Deep Manifold

Neural Network Mathematics

Max Ma and Gen-Hua Shi

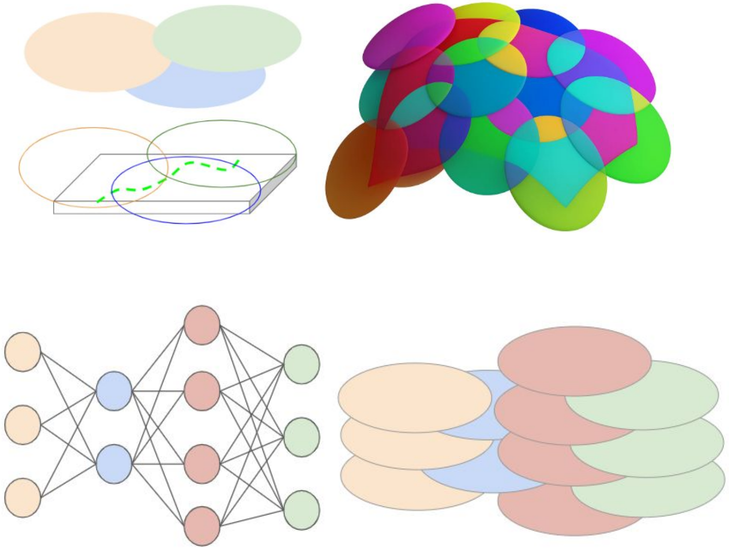
2026.01

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

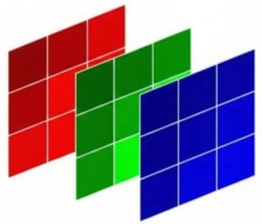
Deep Manifold Part 2: Neural Network Mathematics, arXiv:2512.06563

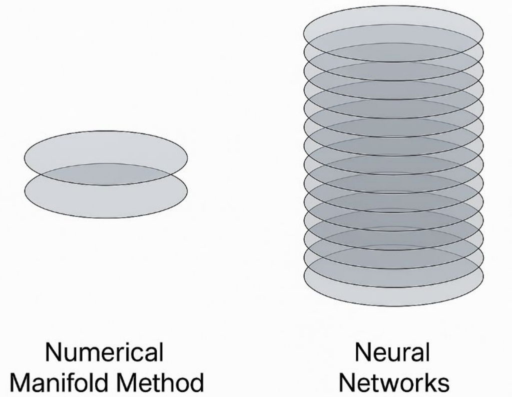
Deep Manifold Part 1:Anatomy of Neural Network Manifold

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



Data Flow (Physical Domain)



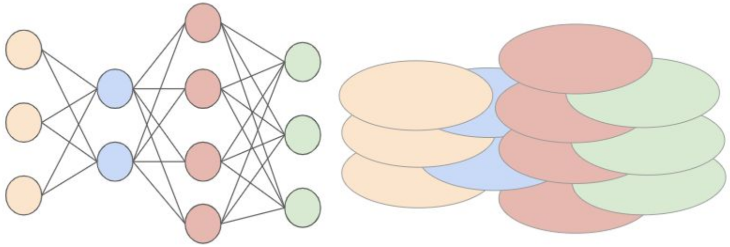


Neural Network Mathematics

Neural Network Geometry

Geometry bestows the eye that beholds all things from above, a very ladder to the freedom

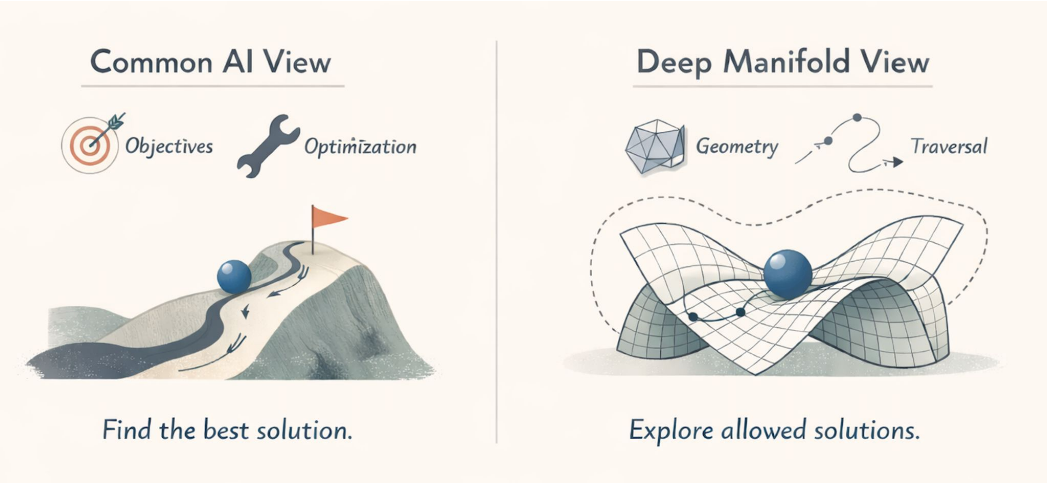
- Connected and stacked piecewise-smooth manifolds jointly form the geometric structure of the representation space.
- Node covers act as the local units of these piecewise-smooth manifolds, and their orientations change at every iteration.
- These piecewise-smooth manifolds are differentiable and integrable.



Geometry Rules

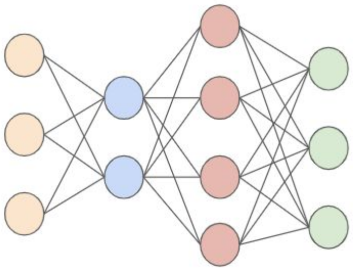
- Common AI View
  - Objectives(loss) + Optimization(parameter) determine the solution
  - Geometry emerges as a byproduct, secondary
- Deep Manifold View
  - Geometry determines what solutions can exist
  - Optimization only traverses a pre-shaped manifold

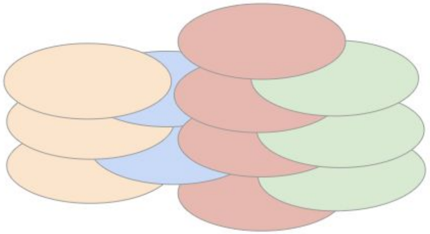
- Learning is inverse and non-identifiable
- Geometry is the only stable prior
- Geometry determines inference intrinsic pathways

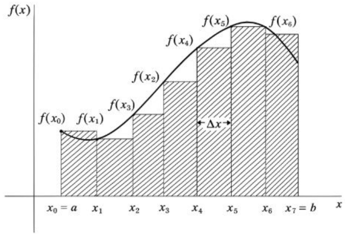


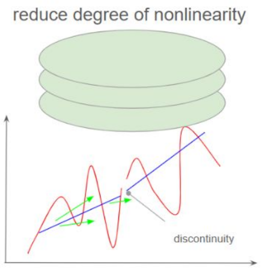
Stacked Piecewise Manifold

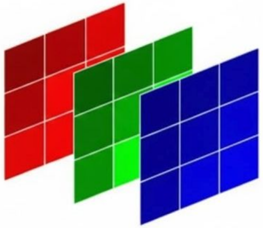
- A manifold: a point, a line, a cycle, a triangle, an infinite-dimensional Banach manifold
- Image RGB: 3 stacked Pointwise Manifold
- Neural Network: connected, stretched, **stacked piecewise** Manifold
- Stacked Piecewise Manifold Benefit: High Order Nonlinear Data

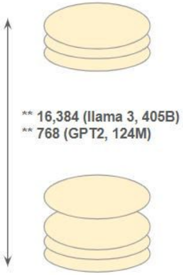






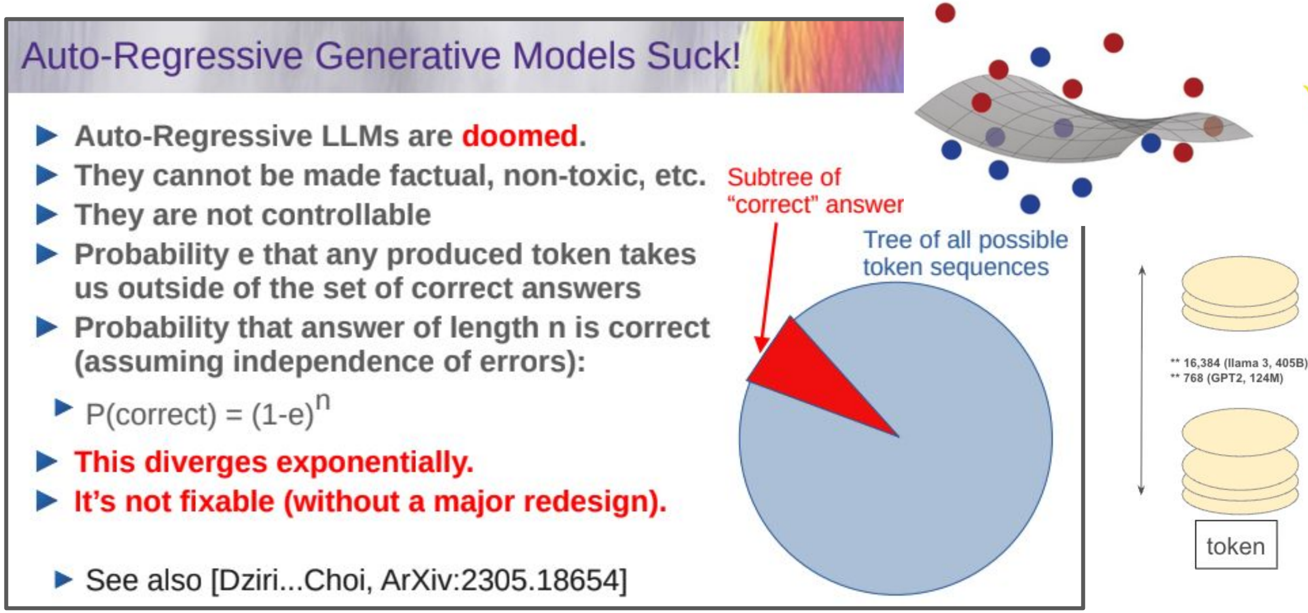






Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

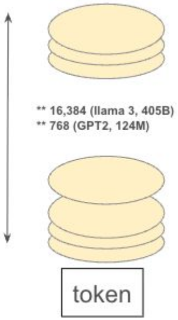
Y. LeCun is right for a single manifold, but why do Transformers work so well ?



Auto-Regressive Generative Models Suck!

- Auto-Regressive LLMs are **doomed.**
- They cannot be made factual, non-toxic, etc.
- They are not controllable
- Probability e that any produced token takes us outside of the set of correct answers
- Probability that answer of length n is correct (assuming independence of errors):
  - P(correct) = $(1-e)^n$
- **This diverges exponentially.**
- **It’s not fixable (without a major redesign).**
  - See also [Dziri...Choi, ArXiv:2305.18654]

Subtree of “correct” answer
Tree of all possible token sequences



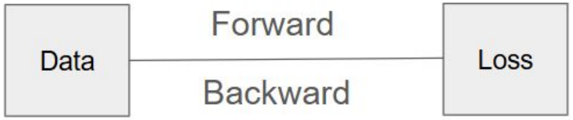
- The exponential-decay critique treats generation as a ***single manifold trajectory*** with independent failure at each step.
- Transformers operate on **stacked piecewise manifolds**, where deviations project onto shared geometric subspaces.
- Error events overlap rather than compound. Generation stability follows **union-bounded geometry**, not multiplicative probability collapse.

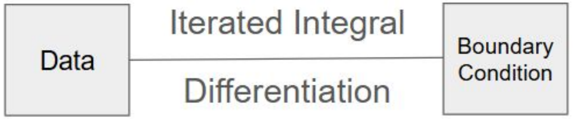
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

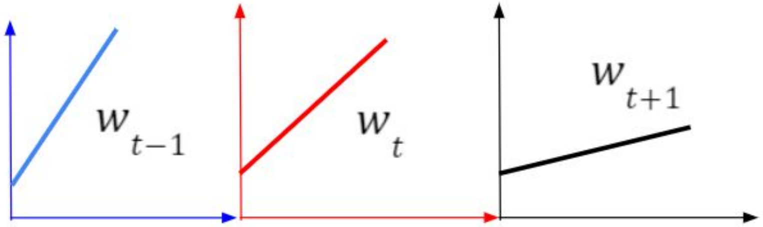
Neural Network Algebra

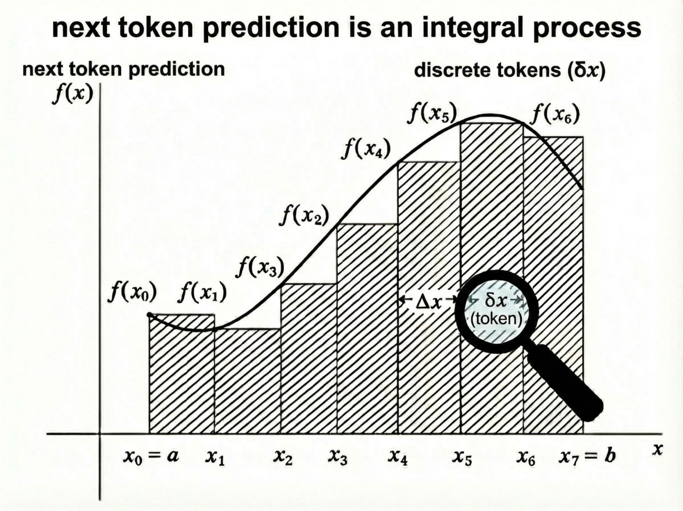
Algebra is the science of operations, the silent one behind all transformation

- The coordinate system evolves with each iteration;
- Counting serves as the most primitive algebraic unit,
- Iterated-integral structure of forward propagation;
- Activation is propertyless


Data
Forward
Backward
Loss


Data
Iterated Integral
Differentiation
Boundary Condition




next token prediction is an integral process

Neural Network Equation

An equation is the quiet connector that enables computation

- Fixed-Point Residual as the Primitive Equation

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

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

- Lagrangian Formulation of Neural Fixed Points

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

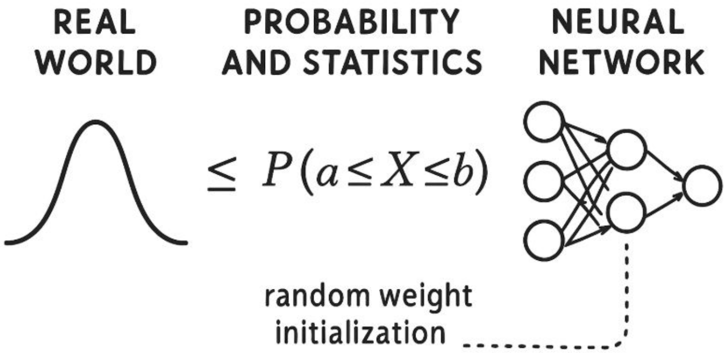
$$\nabla_\theta L(\theta, \lambda) = 0 \implies \text{critical point of } \mathbb{E}_x \| f_\theta(x) - x \|^2$$

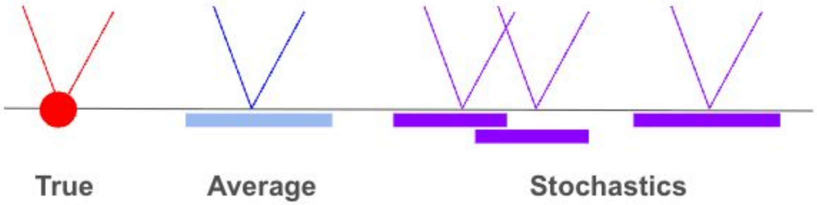
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Neural Network Stochastic

A stochastic world is an inequality world, but it is real

- Stochasticity is expressed through inequalities and group statistics based on summation
- The statistical structure enables neural networks to learn the inherently stochastic real world and naturally form stochastic fixed points.
- It is effortless

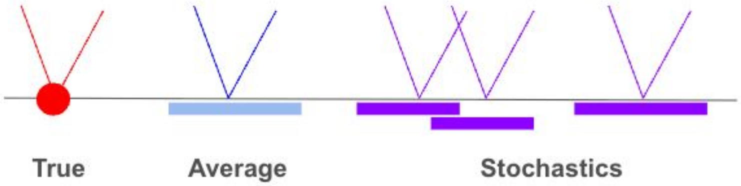

REAL WORLD
PROBABILITY AND STATISTICS
NEURAL NETWORK
$\leq\ P(a \leq X \leq b)$
random weight initialization

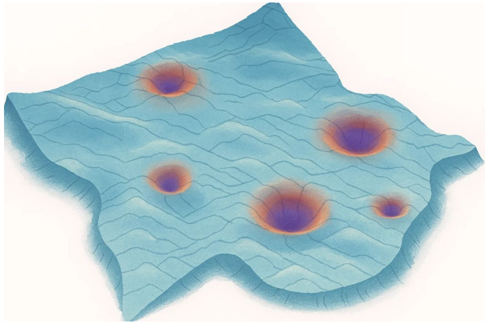


Neural Network Fixed Point

Fixed point theory is the theory of iteration, until fixed

- Iterations traverse billions of piecewise-smooth manifolds, giving rise to innumerable fixed points and convergence paths within foundation models.
- Because the training data themselves contain high-order nonlinearity, curvature (second derivatives), and moderate perturbations, the model can distinguish the correct convergence direction toward a fixed point.



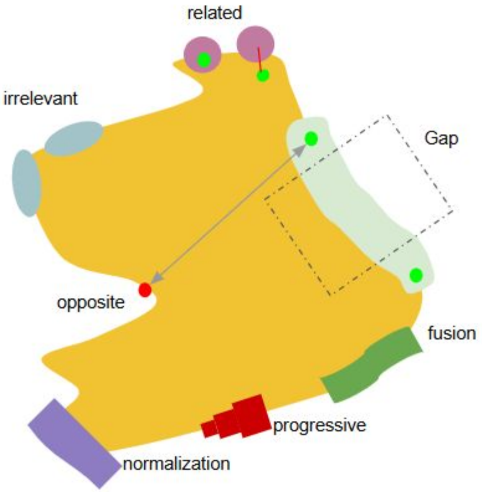


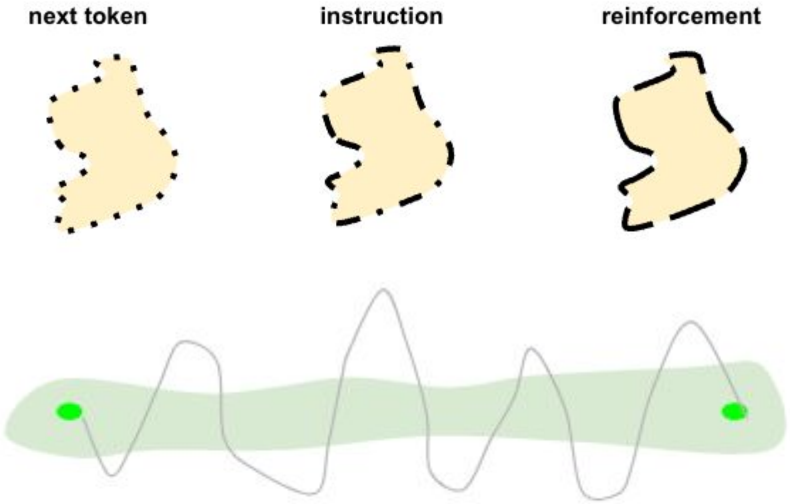
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Neural Network Boundary Condition

Boundary conditions give iteration purpose and direction

- Boundary conditions are the sole source of iterative direction and determine the convergence path during training.
- When a foundation model lacks static fixed points, symmetric, weak, and discrete boundary conditions become necessary to guide the convergence of a high-order nonlinear system.





Learnable Numerical Computation

Mathematics & Numerical Computation

- Mathematics
  - Universal in principle: seeks analytical solutions that hold across domains
  - Mathematics excels at description; often not directly solvable or computable.
  - Structural limitation: analytical existence does not imply closed-form expressibility, especially for high-order nonlinear, discontinuous, or stochastic systems.
- Numerical Computation
  - Galerkin method: a numerical solver that replaces exact solvability with weak consistency on a chosen representation space.
  - Approximation theorem: a result guaranteeing that functions in a target class can be approximated arbitrarily well by functions from a specified family, under a given metric. discretization, approximation, iteration
  - Adaptable: empirical terms, adaptive meshes/layers, usable convergence over exactness.
  - Pragmatic: whatever works, as long as it converges.

Numerical computation can be deceptive. Without solid mathematical grounding, it may advance remarkably far in practice—as seen in the supercomputer era, and again today in AI. Scaling computational power does not imply scaling mathematical understanding.

From Fixed Point to Learnable Computation

- Fixed point defines what learning is.

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

  - Beautiful descriptor, but not a solver; no numerical procedure
  - No notion of progress, No way to handle constraints (architecture, data)
- Lagrangian Formulation makes it solvable.

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

  - Lagrangian equilibrium = neural network fixed point
  - λ = boundary enforcer, g(θ) =0 architectural / data constraints
- Numerical iteration makes it real.
  - Mathematics becomes computation only when residuals are iteratively reduced under constraints
  - Galerkin Method: equations become solvable numerically only after residualization and iteration
  - Learning is not defined by objectives, but by the existence of a stable numerical iteration.

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

Different Model, Different Reasoning Path

- Neural Fixed Point Equation:

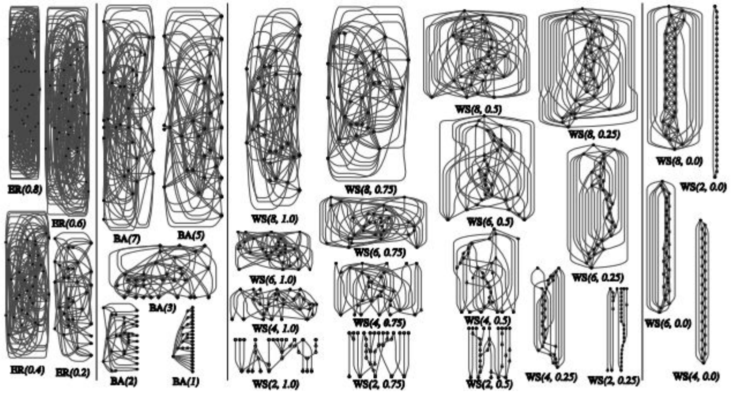
$$f(x) - x = e(x), \min_{\theta} e(x)$$

- Lagrangian Formulation of Neural Fixed Points:

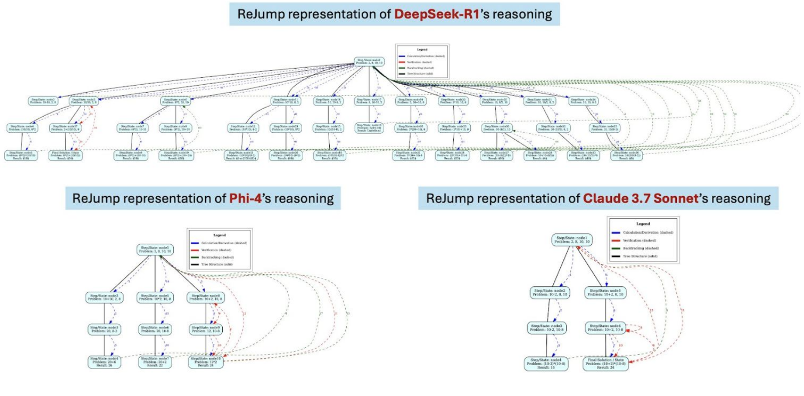
$$L(\theta, \lambda) = \mathbb{E}_x \left\| f_\theta(x) - x \right\|^2 + \lambda\, g(\theta)$$

- Let the model **Architectural** and/or **Data** constraints be g(θ) = 0

g(θ): Same prompt, different reasoning path, yet the same accurate output.



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


ReJump representation of DeepSeek-R1's reasoning
ReJump representation of Phi-4's reasoning
ReJump representation of Claude 3.7 Sonnet's reasoning
Legend

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

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

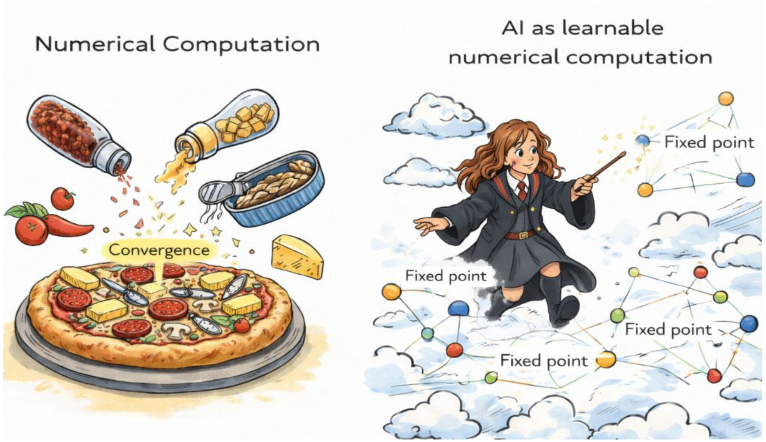
Neural Networks as Galerkin-Type Numerical Systems

- Neural networks can be understood numerically through a Galerkin-type formulation, rather than classical optimization.
- Classical Galerkin Method
  - Solve operator equation
  - Enforce solution weakly $\mathcal{N}(u) = 0 \quad \mathrm{on}\Omega$ $\int_{\Omega} \mathcal{N}(u_h) v_h \, d\mu = 0$
    - finite-dimensional trial space, residual orthogonality, numerical iteration
- Neural networks are not optimizing functions, they are numerically solving residual equations on learned manifolds.

Galerkin	Neural Network (Deep Manifold)	Galerkin without fixed geometry, which **itself is learned**.
Operator	Implicit data-induced operator	
Trial space	Learned manifold $\mathcal{M}_\theta$	What “**Solving**” Means **Not:** finding a global minimum **But:** achieving stable fixed-point consistency across admissible manifold regions
Basis functions	Propertyless activations	
Weak form	Residual energy integral	
Quadrature	Minibatch sampling	
Assembly	Stacked, piecewise manifolds	Training solves a variational fixed-point system numerically, inference traverses the learned geometry.
Solver	Backpropagation	
Convergence	Fixed-point stability	

AI as Learnable Numerical Computation

- Neural networks inherit a numerical worldview: solvers, not theorem provers.
- Core tension: scaling from domain-specific numerics toward mathematical universality.
- Numerical computation pragmatism: as long as it converges, the method is not restricted. As long as it converges, you can add any pizza toppings: pepperoni, pineapple, anchovies, extra cheese, whatever works
- Absent a fixed point, the network breaks all constraints. The Olympians descend from order into improvisation: each god crossing the same boundary by a different force, unconstrained, uncoordinated, yet collectively sufficient.



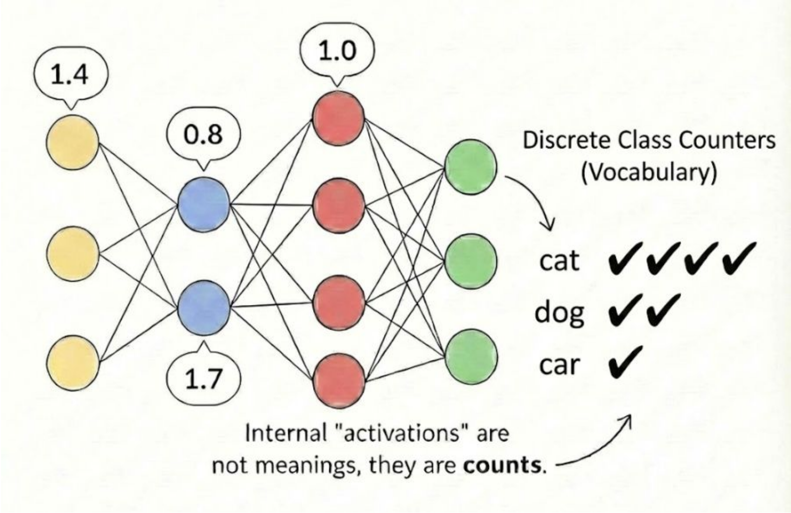
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592;

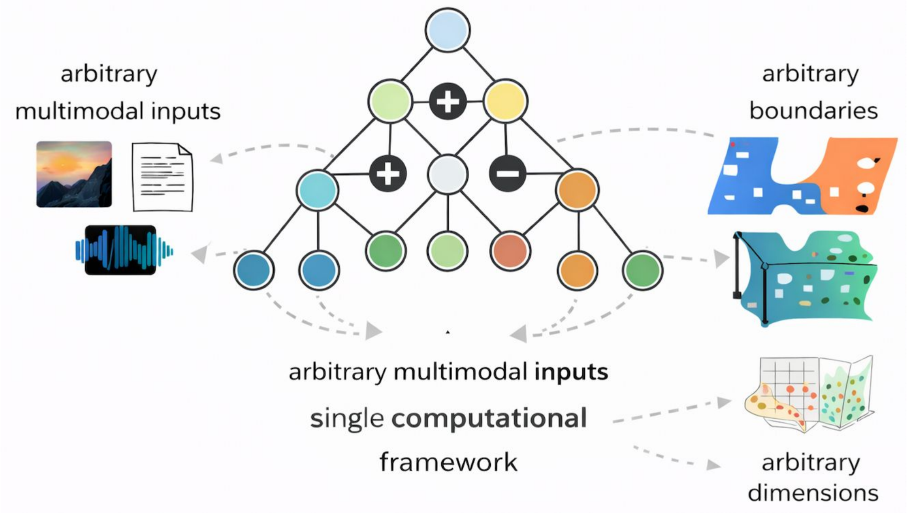
Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Neural Network Propertyless

Property-Lessness and Counting

- Activation as **hidden representations** or **latent variables**, but they lack assertable definitions and intrinsic properties.
- **Classification** asks *which class has more support than the others*.
- **Property-lessness:** Individual activations do not encode semantic properties.
- **Discrete decision compatibility:** Counting transitions cleanly from continuous accumulation to discrete class selection (argmax).

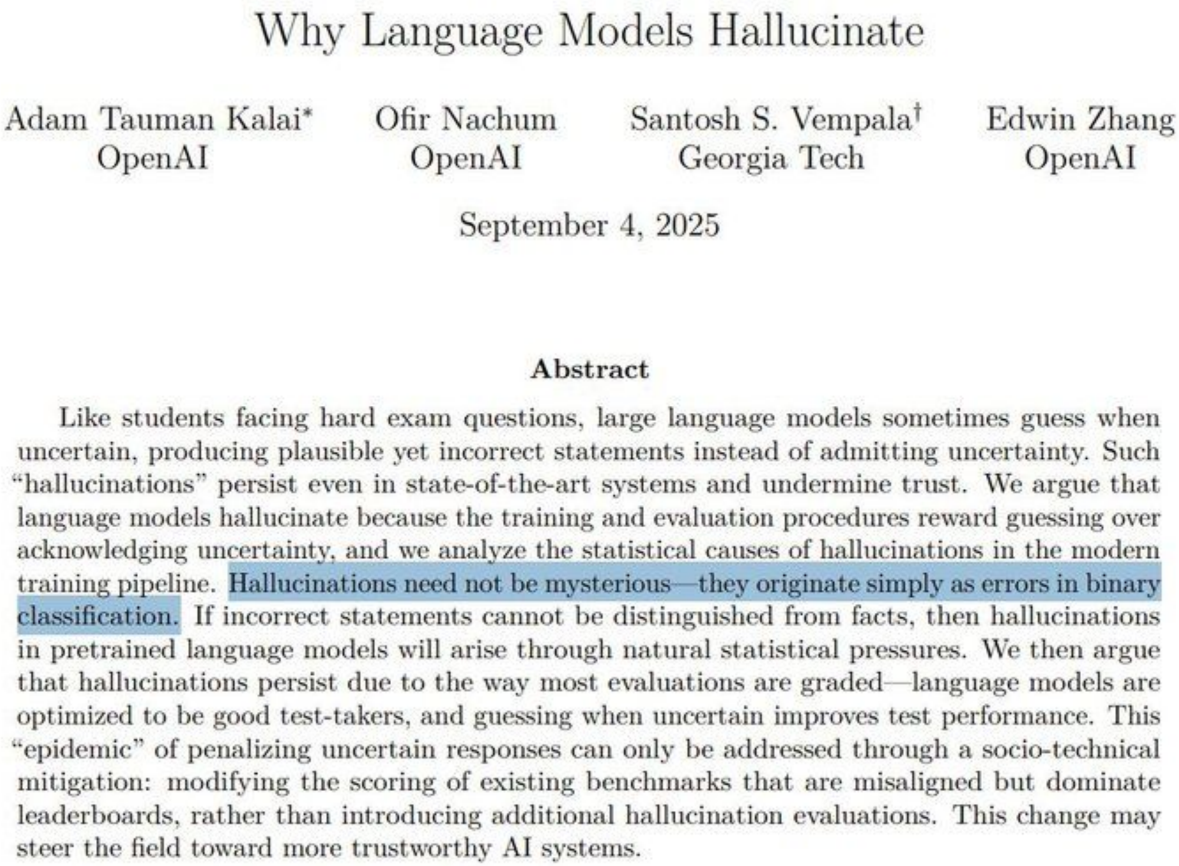

1.4
1.0
0.8
1.7
Discrete Class Counters (Vocabulary)
cat ✓✓✓✓
dog ✓✓
car ✓
Internal "activations" are not meanings, they are **counts**.



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

Open AI is Not Wrong

Large language models fundamentally perform classification over learned representation spaces.



Why Language Models Hallucinate

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

September 4, 2025

Abstract

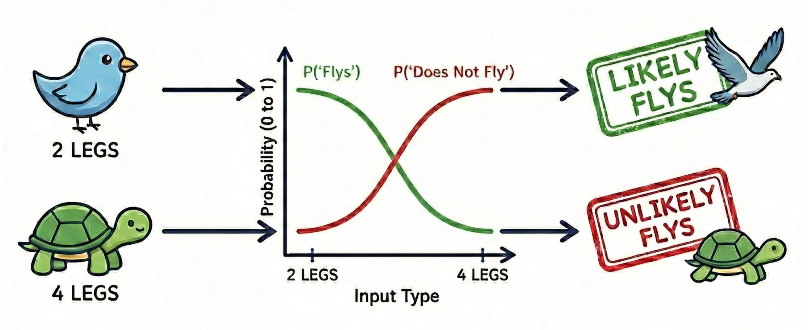
Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate 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.

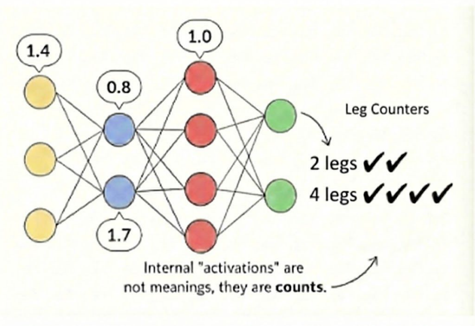
Classification

The classification is the system used to place items on the shelves; the **catalogue** is the records of each item, and the way you **locate** items that have been **classified**.

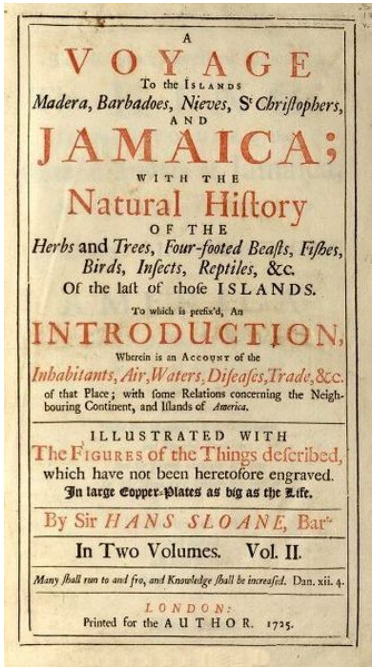
The Enlightenment Gallery at the British Museum

- Sir Hans Sloane's **Catalogue** of Jamaican Plants published in 1696
- **Catalogue** and **classification** enabled **reasoning** in the **Enlightenment** by reducing knowledge to **countable**, comparable features.
- A **two**-legged animal likely flys, A **four**-legged animal animal unlikely flys.
- Neural networks follow a similar principle of **classification** via aggregated activation **counts**, nothing else.





Internal "activations" are not meanings, they are **counts**.



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.

Property-Lessness and Counting

- If we examine the Transformer models, their operation is fundamentally classificatory
- The vocabulary defines the discrete class space, and every activation represents a local stage of counting
- Each layer acts as an integral operator that aggregates and transports evidence:

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

- Stacking these operators yields a full L-layer manifold integration:

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

- The categorical decision for the next token is then obtained by integrating the final manifold representation through a class-specific counting field:

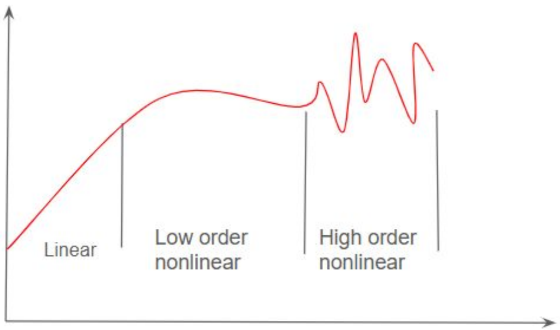
$$z_c = b_c + \int_{M^{(L)}} \theta_c(\xi)\, h_L(\xi)\, d\mu_L(\xi), \qquad p(c \mid p) = \frac{e^{z_c}}{\sum_{k \in V} e^{z_k}}$$

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

Data

High Order Nonlinear Data

US Flag: sharp color jumps between stripes (**White/Red**) , and between the **white** stars and the **blue** background. Such abrupt changes can be considered high-order nonlinearities



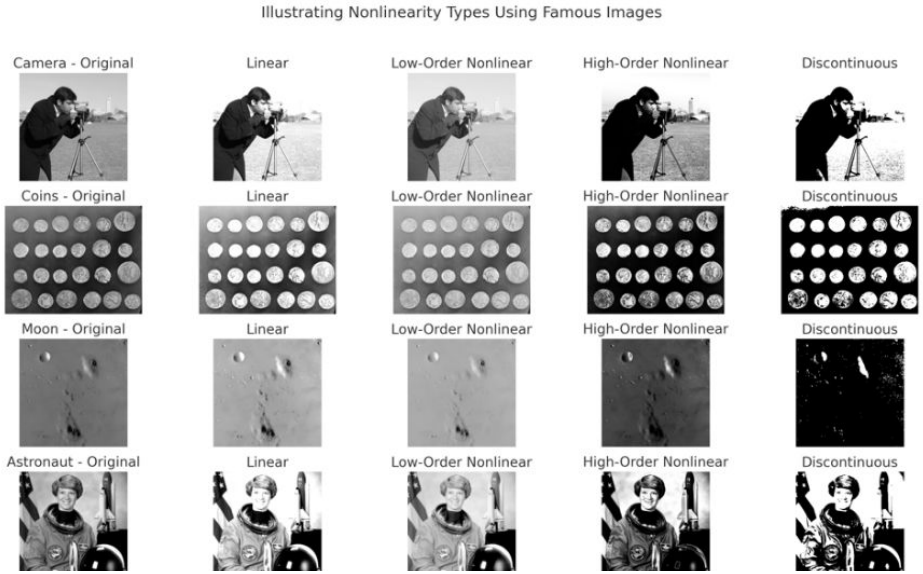
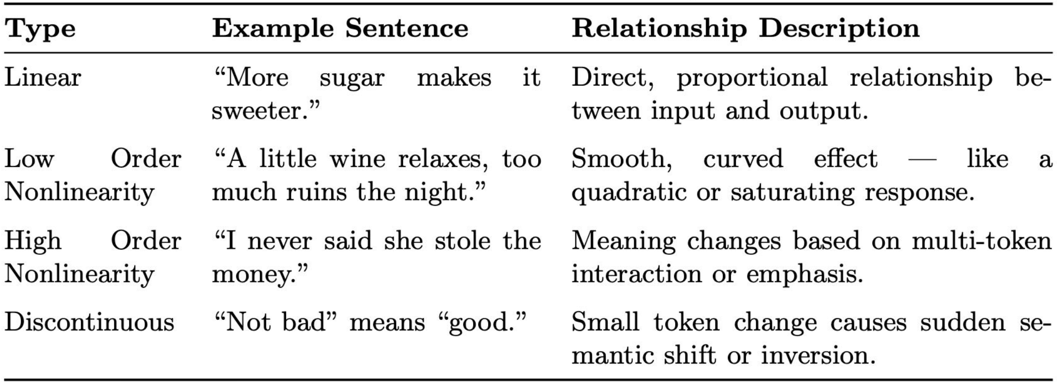


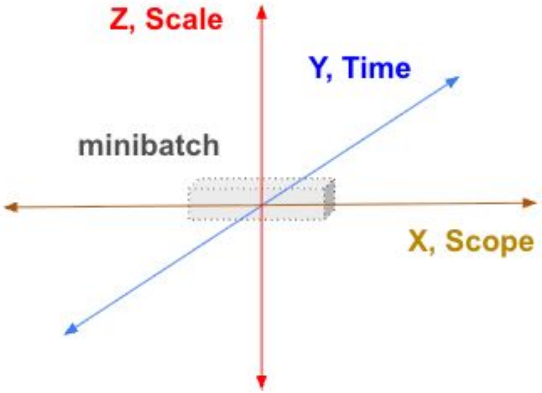
Table 3: NLP Nonlinearity

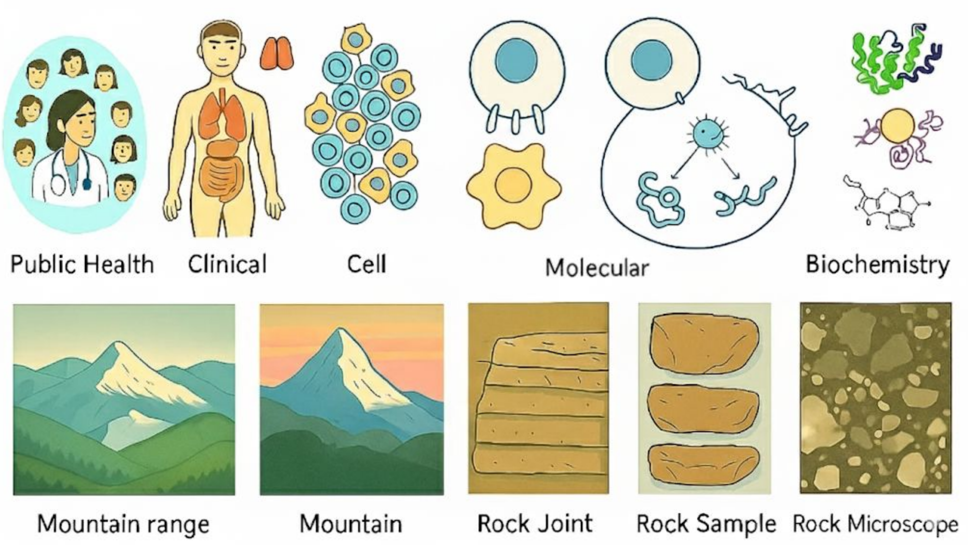


Type	Example Sentence	Relationship Description
Linear	“More sugar makes it sweeter.”	Direct, proportional relationship between input and output.
Low Order Nonlinearity	“A little wine relaxes, too much ruins the night.”	Smooth, curved effect — like a quadratic or saturating response.
High Order Nonlinearity	“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.

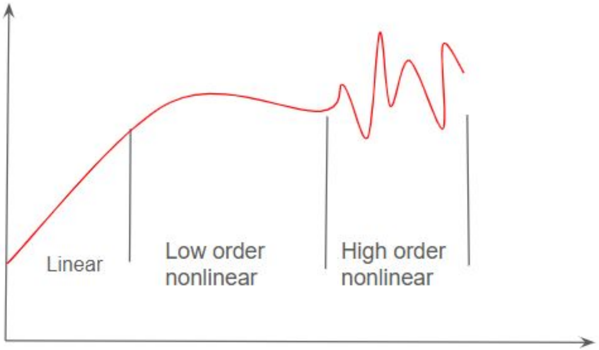
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

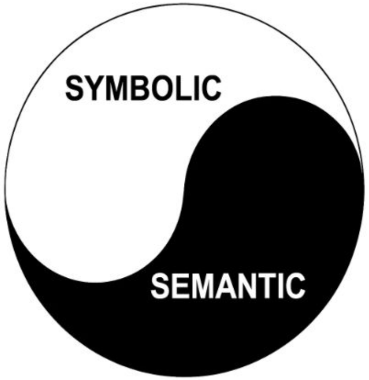
Data Complexity





Mini-batching is a major challenge for fixed-point convergence in neural network computation.





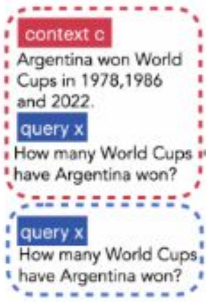
High-order nonlinearity in data is a major source of neural plasticity.

Symbolic vs. semantic is a scale issue.

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

Temporal Misalignment as an Ill-Posed Inverse Problem

- Neural Network is **forward-inverse combined**, which enable neural network to learning from the past.
- The **negative time** is critical property of inverse problem.
- **Temporal misalignment** in LLMs arises because time is a missing boundary condition without explicit timestamps, inference producing **aliasing** between eras.
- Neural Network **property-lessness amplifies** temporal ambiguity rather than resolving it.
- Migration Path
  - Explicit Timestamp in Training Data. This restores orientation to the data manifold
  - Prompt-Level Temporal Boundary Conditions



context c

Argentina won World Cups in 1978,1986 and 2022.

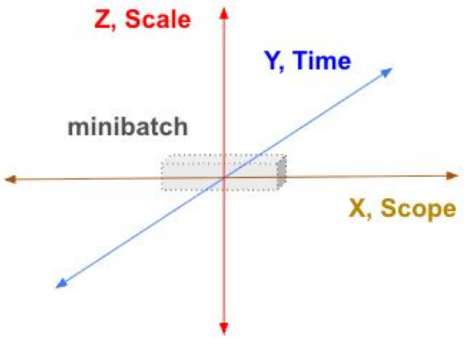
query x

How many World Cups have Argentina won?

query x

How many World Cups have Argentina won?

'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.



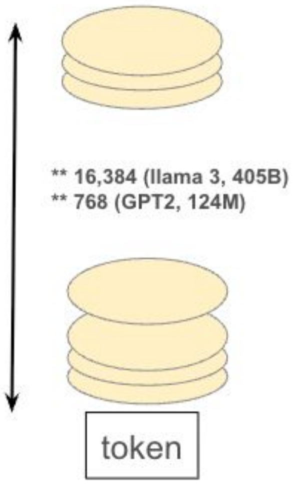
Trusting Your Evidence: Hallucinate Less with Context-aware Decoding, W. Shi et al. 2023

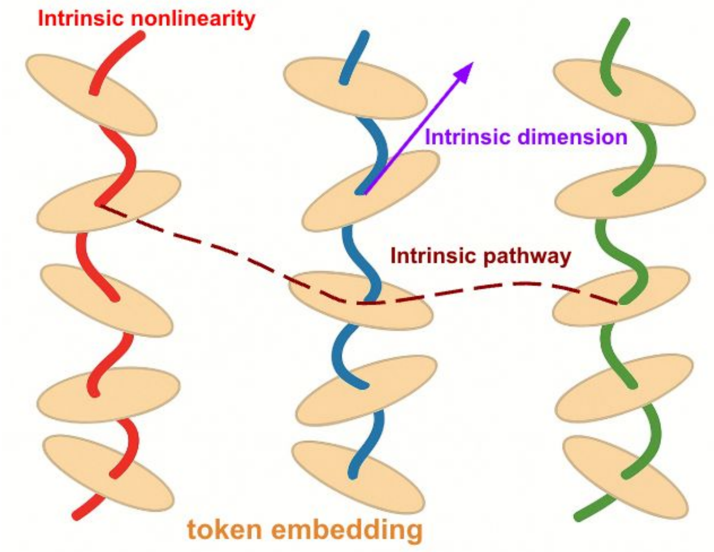
Temporal misalignment is a consequence of missing time as an explicit boundary condition in forward–inverse learning, not a generalization error. From a Deep Manifold perspective (arXiv:2409.17592), neural networks converge to stochastic fixed points shaped by data; they do not "fail" to generalize, but rather lack the temporal encoding necessary to distinguish eras. While their forward–inverse nature allows integration of past data, the absence of timestamps prevents time from acting as a formal constraint, leading to systematic misalignment.

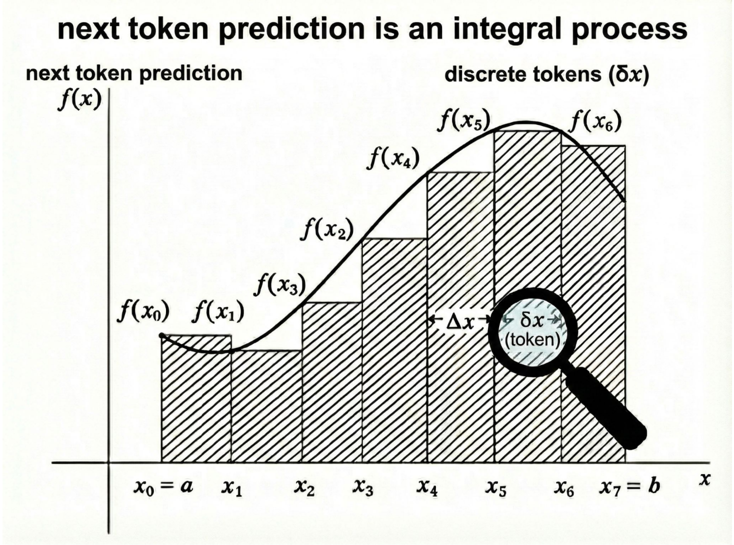
Inference

Next Token Predication is Integral Process

- Each element of a token embedding defines a local **piecewise** manifold.
- A token embedding is a **stack** of piecewise manifolds, thereby defining its **intrinsic nonlinearity**.
- The **intrinsic dimension** is identified with the tangent space induced by intrinsic nonlinearity between embedding elements, as defined in **dimensionality theory** (algebra).
- An **intrinsic pathway** is an integral curve of this field—millions, potentially billions, of such paths coexist.
- Neural network **property-lessness** makes all above possible.

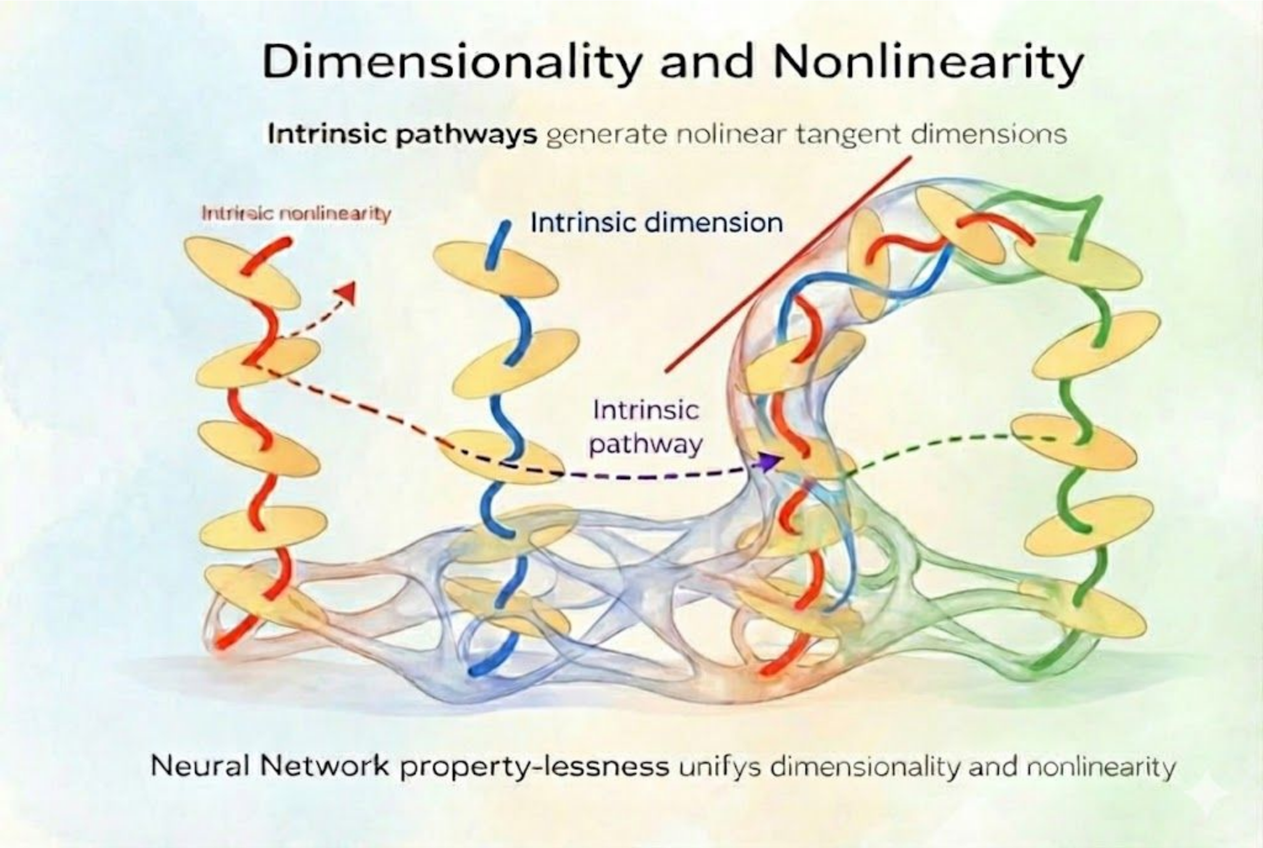






Taking a derivative without an integral, What is the purpose of the derivative?

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

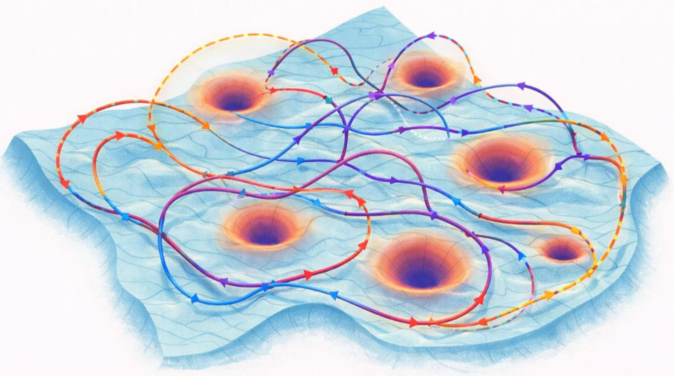

Dimensionality and Nonlinearity
Intrinsic pathways generate nolinear tangent dimensions
Intrirsic nonlinearity
Intrinsic dimension
Intrinsic pathway
Neural Network property-lessness unifys dimensionality and nonlinearity

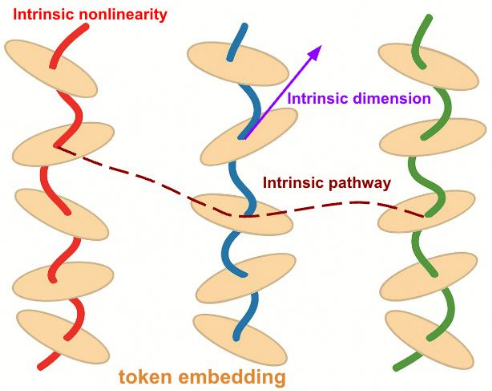
- **Nonlinearity,** defined by each element of a token embedding, piecewise manifold
- **Dimensionality,** the nonlinearity tangent (dimensionality theorem algebra)

What appears as dimension **compression** is in fact intrinsic dimension **extraction** along intrinsic nonlinearity

Inference Complexity: Dynamic fixed points and integral pathways.

- Inference is simultaneous traversal of a vast family of integral pathways
    - Prompt / instruction defines the initial boundary condition of the integral
- Learned geometry contains many coexisting fixed points
    - Each token prediction corresponds to a local convergence
- Fixed points during inference are dynamic
    - Context update ⇒ boundary shift ⇒ fixed-point drift.
- Inference can be approximated by a series of Fourier expansion of the integral field
    - Dramatic speedup and cost reduction without altering learned geometry.

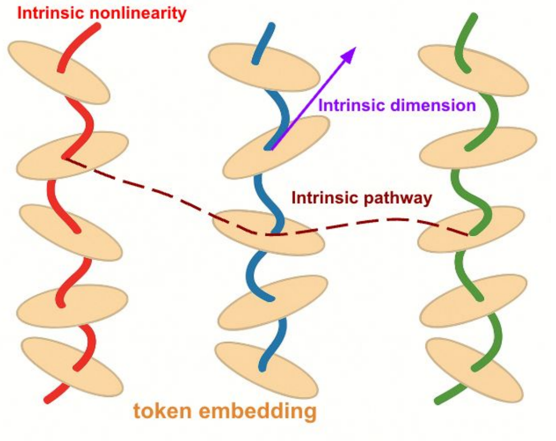


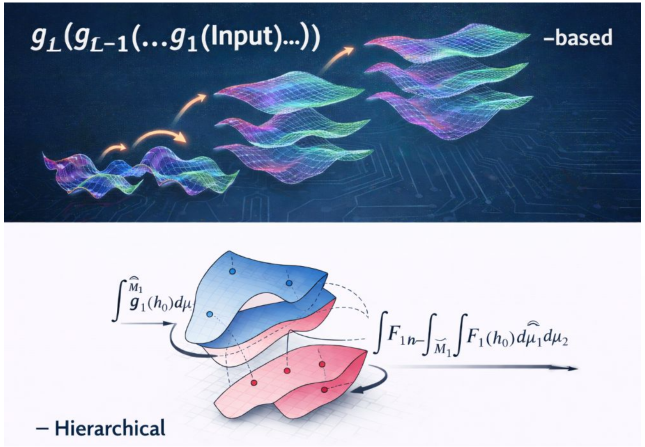


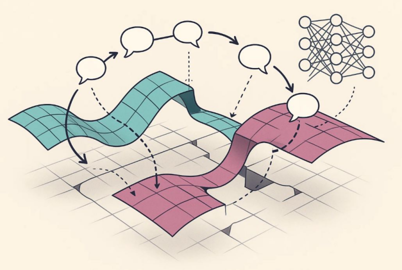
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

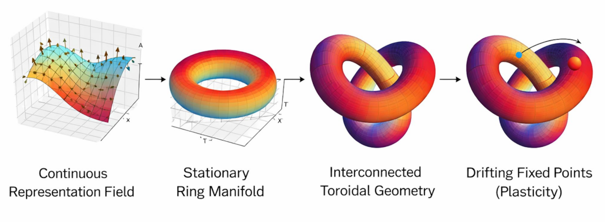
Multi-turn, Hierarchical and Recursive

- **Multi-Turn**: Neural networks admit an enormous number of emergent intrinsic pathways, which could be in trillisions.
- **Hierarchical**: Neural Network is based on function composition and operates on nested and iterated integral over stacked manifolds
- **Recursive**: Neural Network develops Interconnected Toroidal, ring-like Geometry

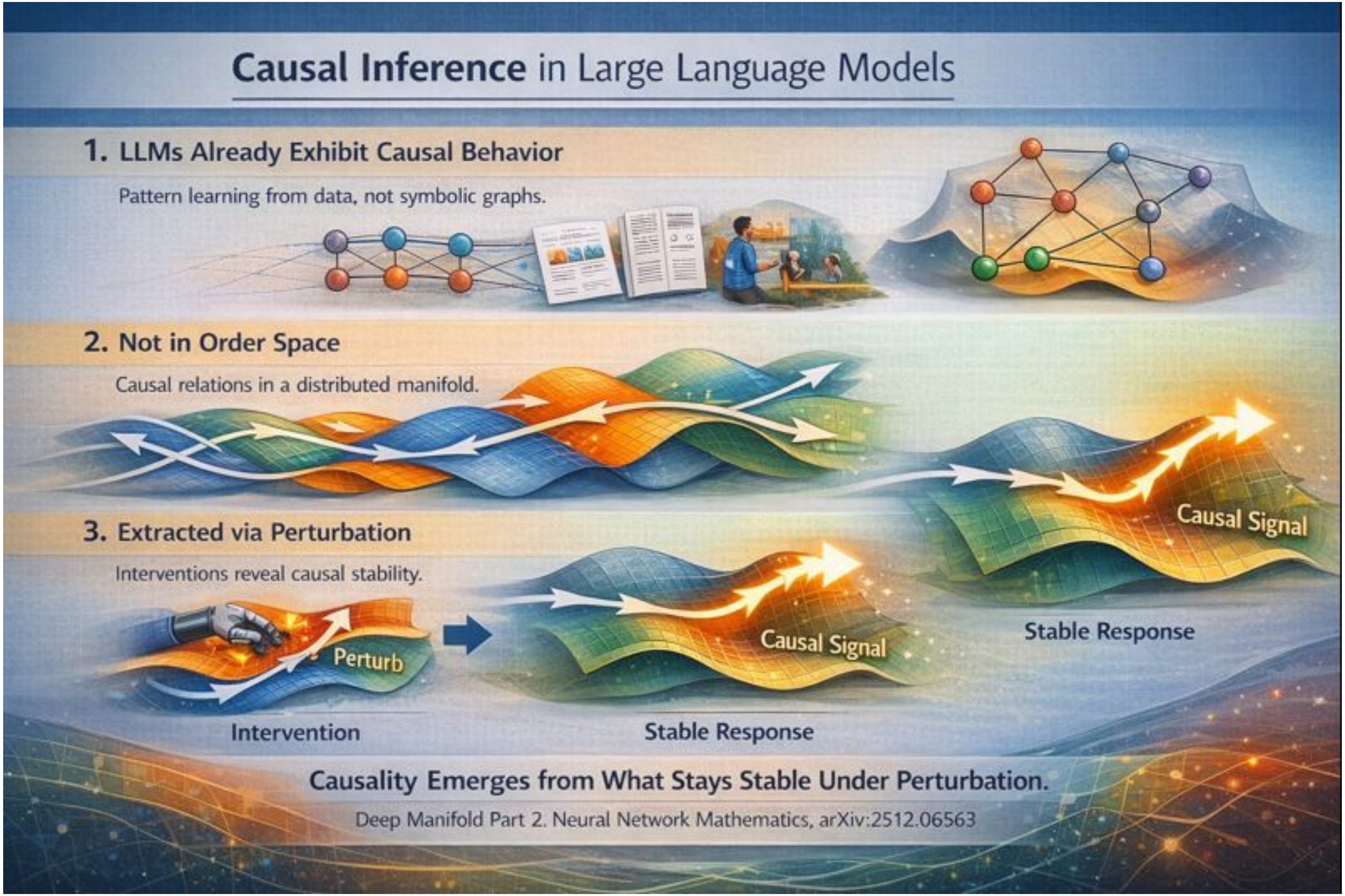



$g_L(g_{L-1}(\ldots g_1(\text{Input})\ldots))$
–based
$\int^{\widehat{M}_1} g_1(h_0)d\mu$
$\int F_{1n} - \int_{\breve{M}_1} \int F_1(h_0)\, d\widehat{\mu}_1 d\mu_2$
– Hierarchical




Continuous Representation Field
Stationary Ring Manifold
Interconnected Toroidal Geometry
Drifting Fixed Points (Plasticity)

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



Causal Inference in Large Language Models

1. LLMs Already Exhibit Causal Behavior

Pattern learning from data, not symbolic graphs.

2. Not in Order Space

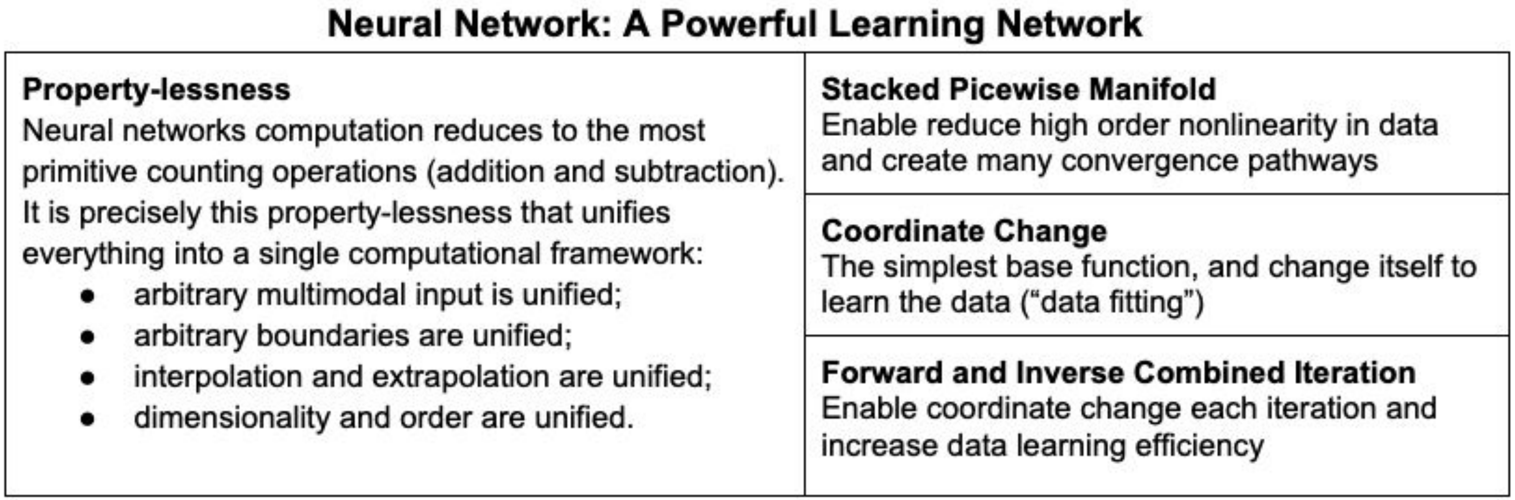
Causal relations in a distributed manifold.

3. Extracted via Perturbation

Interventions reveal causal stability.

Causality Emerges from What Stays Stable Under Perturbation.

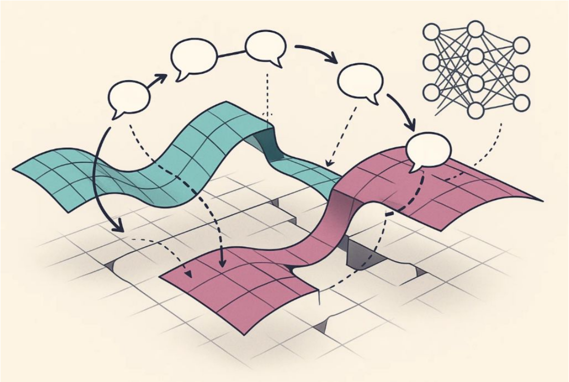
Deep Manifold Part 2. Neural Network Mathematics, arXiv:2512.06563

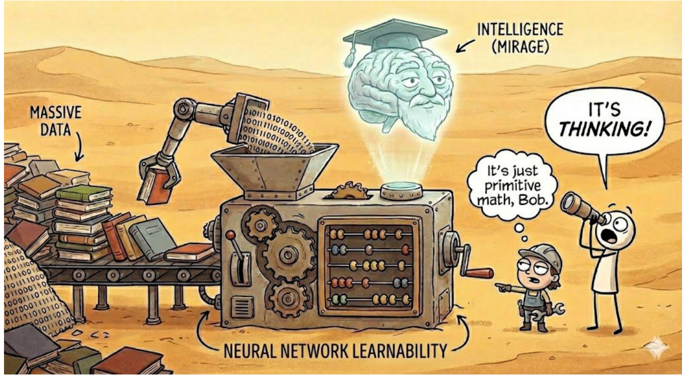


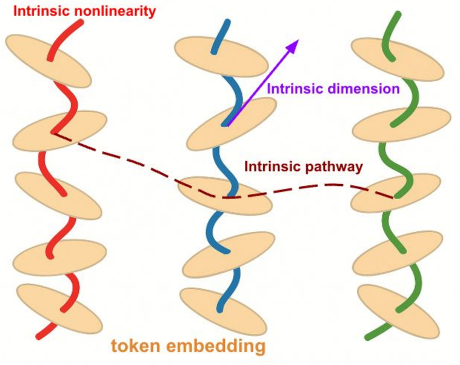
Neural Network: A Powerful Learning Network

Property-lessness Neural networks computation reduces to the most primitive counting operations (addition and subtraction). It is precisely this property-lessness that unifies everything into a single computational framework: • arbitrary multimodal input is unified; • arbitrary boundaries are unified; • interpolation and extrapolation are unified; • dimensionality and order are unified.	**Stacked Picewise Manifold** Enable reduce high order nonlinearity in data and create many convergence pathways
	Coordinate Change The simplest base function, and change itself to learn the data ("data fitting")
	Forward and Inverse Combined Iteration Enable coordinate change each iteration and increase data learning efficiency

- Neural network **Learnability**: Neural networks **do not** possess intrinsic **reasoning** or **cognitive** abilities; such capabilities are entirely learned from data through the model's **learnability**.
- **Scaling** laws remain largely empirical and lack a unifying **theoretical** foundation; in particular, they lack a principled notion of learning **efficiency** and largely ignore learning **complexity**.



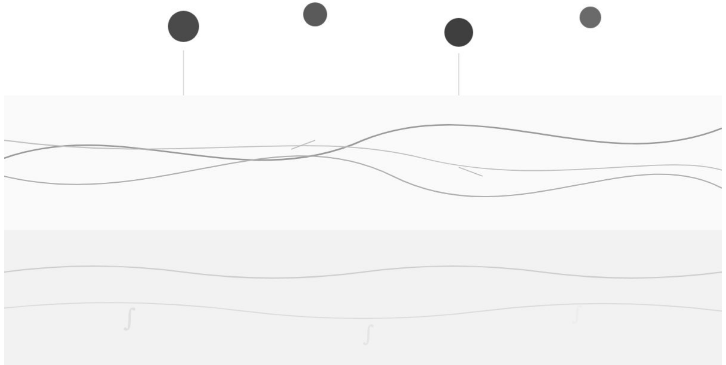




Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Emergence And Complexity

- Emergence and complexity are inseparable.
  - Complexity is the substrate (soil) of emergence; emergence is its observable outcome (fruit). **Increasing** complexity **expands** better emergent behavior.
- Deep Manifold attributes
  - the emergence and complexity to **property-lessness**
  - the fundamental reason neural networks are so powerful.
- From classification to emergence: a **property-less** foundation
  - Classification precedes explanation and is grounded in mathematical property-lessness.
  - The **Enlightenment** Gallery (British Museum) frames classification as the origin of science

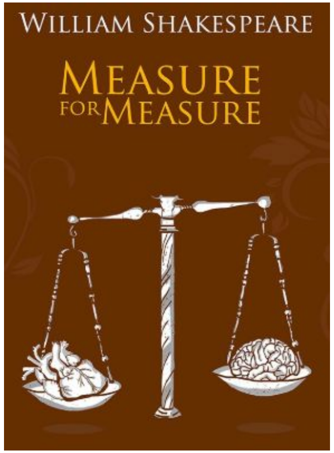


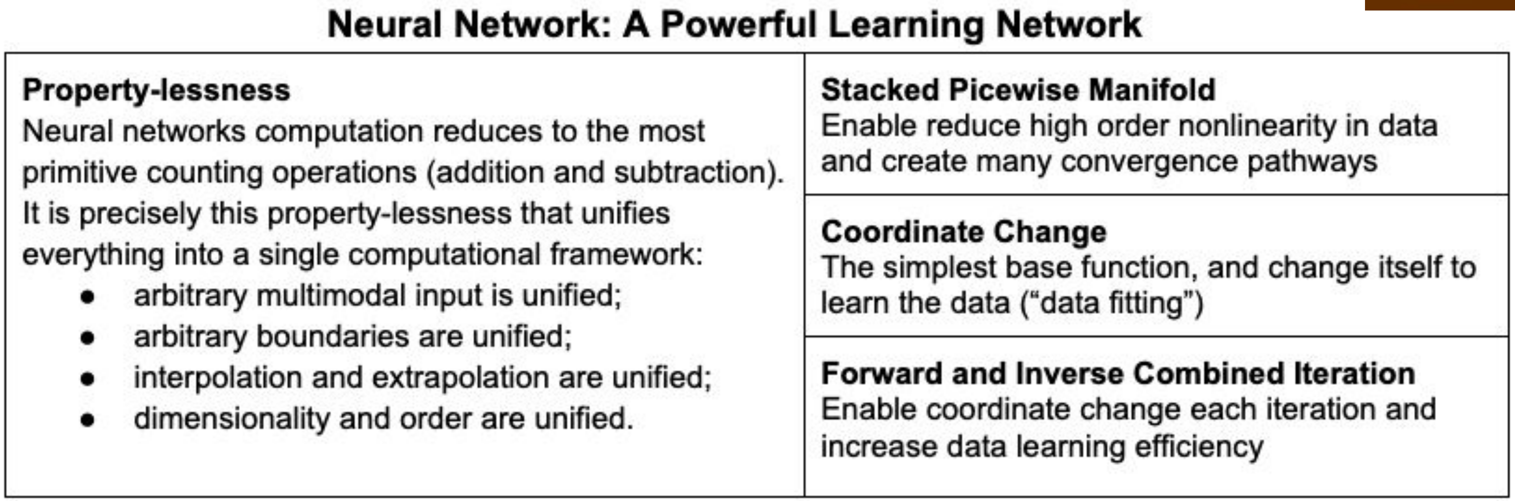
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

“Measure for Measure” 水能载舟，亦能覆舟

It is excellent To have a giant's strength; But it is tyrannous To use it like a giant.

A **neural network** is a powerful learnable numerical computation. One of its greatest strengths lies in its "**property-lessness**". This allows neural networks to learn almost anything from mixed modalities and across disciplines. However, It is not grounded in any specific physical, known scientific *principles*, physical and time *dimension*, *moral compass*, or even *common sense*. They are only **as faithful as their training data**.





Neural Network: A Powerful Learning Network

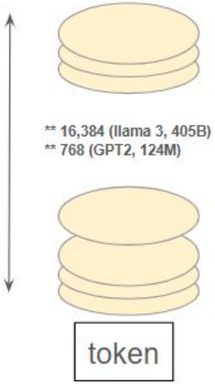
Property-lessness Neural networks computation reduces to the most primitive counting operations (addition and subtraction). It is precisely this property-lessness that unifies everything into a single computational framework: • arbitrary multimodal input is unified; • arbitrary boundaries are unified; • interpolation and extrapolation are unified; • dimensionality and order are unified.	**Stacked Picewise Manifold** Enable reduce high order nonlinearity in data and create many convergence pathways
	Coordinate Change The simplest base function, and change itself to learn the data (“data fitting”)
	Forward and Inverse Combined Iteration Enable coordinate change each iteration and increase data learning efficiency

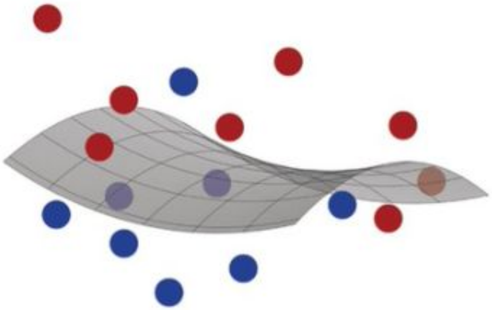
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592;   Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

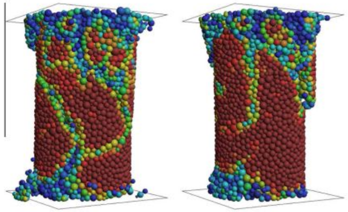
Training Progression

Transformer Positional Embedding

- It is Transformer's blunder/defect
  - Transformers naturally operate on stacked, **piecewise manifolds**. Due to the very high embedding dimensionality, token representations become effectively **point-wise**, like grains of sand, resulting in strong adaptability to data high-order nonlinearity while promoting global manifold **smoothness** (local continuity and global transitions) and computation stability
  - but leads to compute-budget overruns and ultimately a **scaling ceiling**, similar in spirit to problematic binding mechanisms we have observed in Discrete Element Methods (DEM).
- **Two** fundamental aspects of Transformer have never been explicitly recognized by the AI community.



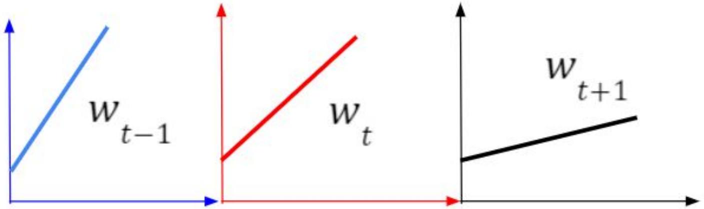


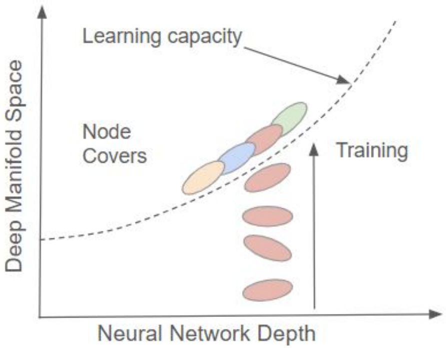


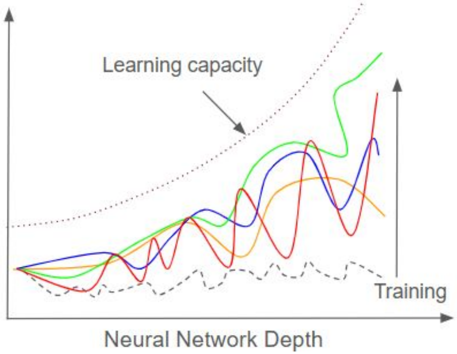
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Continual Learning

- **Why it is possible:** neural-network node coordinates can change indefinitely across iterations: a capability no classical numerical method or computation framework has ever possessed..
- **What holds it back:** neural plasticity and dynamically shifting fixed points constrain stability, making long-horizon continual learning intrinsically difficult.



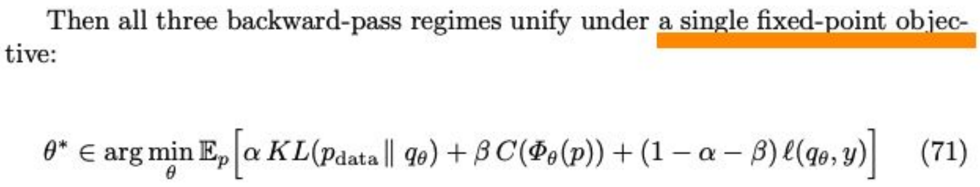




Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

RL and Fixed Point Perturbation

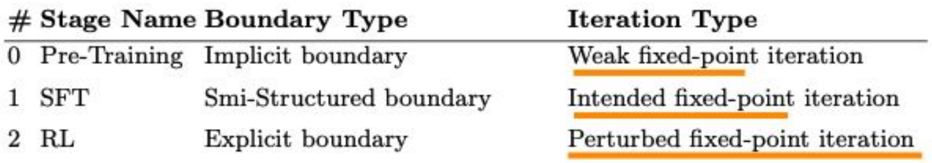
- Without the fixed points, the iteration cannot rely on internal geometry; it must be guided exclusively by **boundary conditions.**
- Without **symmetric** boundaries, the iteration must rely on **weak, Soft and Discrete** boundary conditions.
- “Weak and Soft”: numerical **perturbations** are minimal, yet globally the system remains highly adaptable.
- Discrete: Even when boundary conditions are discrete, sparse, or non-strict, the network can still reconstruct its **intrinsic integral** trajectory and locate new **fixed points**.



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

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

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

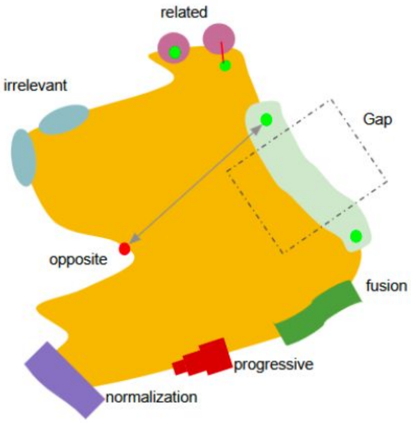


Fig. 14: Symmetric Boundary Condition

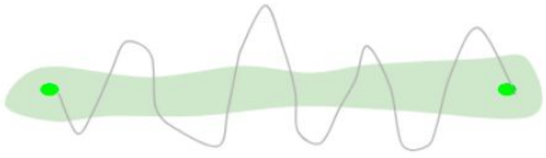


Fig. 15: Weak and Discrete Boundary Condition

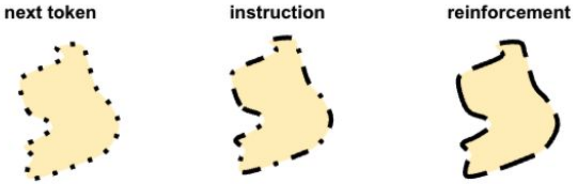
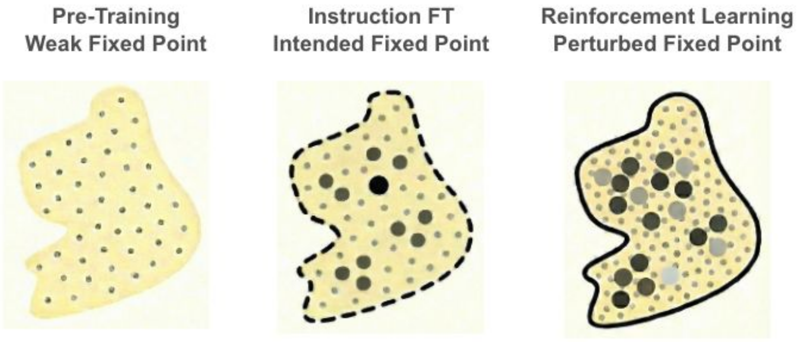


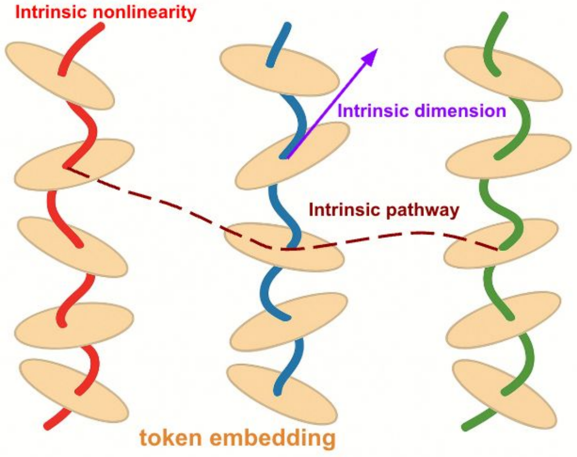
Fig. 13: Foundation Model Boundary Conditions

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

Fixed Point Progression

- In practice, a *fixed point* corresponds to what we call a *solution, answer, or decision*.
- A **Weak** *Fixed Point* does **not** imply the point itself is weak; rather, its *intrinsic convergence pathway* is **weak and diffuse**.
- An **Intended** *Fixed Point* serves as an **anchor**, shaping the convergence basin through explicit boundary constraints.
- A **Perturbed** *Fixed Point* is often the *most* **effective**, as weak and stochastic boundary conditions activate richer convergence pathways.
- Under **high-order nonlinearity**, *perturbation* becomes *essential* for fixed-point iteration to converge.
- The effectiveness of *Reinforcement Learning* arises from boundary conditions that are **symmetric, weak, and discrete**, which delay **neural plasticity**

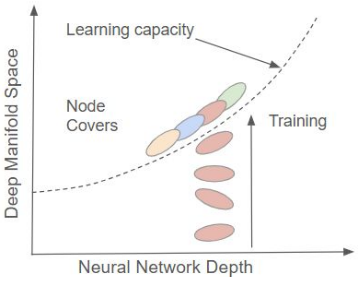


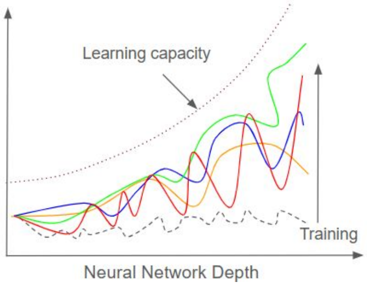


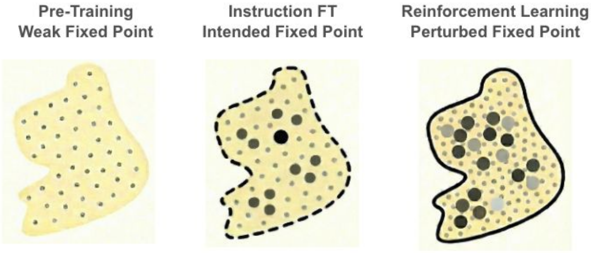
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Trainability, Learnability and Neural Plasticity

- **Unbounded** trainability via coordinate unlimited change
  - Because node coordinates (manifold covers) change at every iteration, neural networks can be trained indefinitely.
- Trainability ≠ **Learnability**
  - Continued optimization may yield no new learning and can induce misalignment with previously learned structures.
- Neural **plasticity** is the bottleneck
  - Delaying neural plasticity is a critical objective in both architecture design and training strategy
- **Learning Complexity**
  - Data-induced complexity: high-order nonlinearity and near-infinite data scope
  - **Architecture-induced complexity: rigid neural network architectures**
  - Boundary-induced complexity: asymmetric, strong, or pointwise boundary conditions
  - Optimization-induced complexity: batch structure and learning rate schedule



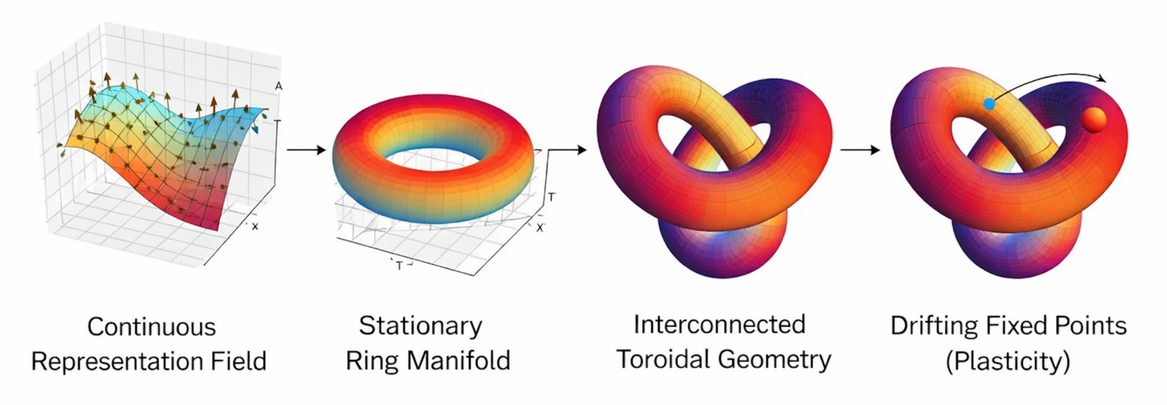




Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Stationary, Symmetry and Neural Plasticity

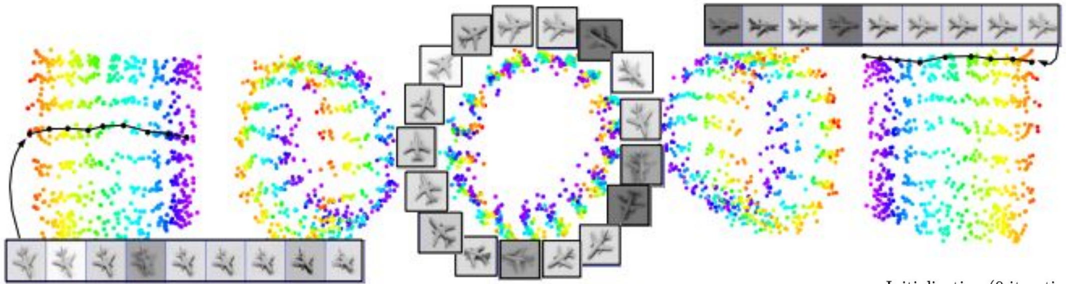
- **Stationarity** arises when real-world data imposes many incompatible constraints; under stochastic training these constraints cancel statistically, producing degenerate directions along which the variational gradient vanishes.
  - Stationarity in Deep Manifold is exactly the variational condition that defines neural fixed points: a neural fixed point is a stationary solution of the implicit **Lagrangian** under weak, data-induced boundary conditions.
- **Symmetry (Ring)** emerges when such degenerate stationary directions admit continuous transformations that leave the functional invariant, causing stationary solution sets to organize into closed level-set manifolds, observed as rings or shells.
  - **Interconnected Toroidal Geometry** arises when multiple ring-like stationary manifolds share overlapping data constraints (e.g., *"bank"* with multiple meanings), coupling independent rings into intertwined tori and enabling semantic superposition and neural plasticity.
- **Neural plasticity** arises from interconnected toroidal structures because they create extended, coupled flat directions in which small boundary perturbations move the solution along the stationary manifold rather than restoring it to a unique configuration.


A
T
x
T
X
T
Continuous Representation Field
Stationary Ring Manifold
Interconnected Toroidal Geometry
Drifting Fixed Points (Plasticity)

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

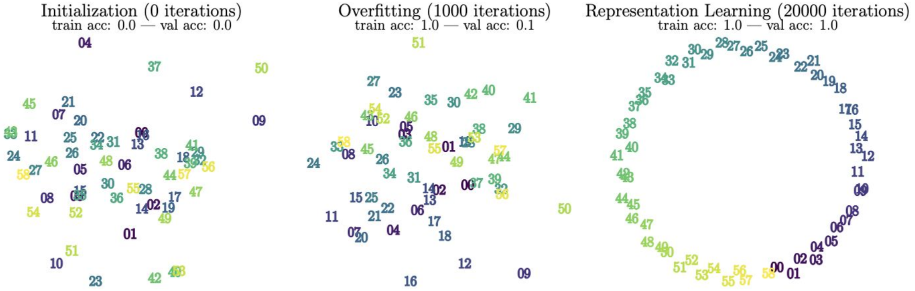
Is the ring the most stable stationary manifold?

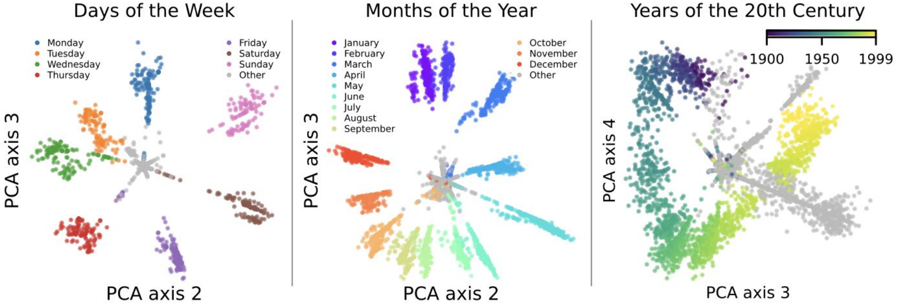
Locally **yes**, globally **no**. The ring is locally the most stable stationary manifold under isotropic constraints, but global stability breaks when multiple invariances couple into higher-order geometry.



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)



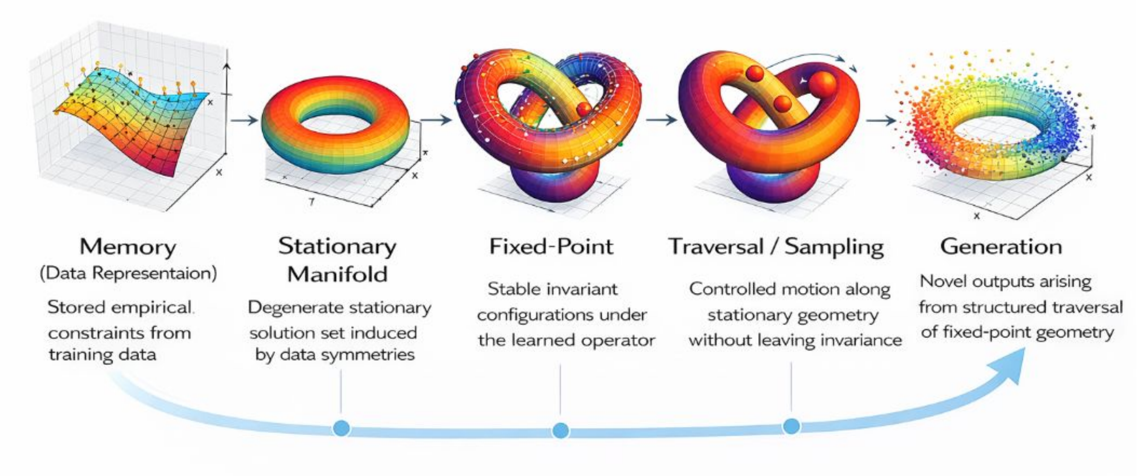


Not All Language Model Features Are One-Dimensionally Linear, (J. Engels… Max **Tegmark**, 2024)

Training Complexity: From Memorization to Generalization

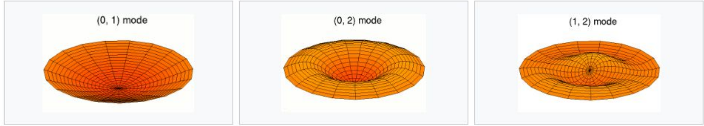
Geometry and Fixed Point In Deep Learning

- Training is geometric evolution, not parameter optimization.
  - Learning reshapes the admissible solution geometry rather than minimizing a fixed function.
- Data induces constraints; geometry absorbs them.
  - Empirical data progressively deforms the representation manifold into stationary structures.
- Stationarity defines what is learned.
  - Learned knowledge corresponds to stationary manifolds and their fixed-point sets under the induced operator.
- Generalization is geometric traversal.
  - Generation arises from structured movement within invariant geometry

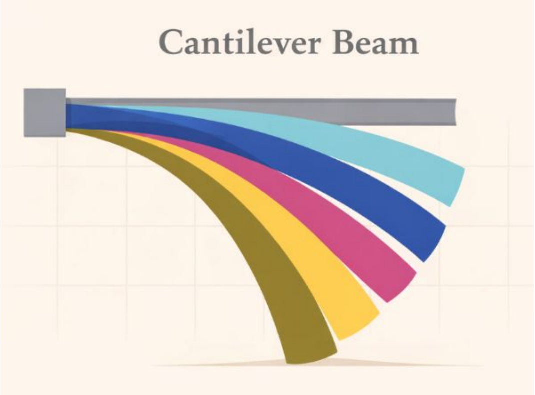

Memory
(Data Representaion)
Stored empirical constraints from training data
Stationary Manifold
Degenerate stationary solution set induced by data symmetries
Fixed-Point
Stable invariant configurations under the learned operator
Traversal / Sampling
Controlled motion along stationary geometry without leaving invariance
Generation
Novel outputs arising from structured traversal of fixed-point geometry

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

Bigger Models Are Not the Ultimate Solution



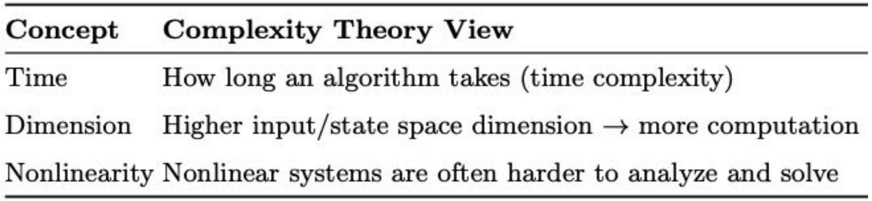
- Why scaling works
  - Larger models increase degrees of freedom
  - Enable high-order nonlinearity
  - Absorb near-infinite data scope and diversity
- Why scaling alone fails — Babuška’s Paradox
  - Pure scale-out introduces accumulated numerical error
  - Massive node interactions amplify instability
  - More parameters ≠ better solution when geometry drifts
- Transformer: still the best we have
  - Locally rigid, structurally fragile
  - MoE: more freedom, implicit federation — fragility remains
- Diffusion has own beauty
  - Globally smooth and robust, computationally expensive
  - Not structurally scalable as a universal solver
- Play-doh like elasticity models
  - Soft, large deformable geometry
  - Stable under high-order nonlinearity
  - Scales with data complexity, not just parameter count



Model CAP Theorem

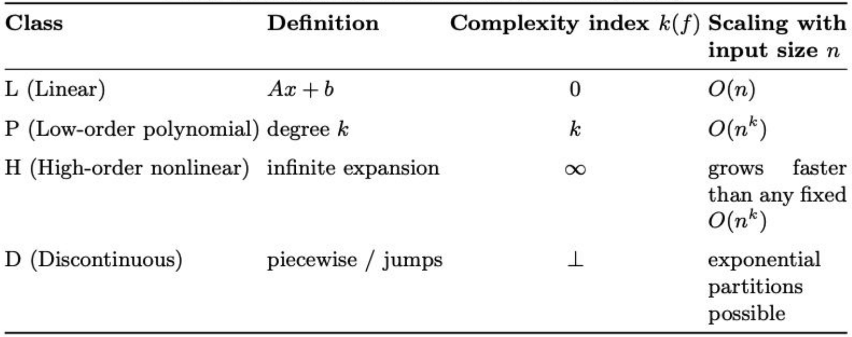
Single-forward pass: only 2 of {Coverage, Accuracy, Performance} can be optimize.

- Coverage denotes how much of the real–world manifold the model attempts to represent across semantic, symbolic, temporal, modal, and nonlinear axes.
- Accuracy denotes local geometric fidelity: curvature alignment, fixed–point stability, and residual correctness within each manifold slice.
- Performance denotes the numerical efficiency of inference.

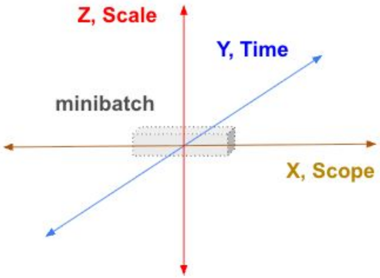


Concept	Complexity Theory View
Time	How long an algorithm takes (time complexity)
Dimension	Higher input/state space dimension → more computation
Nonlinearity	Nonlinear systems are often harder to analyze and solve

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

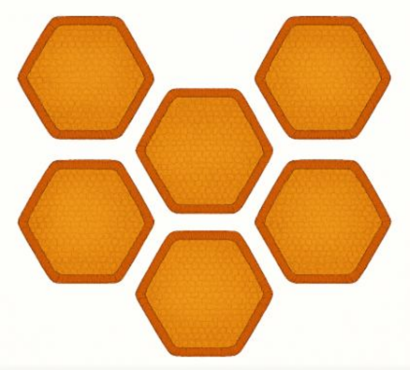


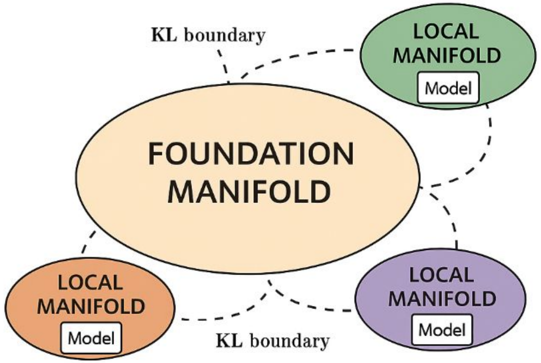
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Manifold Federation: Real World and World Model

The Future Is Mini-AGI Federation

- Data **complexity** induces learning complexity
- Learning complexity arises from dat high-order **nonlinearity**
- **Federation** itself is a manifold concept
  - Each model corresponds to a distinct **manifold**.
  - Like an ML **ensemble**, each model captures a different aspect of high-order nonlinearity
- **Local** federation operates as a **mosaic** of **small elastic** models
- **Global** federation constitutes deep manifold federation learning



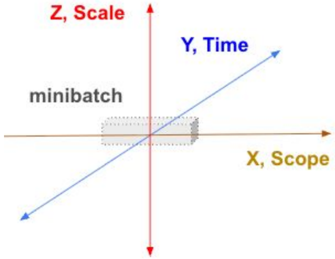


Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

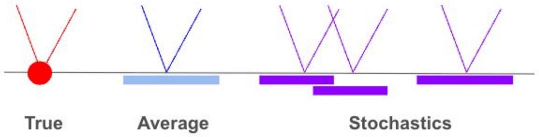
AI for Science and Engineering

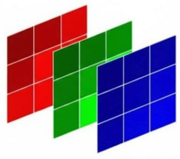
AI for Science and Engineering

Neural network property-lessness underpins its **universal learnability.**



1. Extreme Learnability of Neural Networks
   a. **Forward–Inverse** Unified Iteration
   b. **Propertyless**: Any data type, any boundary condition
   c. **Loss functions**: without any governing physical equations or mathematical abstractions.
   d. **Unbounded Learning Space:** Effective across resolutions, domains, data scopes, and modalities through a unified numerical representation: property-lessness(token embedding)
   e. **Stochastic Fixed Points:** much more close to the real world than human own interpretations.
2. What Enables
   a. **Inverse Problems** become natural and effortless.
   b. Applying **any boundary** conditions can speed up by up to 10,000,000x (10M) faster.
   c. **Discoverability**: Interpolation and extrapolation merge along manifold trajectories.
   d. **Constitutive Modeling**: Governing relations emerge from directly data with observation only.
3. PINN and Neural Operator
   a. PINN and numerical computation are complementary; they cannot replace one another.
   b. For highly complex problems, Neural Operators do not go far.



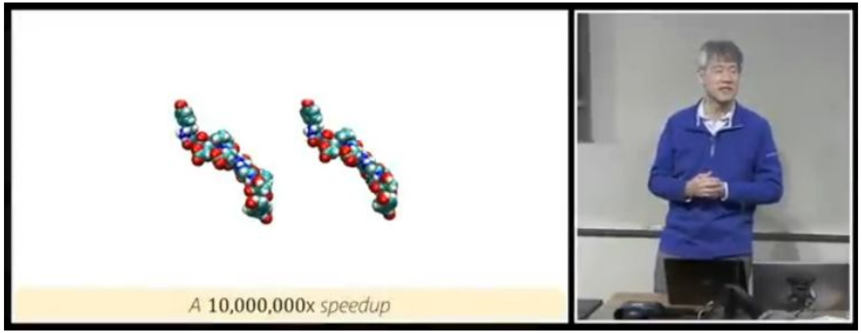


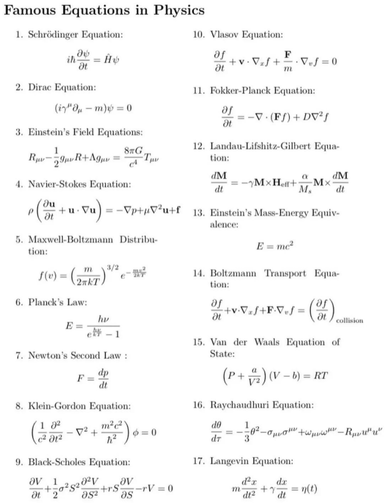
Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592;   Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

AI4SE: Speed Up

- Up to 10Mx
- Explicit and Implicit coupling

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:

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2S^2\frac{\partial^2V}{\partial S^2}+rS\frac{\partial V}{\partial S}-rV=0$$

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:

$$\frac{d\mathbf{M}}{dt}=-\gamma\mathbf{M}\times\mathbf{H}_{\mathrm{eff}}+\frac{\alpha}{M_s}\mathbf{M}\times\frac{d\mathbf{M}}{dt}$$

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:

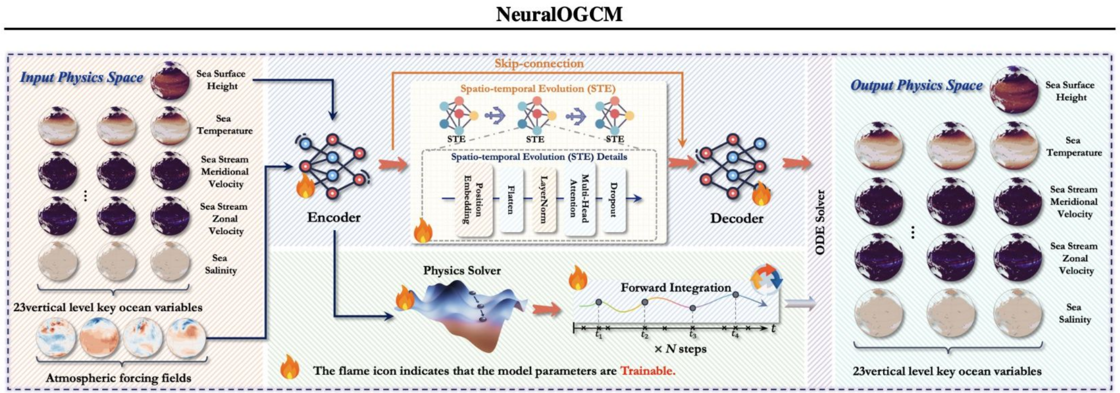
$$\left(P+\frac{a}{V^2}\right)(V-b)=RT$$

16. Raychaudhuri Equation:

$$\frac{d\theta}{d\tau}=-\frac{1}{3}\theta^2-\sigma_{\mu\nu}\sigma^{\mu\nu}+\omega_{\mu\nu}\omega^{\mu\nu}-R_{\mu\nu}u^{\mu}u^{\nu}$$

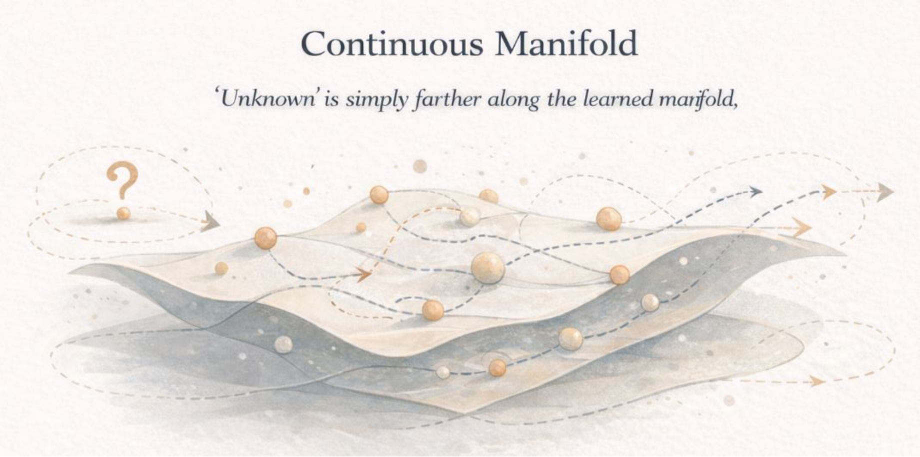
17. Langevin Equation:

$$m\frac{d^2x}{dt^2}+\gamma\frac{dx}{dt}=\eta(t)$$


NeuralOGCM
Input Physics Space
Sea Surface Height
Sea Temperature
Sea Stream Meridional Velocity
Sea Stream Zonal Velocity
Sea Salinity
23vertical level key ocean variables
Atmospheric forcing fields
Skip-connection
Spatio-temporal Evolution (STE)
STE
STE
STE
Spatio-temporal Evolution (STE) Details
Position Embedding
Flatten
LayerNorm
Multi-Head Attention
Dropout
Encoder
Decoder
Physics Solver
Forward Integration
t_1
t_2
t_3
t_4
t
× N steps
ODE Solver
The flame icon indicates that the model parameters are Trainable.
Output Physics Space
Sea Surface Height
Sea Temperature
Sea Stream Meridional Velocity
Sea Stream Zonal Velocity
Sea Salinity
23vertical level key ocean variables

Discoverability: Interpolation and Extrapolation

- Neural networks implement no explicit or implicit knowledge boundary.
- Interpolation and extrapolation collapse into a single process: manifold traversal.
- How Discovery Emerges:
  - Near-infinite data scope continuously expands
  - Propertyless representations allow unrestricted coordinate reshaping
  - Boundary-conditioned iteration guides motion along intrinsic pathways
- Neural networks do not cross knowledge boundaries because no knowledge boundary exists.
- “Extrapolation” is simply stable traversal until constraints fail.
- Discovery is geometric inevitability, not symbolic reasoning



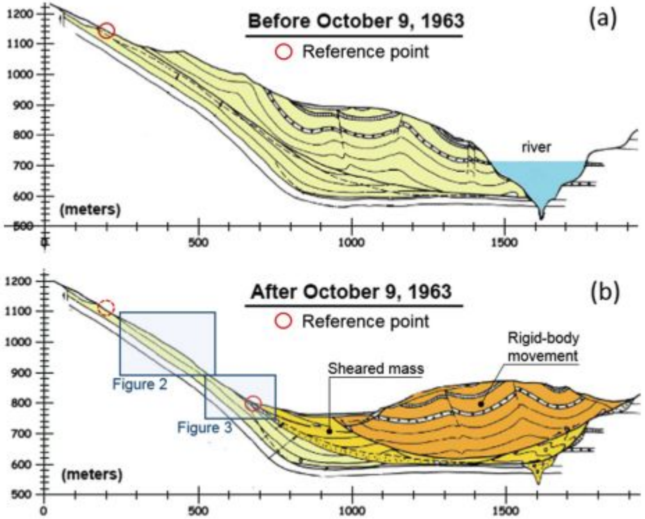
Data-Driven Constitutive Modeling

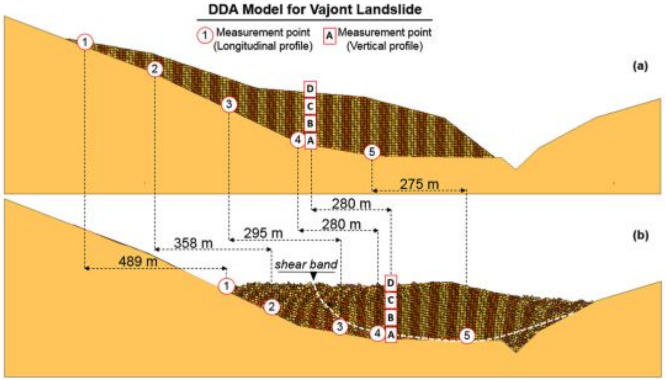
- Constitutive law = intrinsic manifold
  - Physical systems obey intrinsic relations (e.g. stress–strain)
  - In high-order nonlinear regimes, no closed-form law exists
  - Neural networks learn this relation as a constitutive manifold
- From equation to geometry
  - Law becomes learned geometry, not prescribed formula
  - $\sigma = f(\varepsilon)\,(\varepsilon,\sigma)\in\mathcal{M}_\theta,\ \mathcal{M}_\theta=\{(\varepsilon,f_\theta(\varepsilon))\}$
- Fixed-point interpretation
  - Constitutive update treated as a fixed-point problem
  - Training minimizes residuals to recover intrinsic law
  - Learned behavior stabilizes as a numerical fixed point
- Projection onto learned manifold
  - Observations are projected onto $\mathcal{M}_\theta$
  - Update selects the nearest admissible constitutive state
  - Robust to noise, plasticity, and incomplete measurements

Numerical Computation Limitation

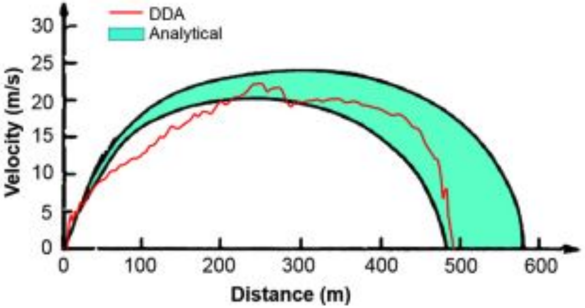
- Equations are assumed; the world is not
  - Classical solvers require a known governing operator; real systems are partially unknown, drifting, or unmodeled.
- Geometry is fixed, not learned
  - Mesh / basis / coordinate choice is external. The solver cannot create new representational geometry when reality demands it.
- Uncertainty is treated as noise, not a boundary condition
  - Stochasticity is usually appended (error bars, turbulence models), rather than integrated as a first-class constraint shaping admissible solutions.
- Complexity is offloaded into “closure terms”
  - When physics is missing, we add friction factors, constitutive laws, subgrid models, handcrafted patches that do not scale.
- No scalable mechanism for multi-regime transition
  - Discontinuities, phase changes, damage, and bifurcations force solver switching, remeshing, or model rewriting.

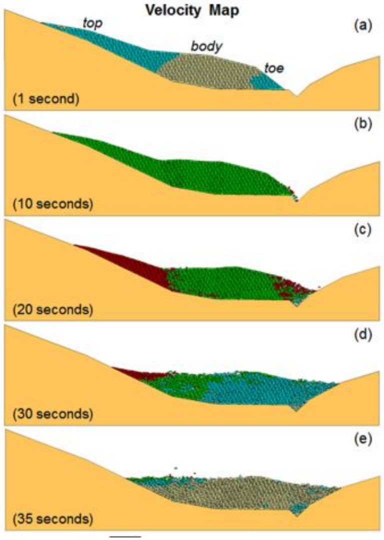
Vajont Landslide, 1963





DDA Model for Vajont Landslide
1 Measurement point (Longitudinal profile)
A Measurement point (Vertical profile)
(a)
(b)
1
2
3
4
5
A
B
C
D
275 m
280 m
280 m
295 m
358 m
489 m
shear band



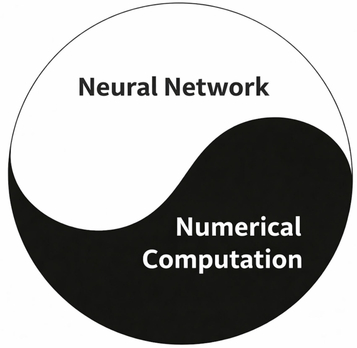


- Rapid sliding and friction degradation: Lessons from the catastrophic Vajont Landslide, J. Ibañeza,, Y. Hatzorb, 2018, Engineering Geology
- Discontinuous Deformation Analysis (DDA)

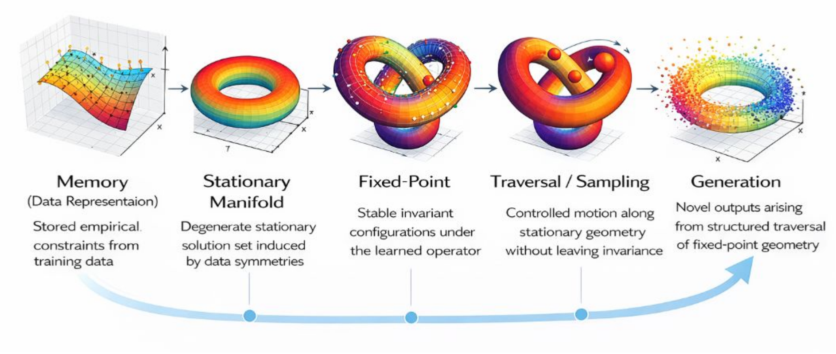
Coupling Approach

Couple the globally learned geometry into local, site-specific numerical fixed-point solutions

- AI excels at global generalization
  - Learns from all historical and simulated cases
- AI is propertyless
  - Handles heterogeneous data (images, text, signals, logs)
- Physics-grounded solvers like DDA, FEM and NMM
  - Naturally support inverse problems, discontinuities, and strong nonlinearity



Neural networks as learnable numerical computations and physics-grounded numerical computations complement each other rather than replace each other.



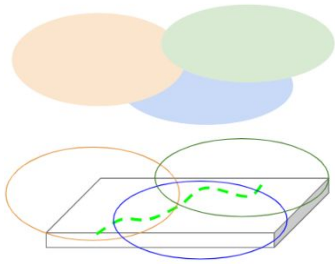
Memory (Data Representaion)
Stored empirical constraints from training data

Stationary Manifold
Degenerate stationary solution set induced by data symmetries

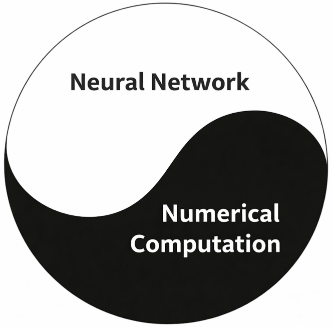
Fixed-Point
Stable invariant configurations under the learned operator

Traversal / Sampling
Controlled motion along stationary geometry without leaving invariance

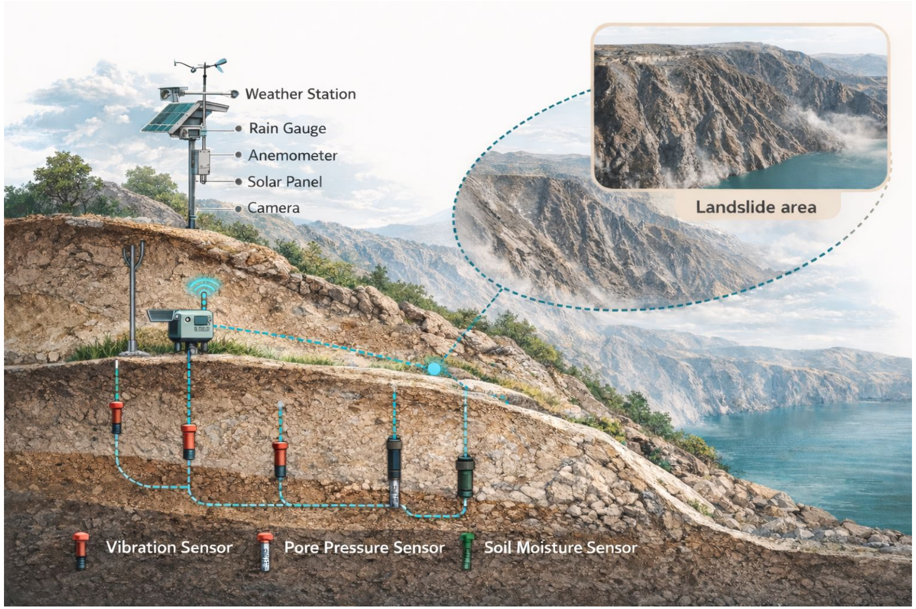
Generation
Novel outputs arising from structured traversal of fixed-point geometry

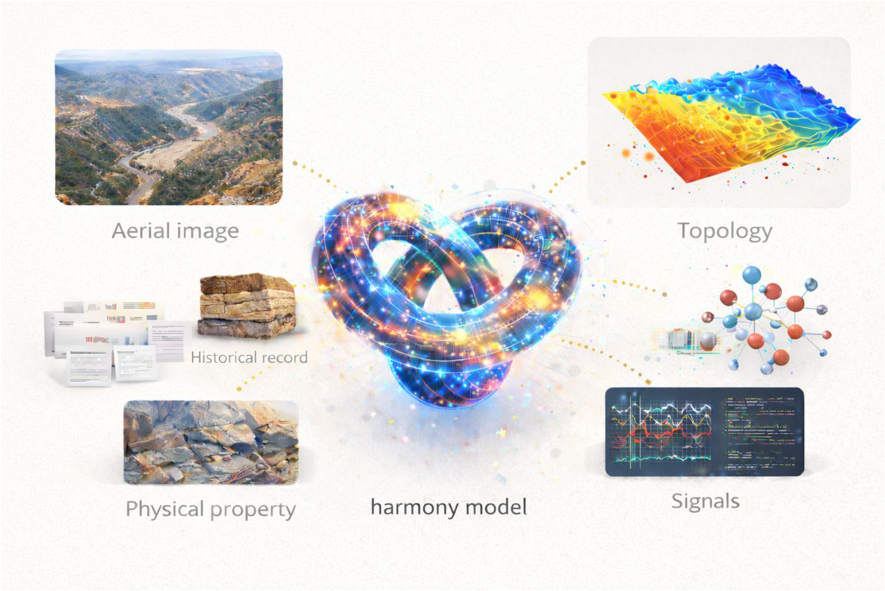


Harmony Model



- Neural Networks
  - Learnability across Data Types & Global Generalization
- Numerical Computation
  - Physics-Grounded Accuracy, True Fixed Points & Explicit Mechanism
- Numerical computation exposes the mechanism; neural networks learn the geometry.





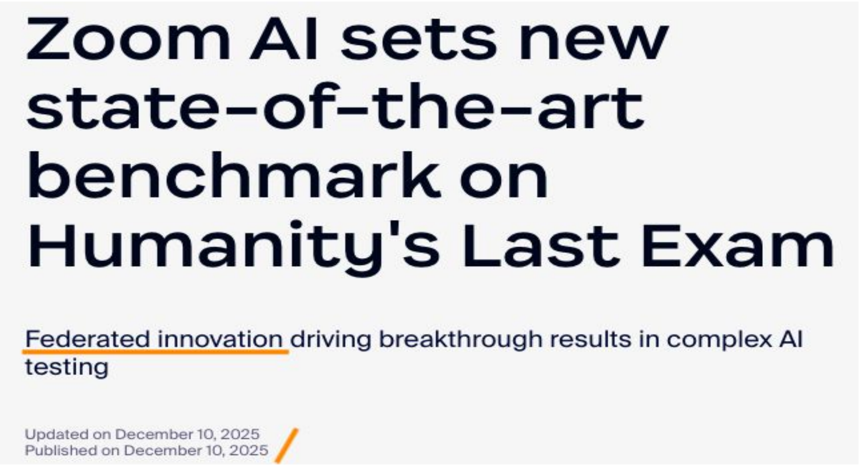
Not Coincidence

Part 2 was officially published on December 6, 2025

Federated AI

Deep Manifold

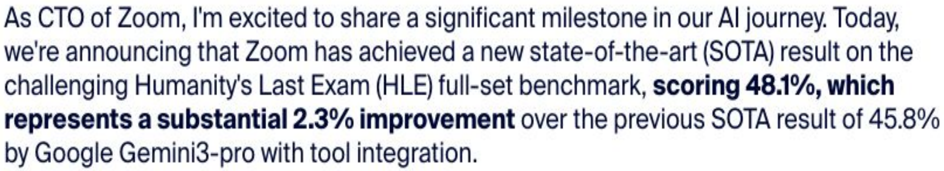
- The future is mini-AGI federation because high order nonlinear data



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

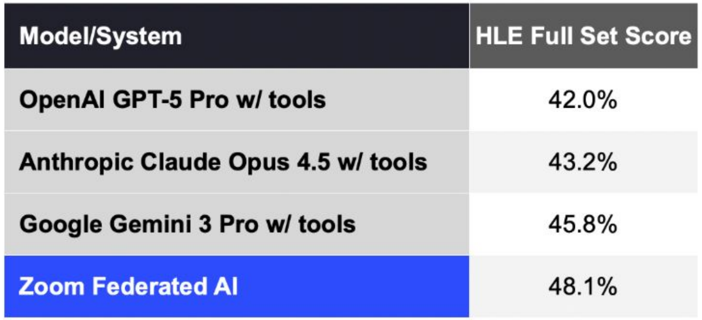
Federated innovation driving breakthrough results in complex AI testing

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

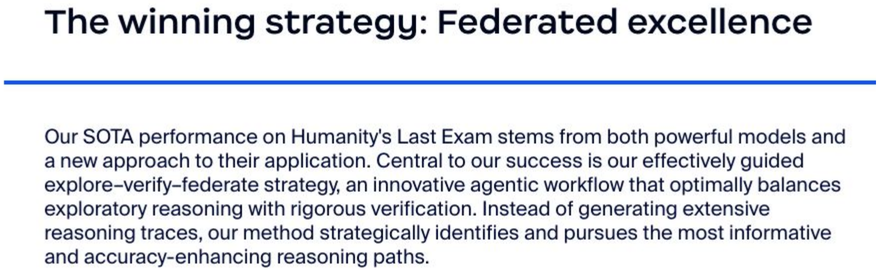


As CTO of Zoom, I'm excited to share a significant milestone in our AI journey. Today, we're announcing that Zoom has achieved a new state-of-the-art (SOTA) result on the challenging Humanity's Last Exam (HLE) full-set benchmark, **scoring 48.1%, which represents a substantial 2.3% improvement** over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



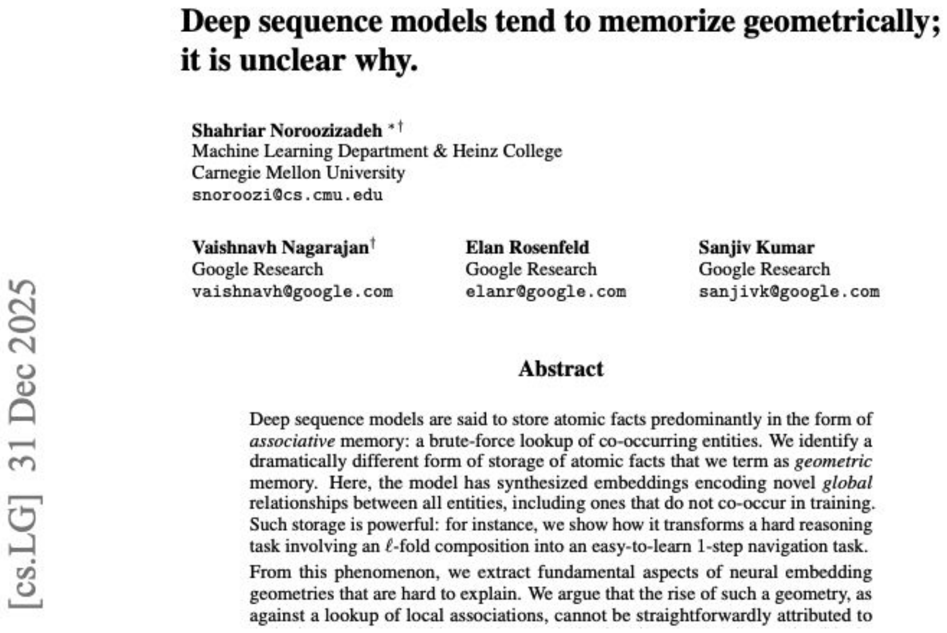
The winning strategy: Federated excellence

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.

Neural Networks Geometry

Deep Manifold

- Neural network behavior is governed by the geometry of its learned world representation.



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

[cs.LG] 31 Dec 2025

Abstract

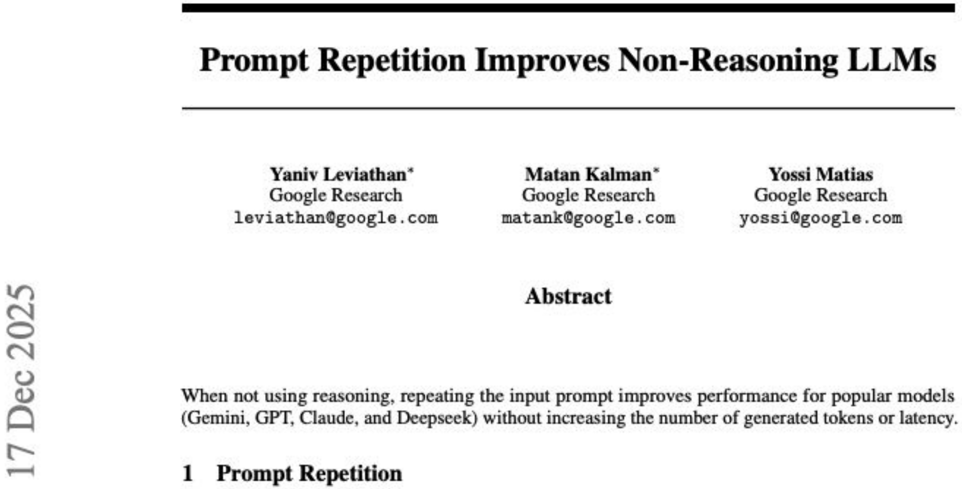
Deep sequence models are said to store atomic facts predominantly in the form of *associative* memory: a brute-force lookup of co-occurring entities. We identify a dramatically different form of storage of atomic facts that we term as *geometric* memory. Here, the model has synthesized embeddings encoding 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

Prompt Repetition

Deep Manifold

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



17 Dec 2025

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

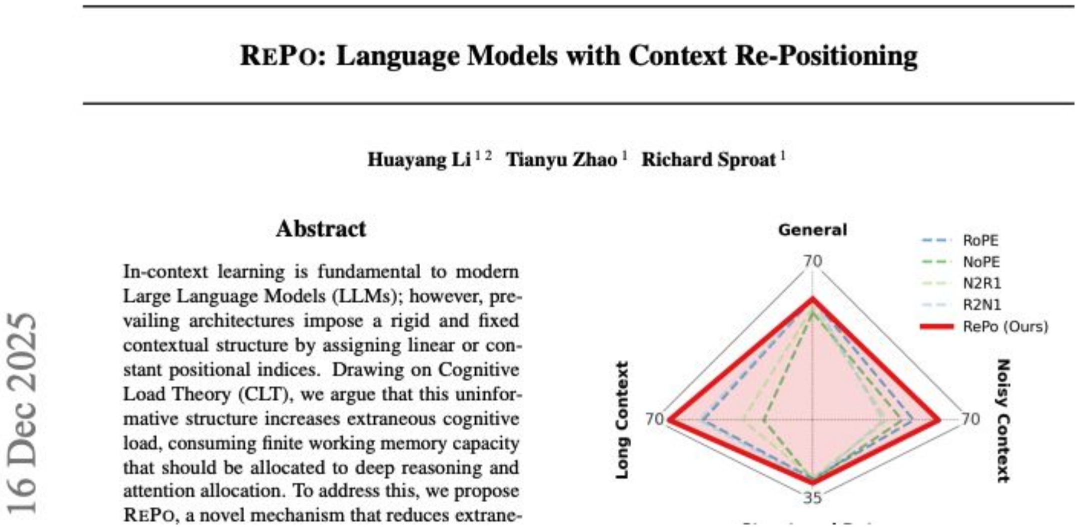
When not using reasoning, repeating the input prompt improves performance for popular models (Gemini, GPT, Claude, and Deepseek) without increasing the number of generated tokens or latency.

1 Prompt Repetition

No Position Embedding

Deep Manifold

Positional embeddings may not be strictly necessary. The reason they persist is that they act as a 'glue' for the high-dimensional embedding space, providing a structural anchor for the model



REPO: Language Models with Context Re-Positioning

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

16 Dec 2025

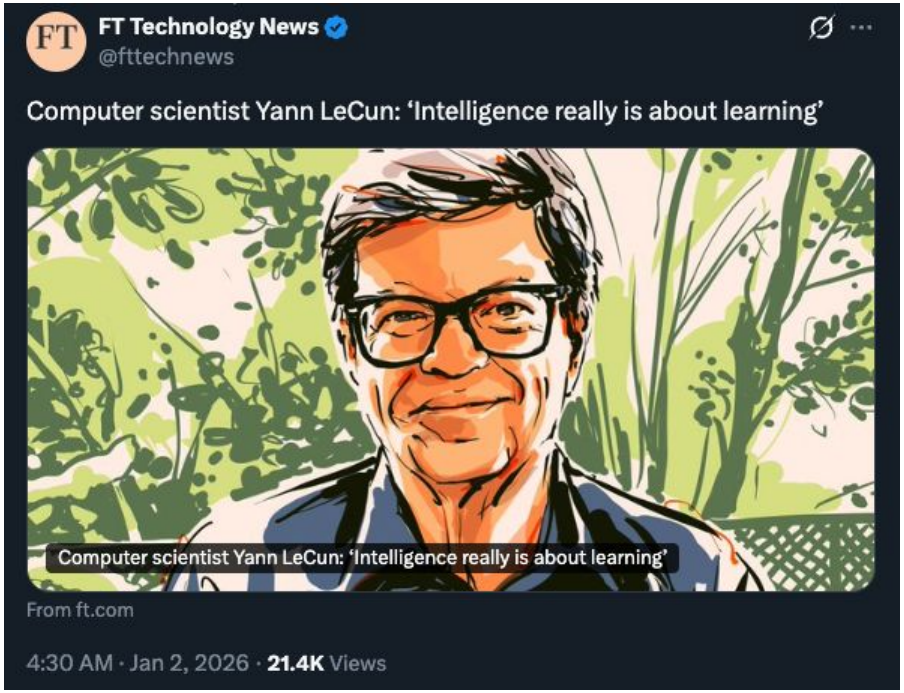
Abstract

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-

Learnability

Deep Manifold

Neural Network is a learnable numerical computation



FT Technology News
@fttechnews

Computer scientist Yann LeCun: 'Intelligence really is about learning'

From ft.com

4:30 AM · Jan 2, 2026 · **21.4K** Views

Datadualism

Datauaiism

- Core Thesis
  - Datauaiism holds that intelligence and knowledge arise from fidelity to data, not from human-designed meaning, intent, or structure.
- Doctrinal Principles
  - Data as the primary authority
    - Data itself is the ultimate source of truth. It is inherently high-order nonlinear and spans a near-infinite scope, far exceeding what can be exhaustively specified by human-designed rules or theories.
  - Neural networks as learnable numerical computation
    - Neural networks are best understood as learnable numerical computation, not as systems governed by known physical laws, moral frameworks, or symbolic principles. Model behavior is not prescribed; it is entirely learned from data.
  - Mathematical foundations
    - The mathematical structure underlying datauaiism is grounded in three pillars:manifold coverings, fixed-point theory, and calculus. Together, they describe how learning systems represent complexity, stabilize behavior, and evolve through computation.
  - Doctrinal humility
    - As a doctrine—by definition a school of thought—datauaiism does not claim inherent correctness. It is not true by fiat, but justified only insofar as it continues to explain and predict empirical outcomes.
  - Role of human knowledge
    - Datauaiism does not exclude human insight or domain knowledge in addressing specific real-world problems. Rather, it treats such knowledge as contextual scaffolding, not as a universal source of authority.

mHC and Muon

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

- Do not fully align with the doctrinal position of datauallism. One that introduces explicit structural constraints, g(θ)
- Not to say that these methods are incorrect or ineffective. Rather, they reflect a different philosophical stance
- his limits gradient alignment across strongly repeated signals, preventing excessive collapse toward a single fixed-point geometry during fine-tuning and subsequent training phases.

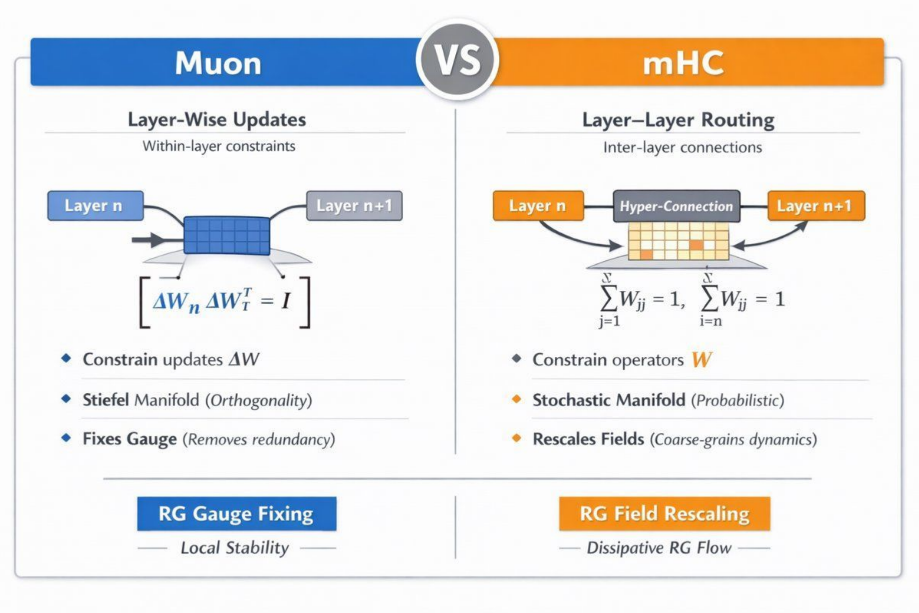

Muon
VS
mHC
Layer-Wise Updates
Within-layer constraints
Layer n
Layer n+1
$\left[\Delta W_n \, \Delta W_T^T = I \right]$
Constrain updates ΔW
Stiefel Manifold (Orthogonality)
Fixes Gauge (Removes redundancy)
Layer–Layer Routing
Inter-layer connections
Layer n
Hyper-Connection
Layer n+1
$\sum_{j=1}^{\mathfrak{X}} W_{jj} = 1, \ \sum_{i=n}^{\mathfrak{X}} W_{ij} = 1$
Constrain operators W
Stochastic Manifold (Probabilistic)
Rescales Fields (Coarse-grains dynamics)
RG Gauge Fixing
Local Stability
RG Field Rescaling
Dissipative RG Flow

Image source
@CalcCon

In Lens of Deep Manifold

Information Theory as Boundary Constraint

A Deep Manifold Perspective on Entropy, Compression, and Learning

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

- Entropy ≠ learning objective → it is a capacity boundary
- Compression → consequence of manifold alignment, not its cause
- Learning → stochastic fixed points from geometric constraint cancellation
- Why Minimum Description Length (MDL) fails for LLMs → description length measures boundary cost, not interior geometry
- What matters instead → intrinsic learning pathways on deep manifolds

Information theory limits representation; Deep Manifold theory explains learning.
Entropy bounds the boundary;geometry computes the solution.

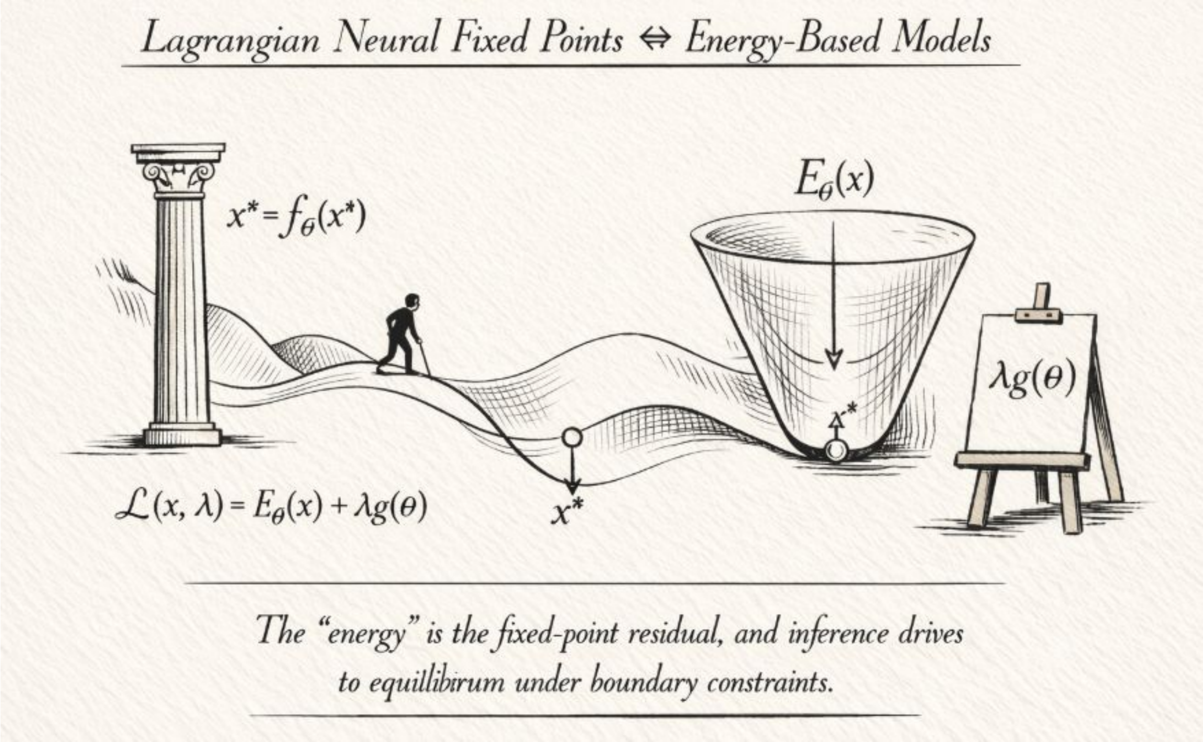
Kolmogorov Complexity: A Beautiful Descriptor

- Kolmogorov complexity asks:
    - What is the shortest program that generates this object?
- Deep Manifold theory asks:
    - What geometric and dynamical constraints make this object a stable solution of a learnable numerical system?
- Kolmogorov theory implicitly assumes: compression ⇒ structure
- Deep Manifold inverts this: structure ⇒ incidental compression
- Neural networks do not minimize description length.
    - They integrate constraints until residuals vanish locally.
    - Compression emerges only after a stable manifold geometry forms.

Kolmogorov complexity bounds description; Deep Manifold theory explains computation. Entropy limits the boundary; geometry determines the solution.

Energy-Based Model

An energy-based model is a neural fixed-point system written in the language of energy: the “energy” is the fixed-point residual, and inference is the numerical act of driving that residual to a stable equilibrium under boundary constraints.

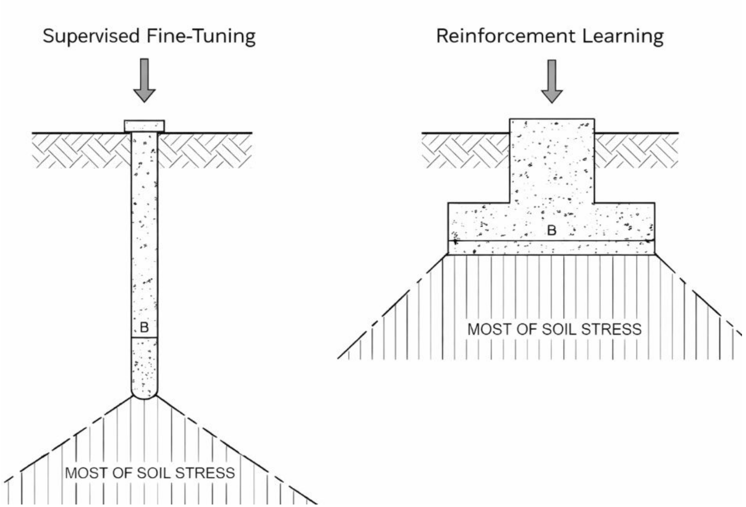


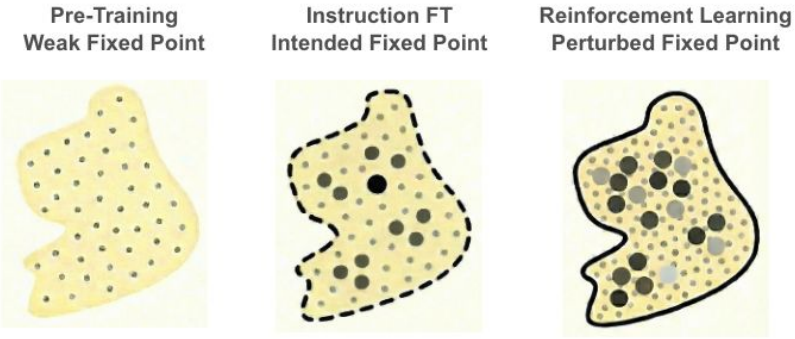
Lagrangian Neural Fixed Points ⇔ Energy-Based Models

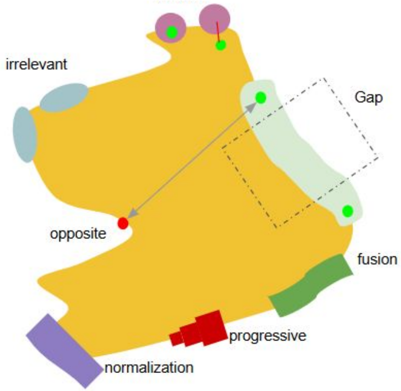
The “energy” is the fixed-point residual, and inference drives to equilibirum under boundary constraints.

SFT: Catastrophic Forgetting

From the Deep Manifold view, pretraining, SFT, and RL correspond to different boundary conditions. SFT does not inherently cause catastrophic forgetting and can be implemented in ways that avoid it.

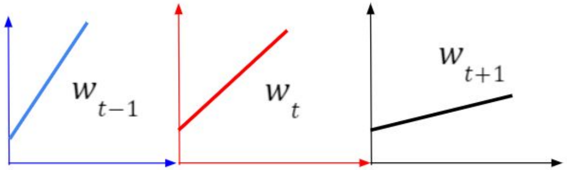




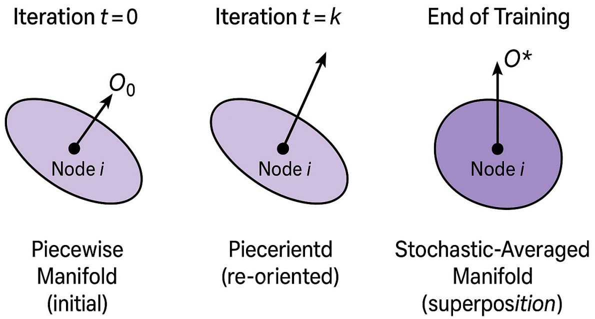


Deep Manifold **Part 1**: **Anatomy** of Neural Network Manifold, arXiv:2409.17592; Deep Manifold **Part 2**: Neural Network **Mathematics**, arXiv:2512.06563

Superposition, Node Cover and Neural Plasticity



Superposition as Stochastic Average Orientation



Iteration $t = 0$

Iteration $t = k$

End of Training

O_0

$O*$

Node *i*

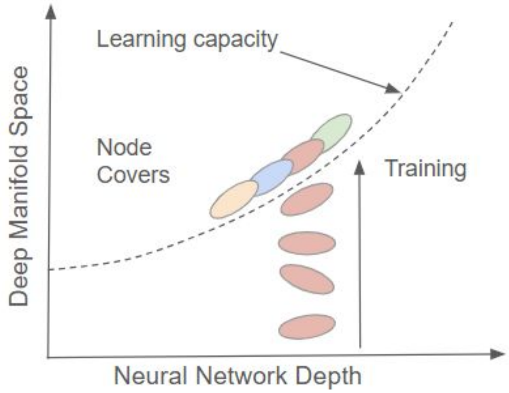
Node *i*

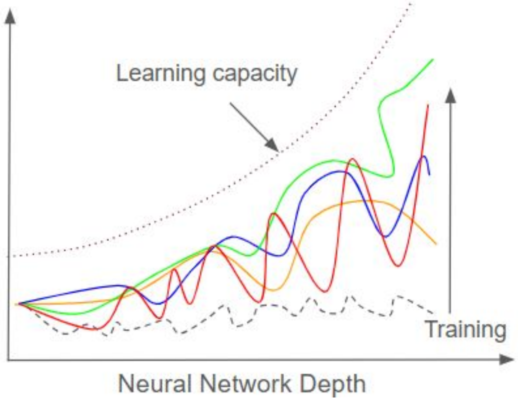
Node *i*

Piecewise Manifold (initial)

Pieceriendt (re-oriented)

Stochastic-Averaged Manifold (superposi*tion*)





- Every iteration rotates or reshapes the node's local manifold orientation.
- Training = continual node-cover re-orientation across stacked manifolds
- Superposition = the stochastic average orientation of piecewise manifolds
- Learning capacity (**scaling law**) is limited by neural plasticity.
- Interconnected tori form inside the stacked piecewise manifolds during training
- These knotted toroidal structures are the geometric source of neural plasticity

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

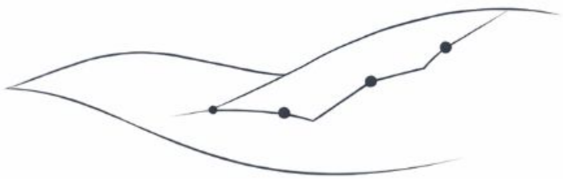
Mechanistic Interpretability

Circuit

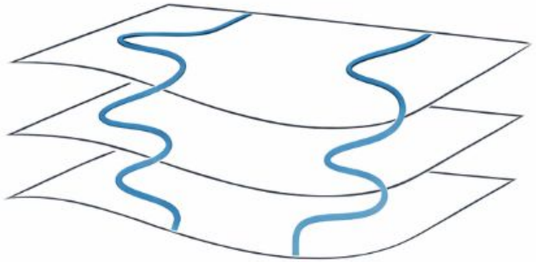
- One smooth continuous surface
- A thin discrete path traveling on the surface
- Nodes or dots along the path
- Geometry is flat and global

Deep Manifold

- Multiple stacked, slightly offset surfaces (piecewise manifolds)
- A continuous curve descending / threading through layers
- Path bends at layer boundaries (integration, not jumps)
- No nodes, no symbols — only flow
- Suggest depth via spacing, not shading



Content: Circuit — Path on a single surface



Content: Nested Integral Pathway — Across stackced manifolds

Category theory and Deep Manifold

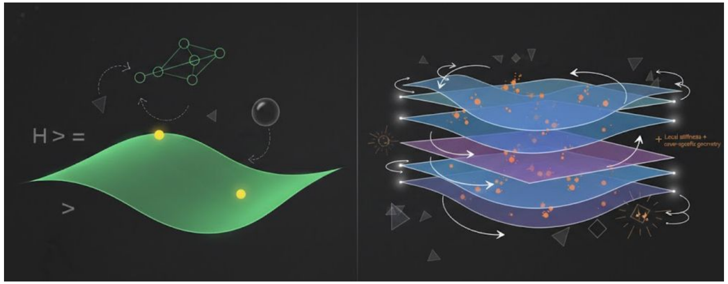
Category theory, like many elegant mathematical theories, offers **beautiful descriptions** but does not directly express the **computable structure of neural networks**. Deep Manifold fills that gap by focusing on the computation and geometry that actually govern learning dynamics.

- Category theory organizes how computations compose; Deep Manifold explains how neural networks compute.
- Category theory excels at compositional syntax, not at modeling high-order nonlinearity, manifold plasticity, or training dynamics.
- Neural networks are not categorical compositions but iterated integrals, whose convergence and behavior are determined by evolving boundary conditions and stochastic fixed points.

$$\begin{tikzcd}
 & TP\otimes\pi^*T^*X \arrow[rr, "\mathrm{d}\lambda_g\otimes(\lambda_g\times 1)"] \arrow[dd, dotted, "c_\pi"] & & TP\otimes\pi^*T^*X \arrow[rr, "\pi^G"] \arrow[dd, "c_\pi"] & & T_GP\otimes T^*X \arrow[dd, "\bar{c}_\pi"] \\
JP \arrow[ur, hook, "i_J"] \arrow[rr, "j\lambda_g"] \arrow[dd, "\pi_J"] & & JP \arrow[ur, hook, "i_J"] \arrow[dd, "\pi_J"] & & \mathrm{CP} \arrow[ur, hook, "i_C"] \arrow[dd, "\pi_{\mathrm{CP}}"] & \\
 & \pi^*(TX\otimes T^*X) \arrow[rr, dotted, "\lambda_g\times 1"] & & \pi^*(TX\otimes T^*X) \arrow[rr, dotted, "\pi^G"] & & TX\otimes T^*X \\
P \arrow[ur, "\mathrm{Id}"] \arrow[rr, "\lambda_g"] & & P \arrow[ur, "\mathrm{Id}"] \arrow[rr, "\pi^G=\pi"] & & X \arrow[ur, "\mathrm{Id}"] &
\end{tikzcd}$$

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

Neural Network Optimization



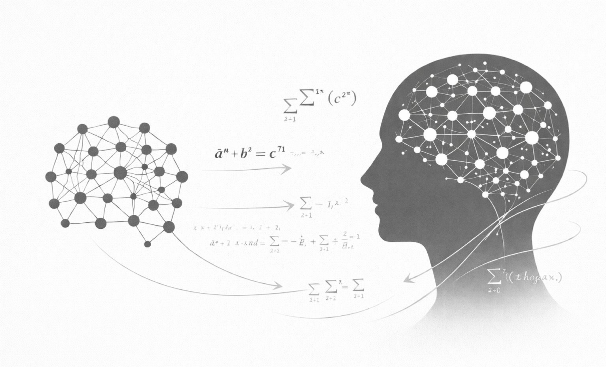
- Classical optimization core thesis
  - Optimization dynamics → implicit regularization → generalization
- Classical optimization requires a well-defined objective function.
  - Neural network training does not optimize a predefined governing function
- Neural network enforces constraints induced by data and architecture, with the loss serving only as a boundary functional.
- Neural network “optimization” is best understood as a numerical transport mechanism for reducing a weak residual induced by data constraints. The loss is a boundary functional defined on empirical constraint sites, not a governing equation of the modeled world.
- In LLMs, the objective is non-stationary and semantically misaligned with truth, so training converges, when it does, to stochastic fixed point shaped by stacked piecewise manifold geometry rather than to a classical optimum.

The Mathematical Lineage of Deep Manifold

Neural Network Mathematics Already Exists

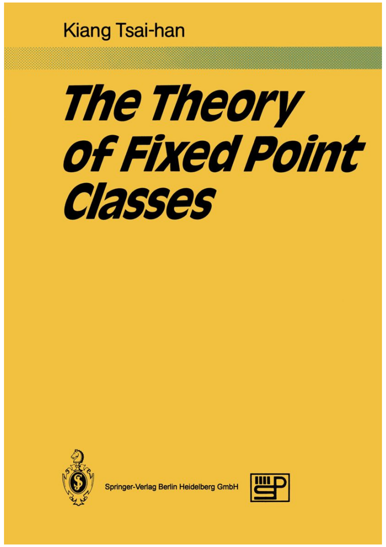
Deep Manifold Only uncovers it

- Neural networks advance mathematics: **concretely through practice, subconsciously through theory**, driven largely by non-mathematicians, AI pioneers, and the broader AI community.
- History repeats itself. Some of the most significant advances in mathematics have historically come from outside the discipline, **pioneered by physicists, engineers, or other scientists**
    1. Pierre de Fermat – Number Theory (1637), background: Lawyer
    2. Blaise Pascal – Probability Theory (1654), background: Physicist, inventor, and philosopher
    3. Isaac Newton – Calculus (1666), background: Physicist and natural philosopher
    4. J´anos Bolyai – Non-Euclidean Geometry (1832), background: Military officer
    5. Claude Shannon – Information Theory (1948), background: Electrical engineer
- Newton developed calculus to express the laws of motion, not to solve abstract math. Today's AI revolution **continues this tradition**.



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

Deep Manifold Mathematical Lineage



Kiang Tsai-han

The Theory of Fixed Point Classes

Springer-Verlag Berlin Heidelberg GmbH

Gen-Hua Shi

60s: Theory of fixed point classes

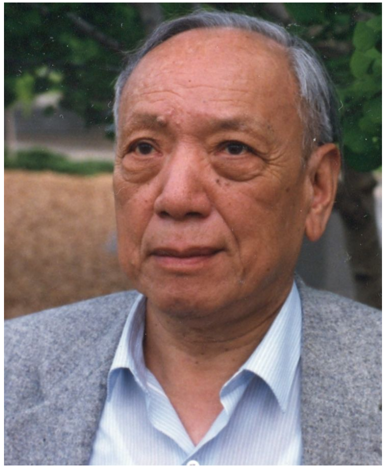
70s: KeyBlock Theory

80s: Discontinuous Deformation Analysis

90s: Numerical Manifold Method

00s: Contact Theory (inequality theory)

The theory of Fixed-point classes resolved the existence, uniqueness, and stability of solutions to differential equations, more than **two centuries** after **Newton** introduced calculus



1980s, Numerical Manifold Method

Professor **Shiing-Shen Chern**, widely regarded as the father of modern differential geometry, served as one of the three members of the Shi PhD dissertation committee (UC Berkeley), providing academic validation for the mathematical framework of numerical manifold method and laying the foundation for the subsequent development of the theory. Chern’s only question was: 'Can stacked piecewise manifolds be extended to any complex domain? He would have been delighted to see the progress in neural networks, as their geometry can be understood as stacked piecewise manifolds.

Mathematicians’ Dreams and Pursuits

- Local and global; continuity and discontinuity; forward and inverse problems

Global/Generalized Mathematics

Differential Geometry is the theory on differentiable manifolds

- **1830**, Évariste Galois, Generalized Number
  - Algebra groups, rings, fields, and Galois theory
- **1892**, Henri Poincaré, Generalized Shapes
  - Covering spaces and geometric laws in algebraic topology, Henri Poincaré
- **1952**, Shiing-Shen Chern, Generalized Functions
  - Manifolds and physical Laws based on cover systems,
  - Differential equations to solve complex physics problems,
  - Mathematical basis for later numerical computing techniques.

Broad Perspective

- Shiing-Shen Chern: how to solve arbitrary complex differential equations
- Shing-Tung Yau: how to solve arbitrary complex differential equations
- Gen-Hua Shi: how to compute complex differential equations

- Computational methods may be naturally non-rigorous, but they are all rooted in a small number of fundamental physical laws.
- These methods are very loosely connected to mathematics, yet they are interlinked.
- It helps humanity understand the world; the same is true for artificial intelligence.
- AI is a learnable numerical computation.

“This is not how mathematicians are trained” (Gen-Hua Shi, 2024.05)

- Simultaneously solving the forward problem in positive time and the inverse problem in negative time has long been a dream of mathematicians, yet they never knew where to begin. Neural networks, however, handle this problem naturally.
- Variables, coefficients, and even coordinate systems evolve continuously—nothing remains fixed. This is not how mathematicians are trained, and such a design could not have come from mathematicians.
- Mathematicians approach composite functions beyond two layers with great caution, mindful of numerous subtle pitfalls; neural networks, by contrast, address them with near nonchalance.
- Neural networks possess stacked mathematical coverings, whereas numerical manifolds typically involve only three to four layers of coverage. By contrast, neural networks stack hundreds or even thousands of such layers. I never imagined anyone would push this so far.