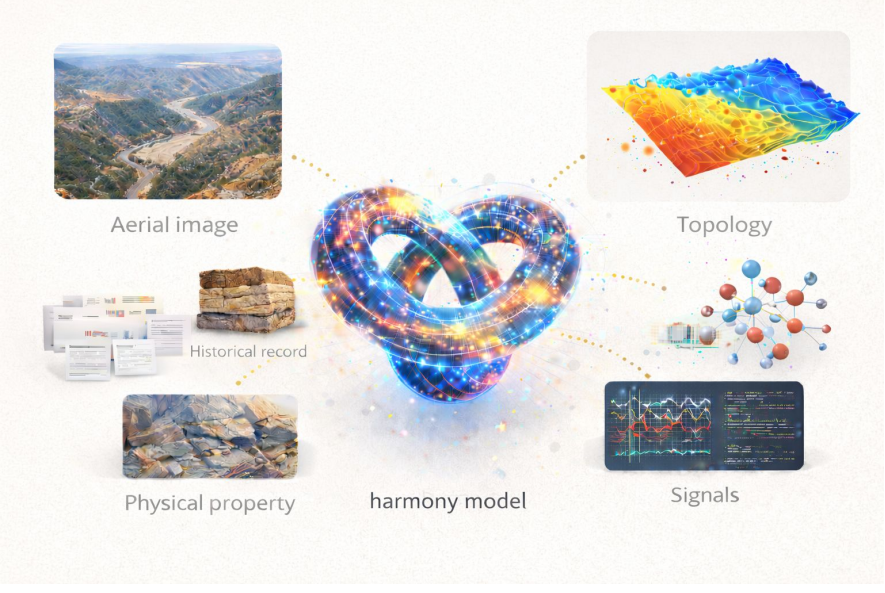


和谐模型

- 神经网络
  - 跨数据类型的可学习性与全局泛化能力
- 数值计算
  - 以物理为基础的精确性、真实不动点与显式机制
- 数值计算揭示机制；神经网络学习几何







不是巧合

Part 2 was officially published on December 6, 2025

人工智能联邦

深度流形

- 深度流形视角下，由于数据的高阶非线性特性，未来必然走向小型AGI 的联邦化。

Zoom AI sets new state-of-the-art benchmark on Humanity's Last Exam

Federated innovation driving breakthrough results in complex AI testing

Updated on December 10, 2025
Published on December 10, 2025

As CTO of Zoom, I'm excited to share a significant milestone in our AI journey. Today, we're announcing that Zoom has achieved a new state-of-the-art (SOTA) result on the challenging Humanity's Last Exam (HLE) full-set benchmark, **scoring 48.1%, which represents a substantial 2.3% improvement** over the previous SOTA result of 45.8% by Google Gemini3-pro with tool integration.

Xuedong Huang
Chief Technology Officer

Model/System	HLE Full Set Score
OpenAI GPT-5 Pro w/ tools	42.0%
Anthropic Claude Opus 4.5 w/ tools	43.2%
Google Gemini 3 Pro w/ tools	45.8%
Zoom Federated AI	48.1%

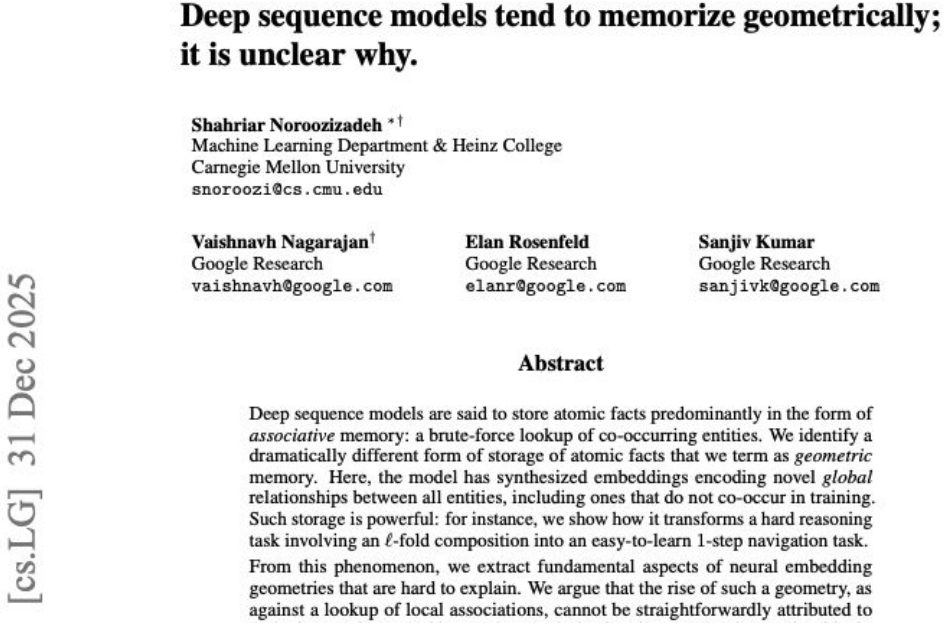
The winning strategy: Federated excellence

Our SOTA performance on Humanity's Last Exam stems from both powerful models and a new approach to their application. Central to our success is our effectively guided explore-verify-federate strategy, an innovative agentic workflow that optimally balances exploratory reasoning with rigorous verification. Instead of generating extensive reasoning traces, our method strategically identifies and pursues the most informative and accuracy-enhancing reasoning paths.

神经网络的几何结构

深度流形

- 神经网络的行为本质上由其学习到的世界表征几何所决定



[cs.LG] 31 Dec 2025

Deep sequence models tend to memorize geometrically; it is unclear why.

Shahriar Noroozizadeh *†
Machine Learning Department & Heinz College
Carnegie Mellon University
snoroozi@cs.cmu.edu

Vaishnavh Nagarajan†
Google Research
vaishnavh@google.com

Elan Rosenfeld
Google Research
elanr@google.com

Sanjiv Kumar
Google Research
sanjivk@google.com

Abstract

Deep sequence models are said to store atomic facts predominantly in the form of *associative* memory: a brute-force lookup of co-occurring entities. We identify a dramatically different form of storage of atomic facts that we term as *geometric* memory. Here, the model has synthesized embeddings encoding novel *global* relationships between all entities, including ones that do not co-occur in training. Such storage is powerful: for instance, we show how it transforms a hard reasoning task involving an ℓ-fold composition into an easy-to-learn 1-step navigation task.

From this phenomenon, we extract fundamental aspects of neural embedding geometries that are hard to explain. We argue that the rise of such a geometry, as against a lookup of local associations, cannot be straightforwardly attributed to

神经提示词重复

深度流形

- 推理是一种迭代积分过程；当积分路径规模达到数十亿时，这一架构早已在现实系统中运行，无需区分所谓的推理或非推理模型

Prompt Repetition Improves Non-Reasoning LLMs

Yaniv Leviathan* Google Research leviathan@google.com

Matan Kalman* Google Research matank@google.com

Yossi Matias Google Research yossi@google.com

Abstract

When not using reasoning, repeating the input prompt improves performance for popular models (Gemini, GPT, Claude, and Deepseek) without increasing the number of generated tokens or latency.

1 Prompt Repetition

无位置嵌入

Deep Manifold

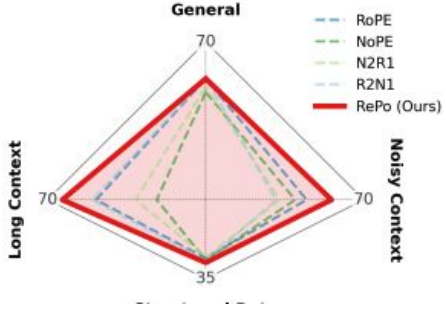
- 位置嵌入在严格意义上并非必要，其持续被采用的原因在于它们在高维嵌入空间中起到结构性粘合作用，为模型提供稳定的结构锚点。

REPO: Language Models with Context Re-Positioning

Huayang Li [1 2] Tianyu Zhao [1] Richard Sproat [1]

Abstract

In-context learning is fundamental to modern Large Language Models (LLMs); however, prevailing architectures impose a rigid and fixed contextual structure by assigning linear or constant positional indices. Drawing on Cognitive Load Theory (CLT), we argue that this uninformative structure increases extraneous cognitive load, consuming finite working memory capacity that should be allocated to deep reasoning and attention allocation. To address this, we propose REPO, a novel mechanism that reduces extrane-



学习能力

深度流形

- 神经网络是一种可学习的数值计算

数据主义

数据主义(数据本体论)

- **核心论断**
  - 数据主义认为，智能与知识源自对数据的忠诚，而非源自人类设计的意义、意图或结构。
- **数据主义学说原则**
  - 数据作为首要权威
    - 数据本身是最根本的知识来源。它天然具有高阶非线性特征，并覆盖近乎无限的范围，远非任何人类设计的规则或理论所能穷尽。
  - 神经网络作为可学习的数值计算
    - 神经网络应被理解为一种可学习的数值计算体系，而非受制于既有物理定律、道德框架或符号化原则的系统。模型的行为不是被预先规定的，而是完全由数据所学习得到的。
  - 数学基础
    - 数据主义的数学结构建立在三大支柱之上:流形覆盖，不动点理论与微积分。它们共同刻画了学习系统如何表征复杂性、实现行为稳定，并通过计算不断演化。
  - 学说的认识论克制
    - 作为一种学说，即一种思想流派，数据主义并不主张自身在定义上天然正确。它并非因宣称而成立，而仅在其持续解释并预测经验现象的能力中获得正当性。
  - 人类知识的角色
    - 数据主义并不排斥人类洞见或领域知识在特定现实问题中的作用，而是将其视为情境性的支撑结构，而非普遍且至上的权威来源。

mHC and Muon

$$L(\theta, \lambda) = \mathbb{E}_x \left\| f_\theta(x) - x \right\|^2 + \lambda\, g(\theta)$$

- 不完全符合数据主义(Dataualism)的教义立场。这类方法引入了显式的结构性约束，记为 g(θ)
- 这并不是说这些方法是错误的或无效的，而是它们体现了一种不同的哲学立场。
- 这将限制由强重复信号驱动的跨阶段梯度对齐，从而避免在微调及其后的训练过程中几何结构过度塌缩。

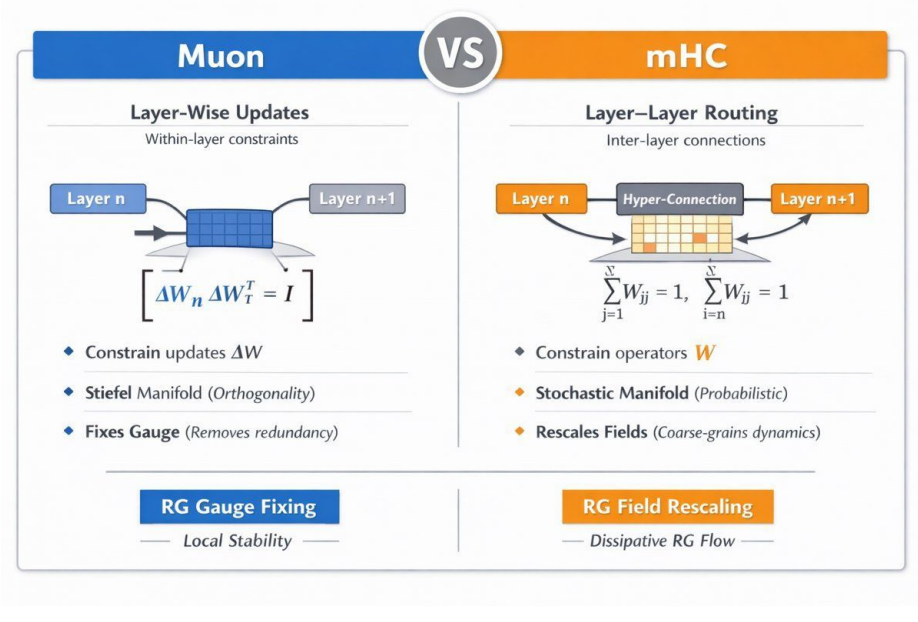


Image source
@CalcCon

在深流形的视角下

信息论是一种边界约束理论

从深度流形视角看熵、压缩与学习

$$\mathcal{J}(\theta) = \underbrace{\int_{\mathcal{M}_\theta} \ell(f_\theta(x), y)\, d\mu(x)}_{\text{learning on stacked nonlinear manifolds}} + \underbrace{\lambda \left(- \int_{\mathcal{M}_\theta} \mu_\theta(x) \log \mu_\theta(x)\, dx \right)}_{\text{entropy as boundary functional}}$$

- 熵 ≠ 学习目标 → 它是一种容量边界
- 压缩 → 流形对齐的结果，而非其原因
- 学习 → 几何约束相互抵消所形成的随机不动点
- 为何最小描述长度(MDL)在大语言模型中失效 → 描述长度衡量的是边界代价，而非内部几何结构
- 真正重要的是 → 深度流形上的内在学习路径

信息论限制的是表征能力；深度流形理论解释的是学习过程。
熵限定边界；几何给出解。

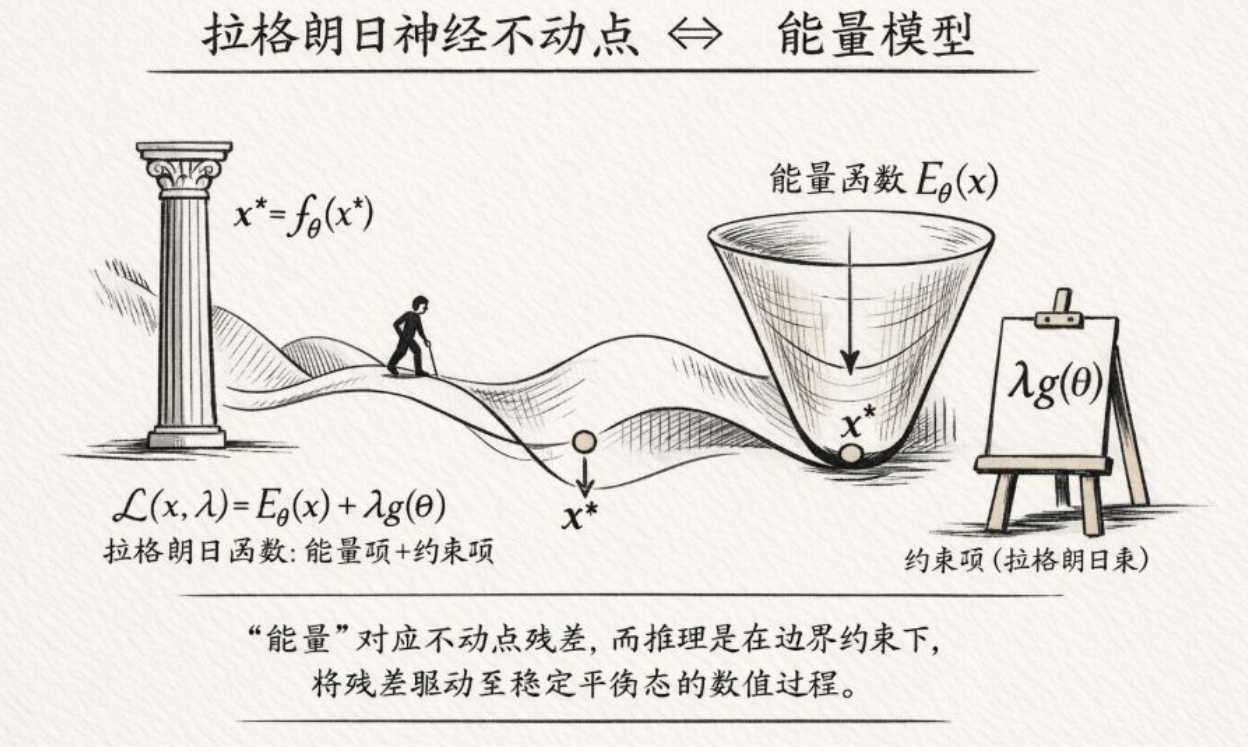
柯尔莫哥洛夫复杂度:一种优美的信息描述理论

- 柯尔莫哥洛夫复杂度所提出的问题是:
  - 生成该对象所需的最短程序是什么？
- 深度流形理论所提出的问题是:
  - 在怎样的几何结构与动力学约束下，该对象能够作为一个可学习数值系统中的稳定不动点而存在？
- 柯尔莫哥洛夫理论:压缩产生结构
- 深度流形理论:结构先于压缩，压缩只是结果
- 神经网络并不最小化描述长度
  - 它们通过整合约束，使残差在局部逐渐消失。
  - 压缩只在稳定的流形几何结构形成之后才会自然出现。

柯尔莫哥洛夫复杂度刻画描述的上界;深度流形理论揭示计算的机制。
熵约束可达的边界条件;几何结构决定系统最终收敛到的解。

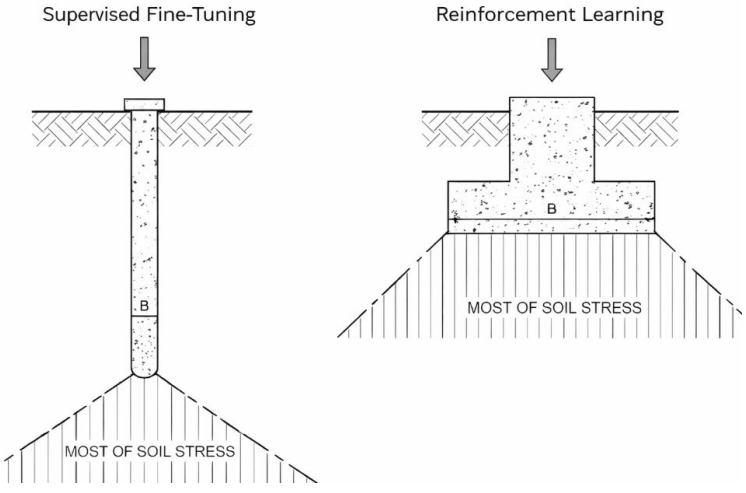
能量模型

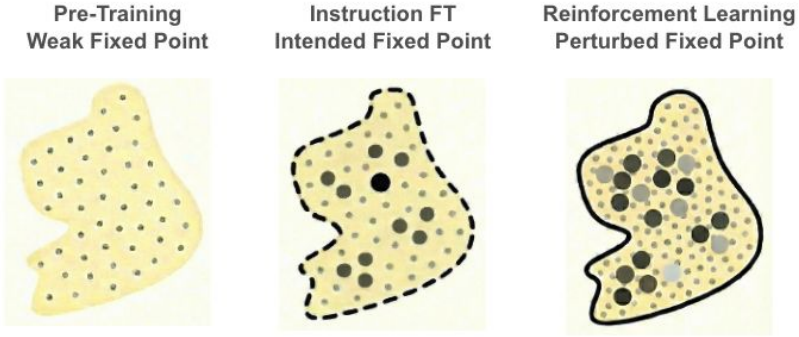
能量模型本质上是用“能量语言”表述的神经不动点系统：其中，“能量”对应不动点残差，而推理过程则是在边界约束下，通过数值计算将该残差驱动至稳定平衡态。

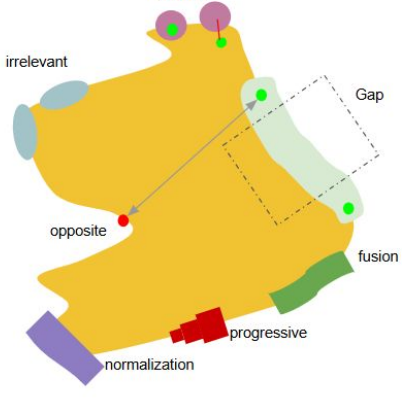


SFT: 灾难性遗忘

从深度流形(Deep Manifold)的视角来看, 预训练, 监督微调(SFT)和强化学习(RL对应着不同的边界条件。
SFT 并非内在地导致灾难性遗忘, 通过恰当的实现方式, 可以避免灾难性遗忘的发生。

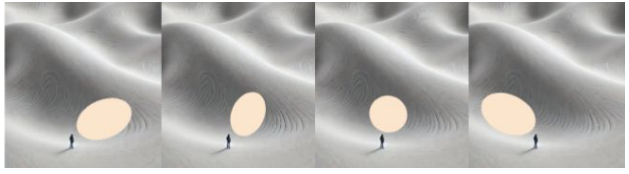


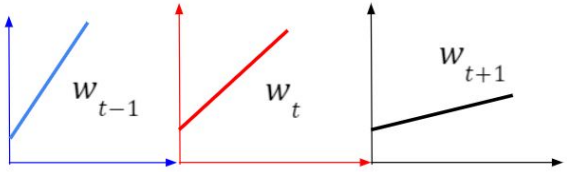


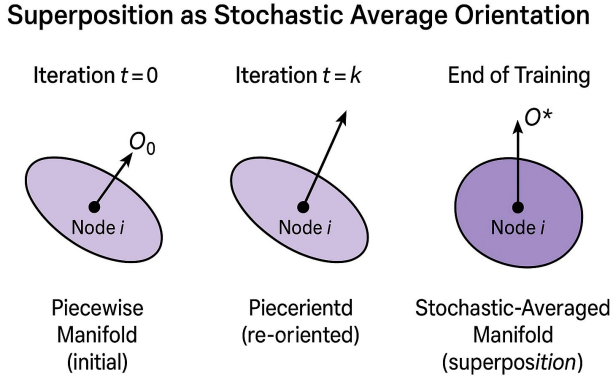


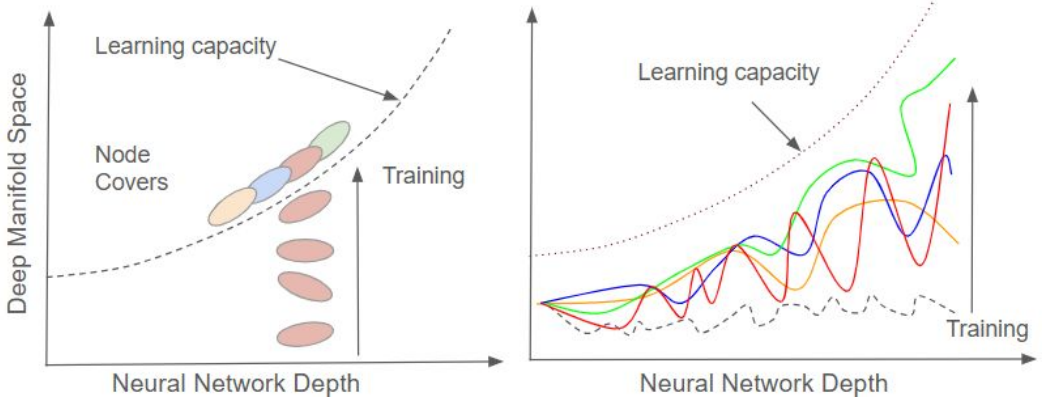
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

叠加，节点覆盖，神经可塑性









- 每一次迭代都会旋转或重塑节点的局部流形取向
- 训练 = 跨越堆叠流形的持续节点覆盖重定向
- 叠加 = 分片流形取向的随机平均
- 学习容量（缩放定律）受神经可塑性所限制
- 在训练过程中，堆叠的分片流形内部会形成相互连通的环面结构
- 这些结状的环面结构是神经可塑性的几何来源

Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

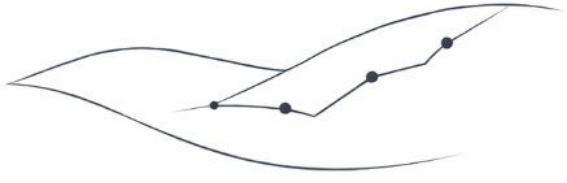
机械可解释性

电路(Circuit)

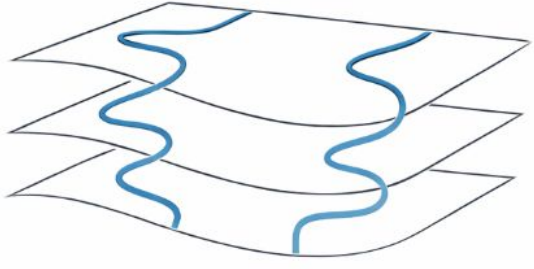
- 单一、平滑、连续的表面
- 一条细的离散路径在表面上行 进
- 路径上有节点或点
- 几何是平坦且全局的

深度流形(Deep Manifold)

- 多个堆叠，轻微错位的表面(分片流形）
- 一条连续曲线向下/穿行于各层之间
- 路径在层边界处发生弯折(积分,而非跳跃）
- 无节点，无符号，只有流动
- 通过层间间距暗示深度，而非明暗阴影



Content: Circuit — Path on a single surface

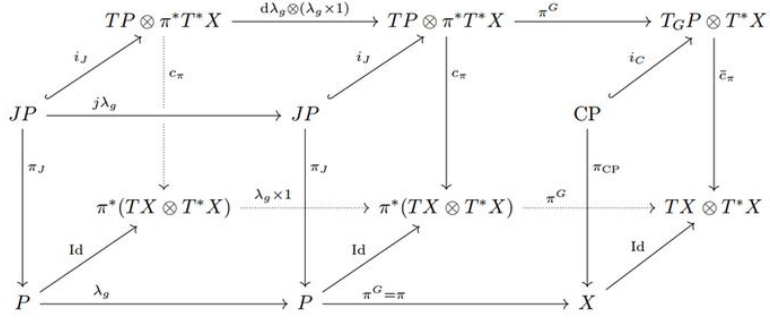


Content: Nested Integral Pathway — Across stacked manifolds

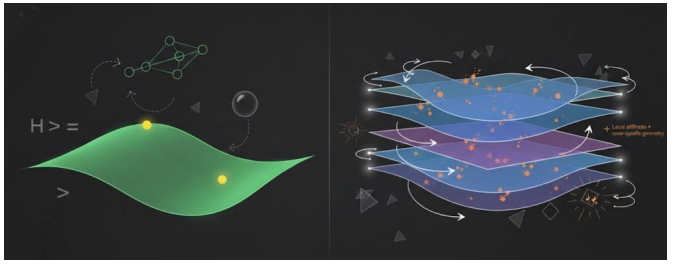
范畴论与深度流形

范畴论与许多优雅的数学理论一样，提供了精美的描述性框架，但并不能直接刻画神经网络中可计算的结构。深度流形通过聚焦于真实支配动态的计算过程与几何结构学习，弥补了这一空缺。

- 范畴论组织的是计算如何进行组合；而深度流形解释的是神经网络如何进行计算。
- 范畴论擅长刻画组合性的语法结构，但并不擅长建模高阶非线性、流形可塑性或训练动力学。
- 神经网络并非范畴意义下的组合结构，而是迭代积分体系，其收敛性与行为由不断演化的边界条件以及随机不动点所决定。



神经网络优化



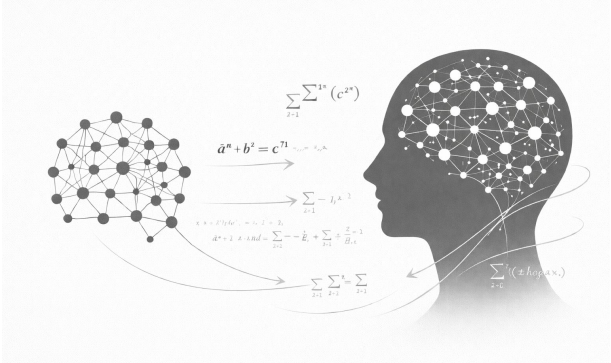
- 经典优化的核心论点：优化动力学 → 隐式正则化 → 泛化
- 经典优化要求存在一个明确定义的目标函数；而神经网络训练并非在优化某个预先给定的支配函数。
- 神经网络执行由数据和架构所引入的约束，而损失函数仅作为边界泛函
- 神经网络的“优化”最好被理解为一种数值输运机制，用于减少由数据约束所引发的弱残差。损失函数是定义在经验约束点上的边界泛函，而非所建模世界的控制方程。
- 在大型语言模型（LLM）中，目标函数是非平稳的，且在语义上与真理并不对齐；因此，当训练收敛时，它并非收敛至经典的全局最优解，而是收敛到由堆叠的分段流形几何所刻画的随机不动点。

深流形的数学传承

神经网络数学早已存在

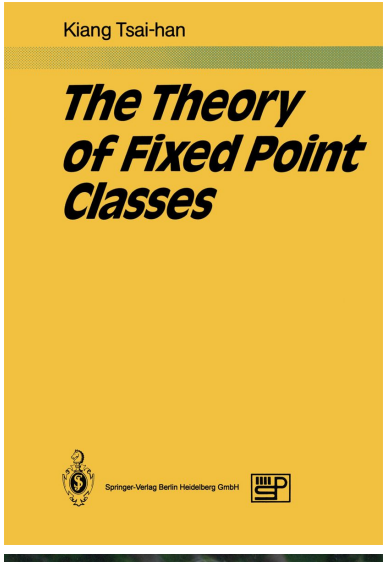
深度流形只是将其揭示出来。

- 神经网络推动着数学的发展：在实践中以具体方式推 进，在理论中以潜移默化的方式深化，而 这一进程主要由非数学家、AI 先驱以及更广泛的 AI 社区所驱动。
- 历史总是在重演。数学史上一些最重要的突破，往往并非源自数学学科内部，而是由物理学家、工程 师或其他科学家率先推 动的。
    1. 皮埃尔·德·费马 —— 数论 (1637)，背景：律师
    2. 布莱兹·帕斯卡 —— 概率论 (1654)，背景：物理学家、发明家与哲学家
    3. 艾萨克·牛顿 —— 微积分 (1666)，背景：物理学家与自然哲学家
    4. 亚诺什·鲍耶伊 —— 非欧几里得几何 (1832)，背景：军官
    5. 克劳德·香农 —— 信息论(1948)，背景：电气工程师
- 牛顿发展微积分是为了表达运动定律，而非为了解决抽象的数学 问题。今天的 AI 革命正延续着这一传统。



Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

深度流形(Deep Manifold)的数学传承



1960s, 不动点类理论

石根华 (Gen-Hua Shi)
60s: Theory of fixed point classes
70s: KeyBlock Theory
80s: Discontinuous Deformation Analysis
90s: Numerical Manifold Method
00s: Contact Theory (inequality theory)

不动点类理论把微积分解的: 存在性, 唯一性, 稳定性 都解决了, 那是牛顿提出微积分后二百多年

1980s, 数值流形

陈省身教授（Professor Shiing-Shen Chern），被广泛视为现代微分几何之父，曾担任石根华博士论文委员会(加州.伯克利) 的三位成员之一，为数值流形的数学框架提供了学术认可，并为该理论的后续发展奠定了基础。

陈教授唯一的问题是:“多层分片光滑覆盖(流形)能否扩展到任意复域？”陈教授一定会很欣慰地看到神经网络的进展，因为其几何本质是多层分片光滑覆盖（流形）

数学家的梦想与追求

- 局部和整体, 连续和非连续, 正反问题

整体数学 : 微分流形

- **1830**, **广义数**: 抽象代数中的群, 环, 域及数律. 埃瓦里斯特·伽罗瓦 (Evariste Galois)
- **1892**, **广义形状**:代数拓扑中的覆盖空间与几何规律. 昂利·庞加莱(Henri Poincare)
- **1952**, **广义函数**: 基于覆盖系统的流形与物理定律, 陈省身(Shiing-shen Chern)
    - 包括微分方程去计算复杂的物理问题,为后来的数值计算提供了数学基础

广义上讲

- 陈省身: **可以例** 任意复杂微分方程
- 丘成桐: **如何例** 任意复杂微分方程
- 石根华: **怎么解** 任意复杂微分方程

- 计算方法虽然数不胜数, 但它们都植根于少数几条基本的物理定律。
- 这些方法背后的数学方程在很大程度上是相通的。
- 这有助于人类理解世界, 对人工智能(AI)而言亦是如此。
- AI是可学习数值计算。

数学家不是这样训练的 (石根华, 2024.05)

- 将正时间的正问题与负时间的反问题同时求解，一直是数学家的梦想，但他们始终不知道从何入手。而神经网络却能自然地处理这个问题
- 变量、系数，甚至坐标都在变化，一切都处于变动之中。数学家不是这样的训练。这样的设计，不可能出自数学家之手
- 数学家在处理超过两层的复合函数时格外谨慎，警惕其中诸多陷阱，而神经网络却几乎漫不经心地轻松应对
- 神经网络在数学上拥有堆叠的覆盖层，而数值流形通常只有三到四层覆盖。相比之下，神经网络则堆叠了上百乃至上千层。我从未想过有人会将其推进到这种程度