Injective and Invertible LLMs

From practice to theory

SAPIENZA
UNIVERSITÀ DI ROMA

EPFL

ARCHIMEDES

TOMMASO MENCATTINI

MSc Data Science @ EPFL

Research Intern @ Gladia / ISTA

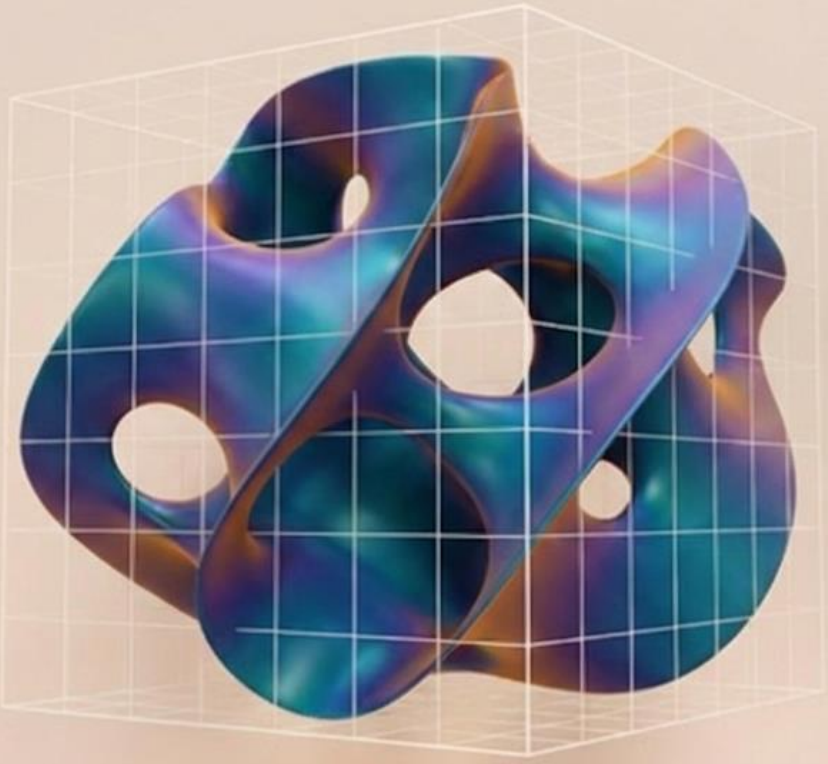
GIORGOS NIKOLAOU

MSc Data Science @ EPFL

Research Intern @ Gladia

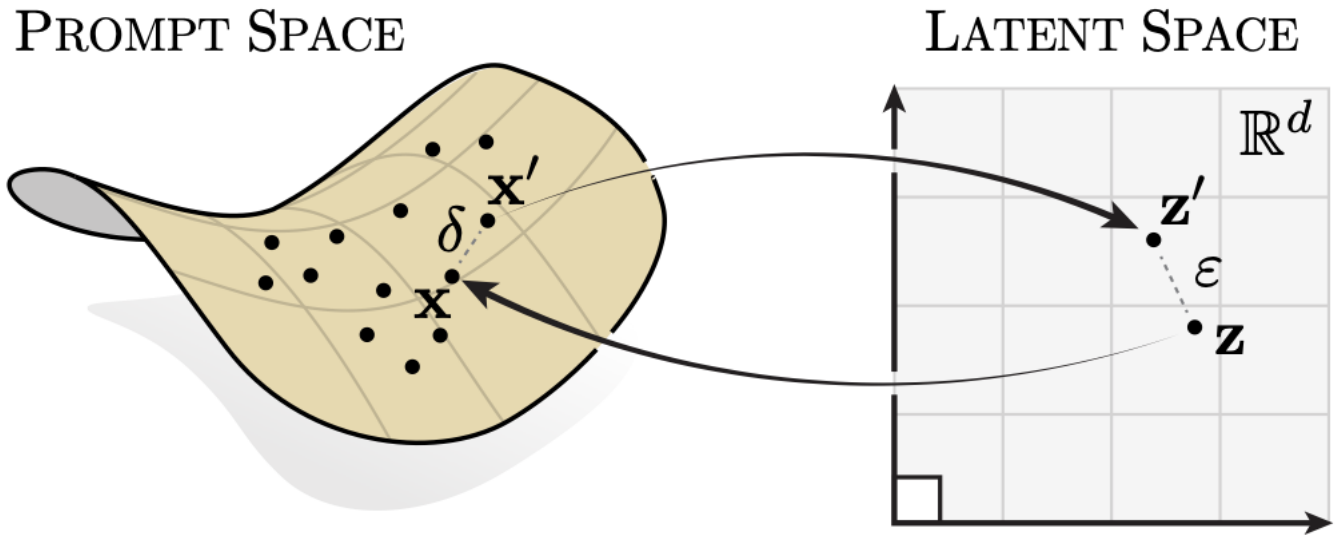
Data Scientist @ Logmind

Main Characters



Language Models are Injective and Invertible

- We **prove** and **verify empirically** that LLMs are injective and invertible!



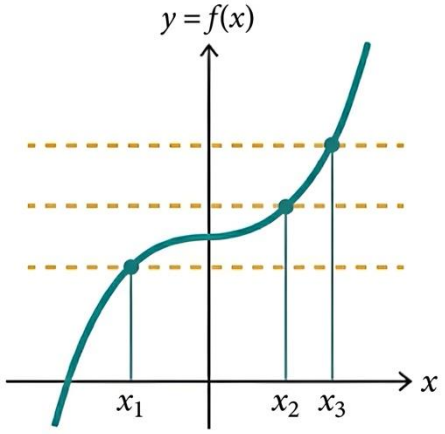
$$\delta > 0 \implies \varepsilon > 0$$

1. What we mean with Injective?
2. What we mean with Invertible?
3. What function is an LLM?

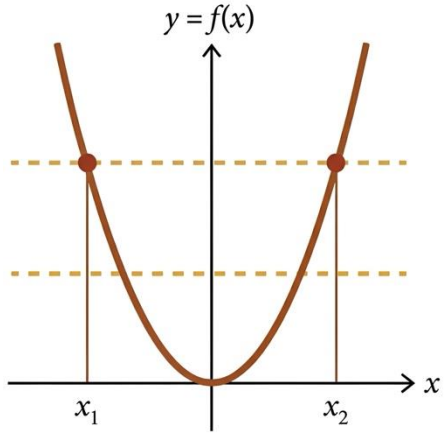
Injective Functions

Injectivity: Distinct inputs map to distinct outputs!

Injective



Non-Injective

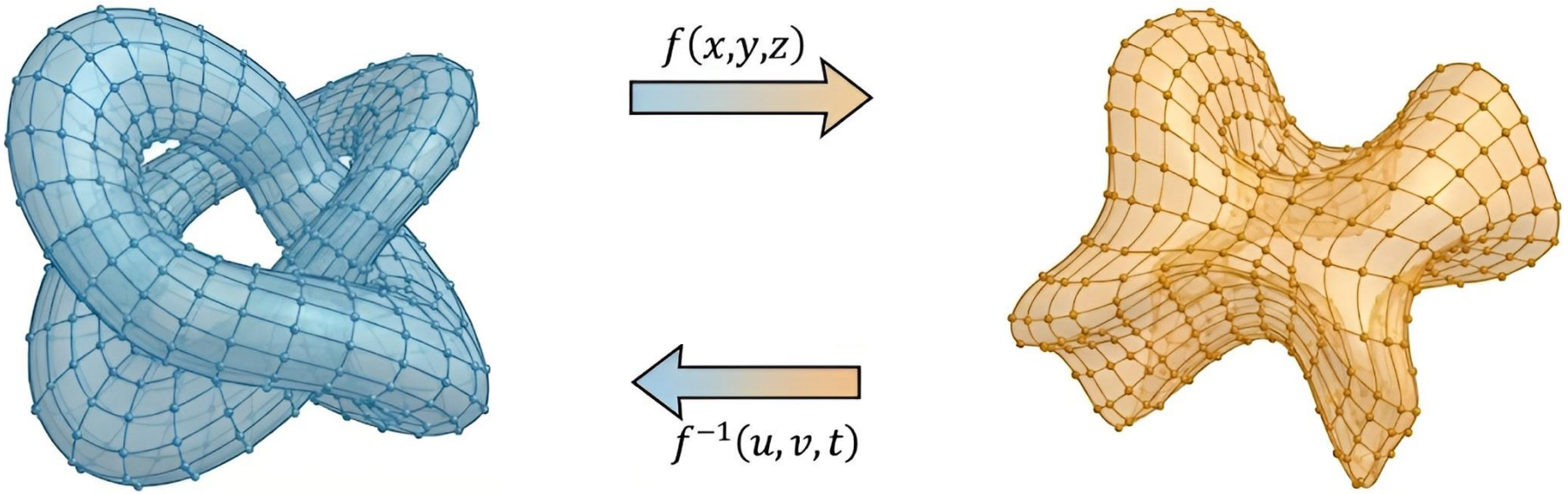


$$x \neq x' \implies f(x) \neq f(x')$$

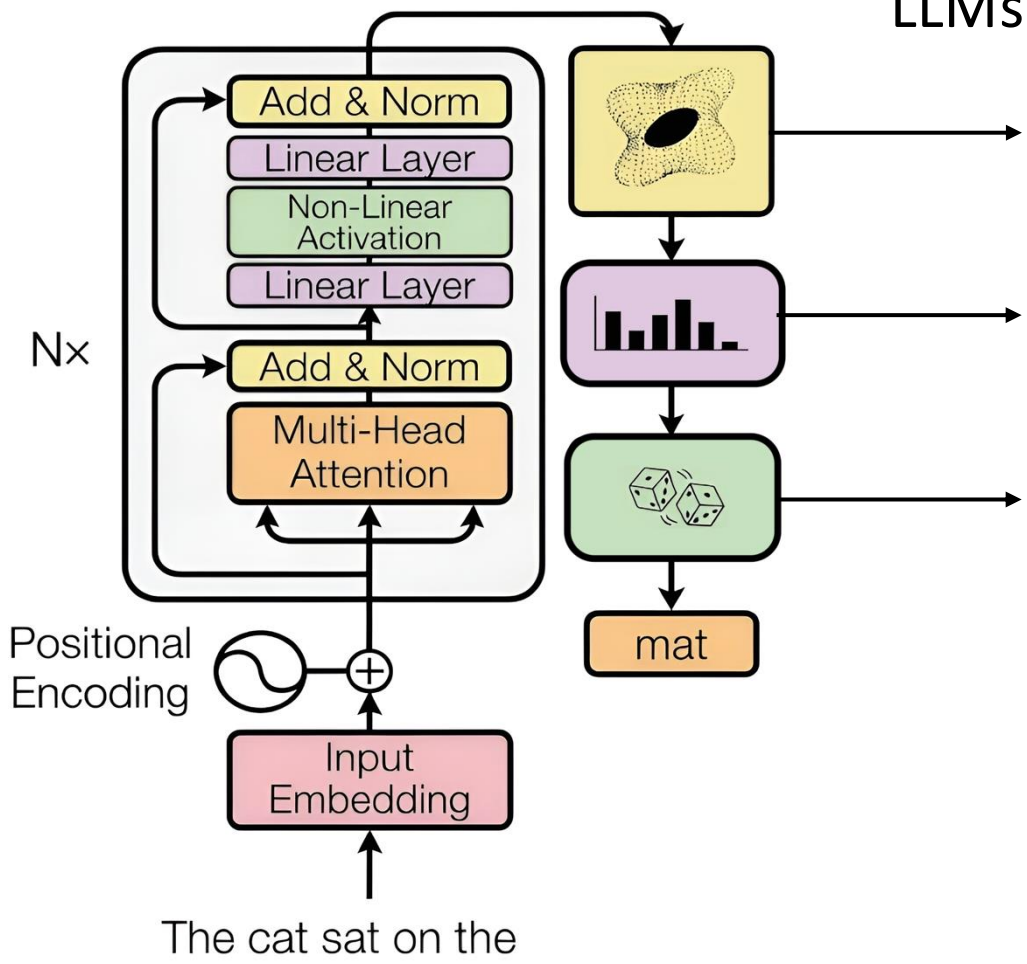
Invertible Functions

Invertibility: Given the output of the function, we can uniquely determine the input.

Injectivity implies that the function is invertible on its image.



What is the output of an LLM?



LLMs can be seen as function from a ***sequence*** to:

1. **Vectors** (Hidden Representations)
2. **Distributions** (Projection)
3. **Tokens** (Sampling)

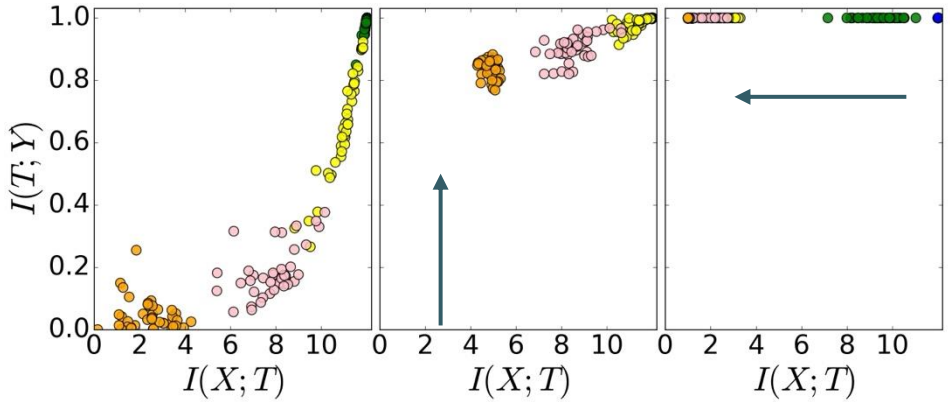
We analyse the mapping from sequence to hidden representations: $\mathrm{s} \mapsto \mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta})$

Invertibility is a form of blasphemy

TISHBY [1] **:** DNN's Learning is **prediction + compression**.

$$\max I(T;Y), \qquad \min I(T;X)$$

ERM phase (short at init): $I(T;Y)\uparrow$ Compression phase: $I(T;X)\downarrow$

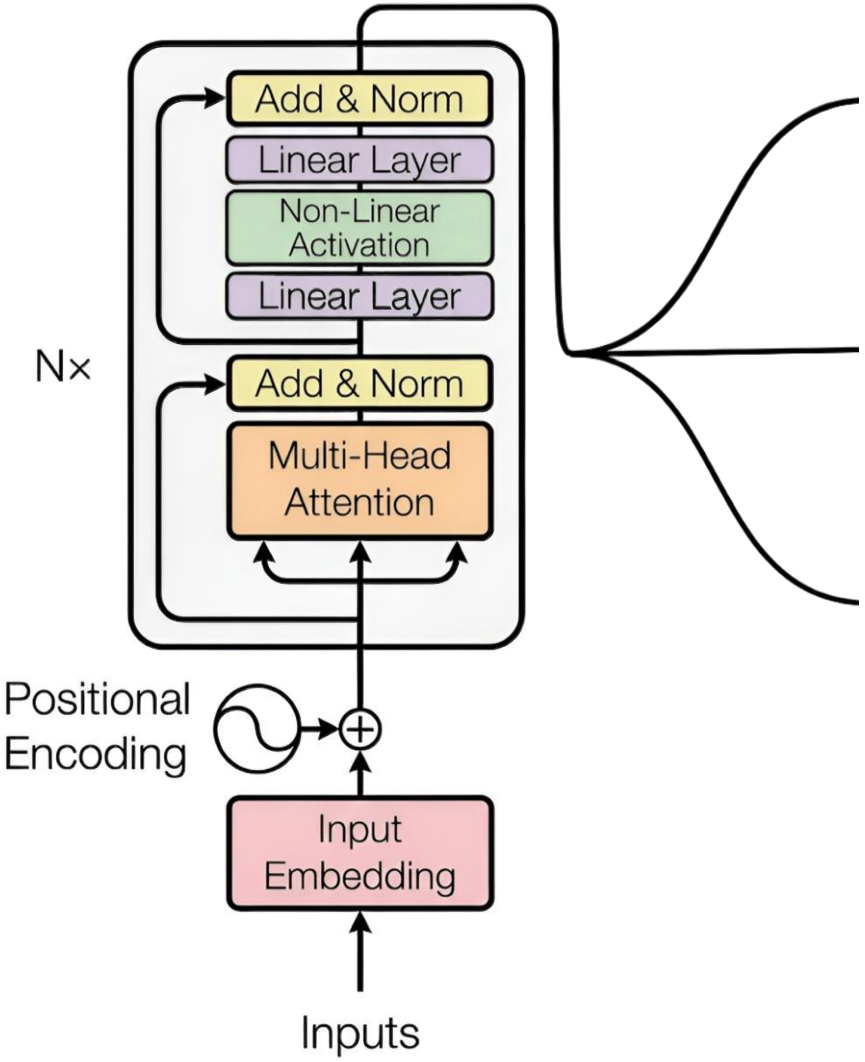


IF LLMs are injective, there is no compression in the activation:

$$r(\cdot;\theta) \text{ is invertible} \rightarrow I(X;Y) = I(r(X;\theta);Y)$$

[1] Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information.

Architectural Suspects for Collisions

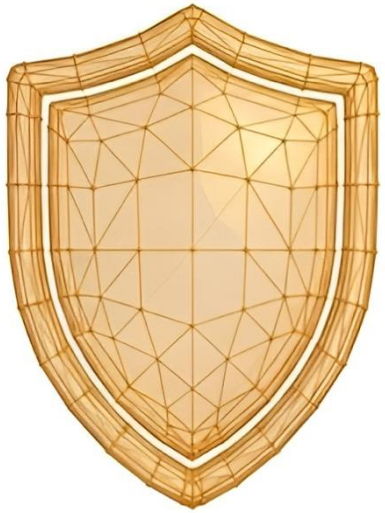


Multi-Head Attention: weighted averaging mechanism, inherently many-to-one.

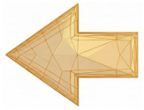
Normalization: (*LayerNorm*) collapses global shift and scale.

Non-Linear Activations: Often many-to-one.
e.g. *ReLU*(−5) = *ReLU*(−1) = 0

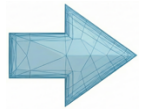
Implications

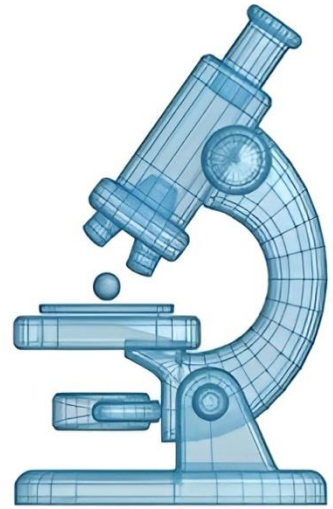


PRIVACY



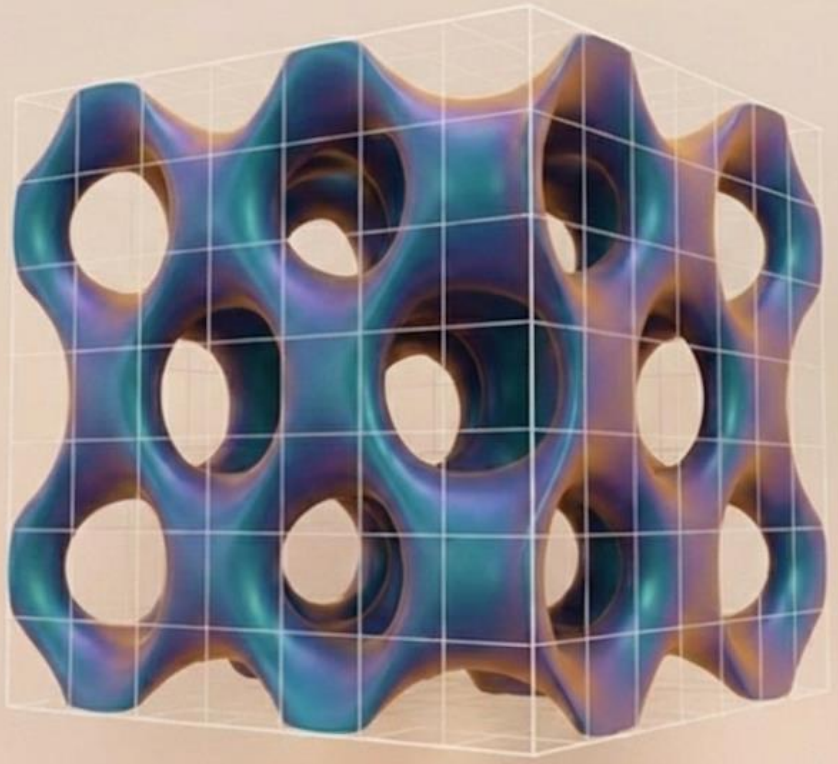
INJECTIVITY &
INVERTIBILITY





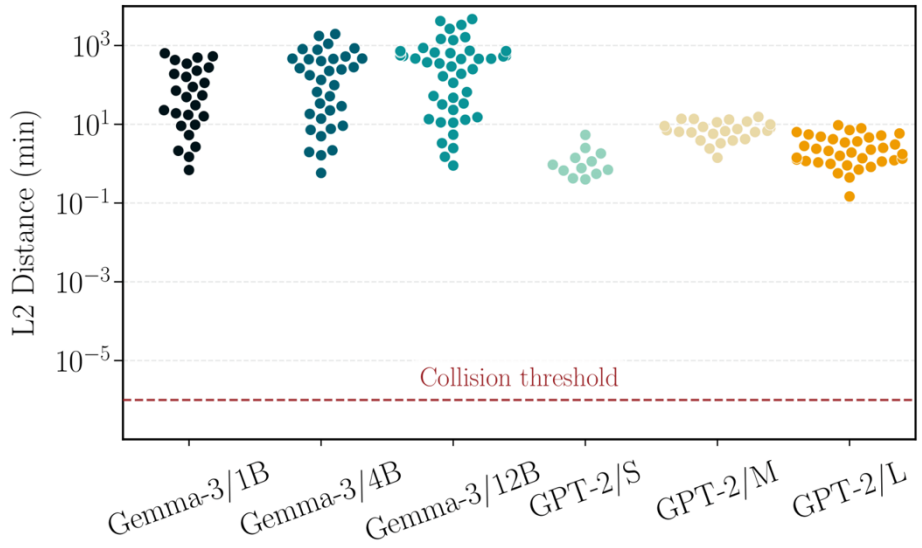
INTERPRETABILITY

Collisions in LLMs



Empirical Surprise: No Collisions

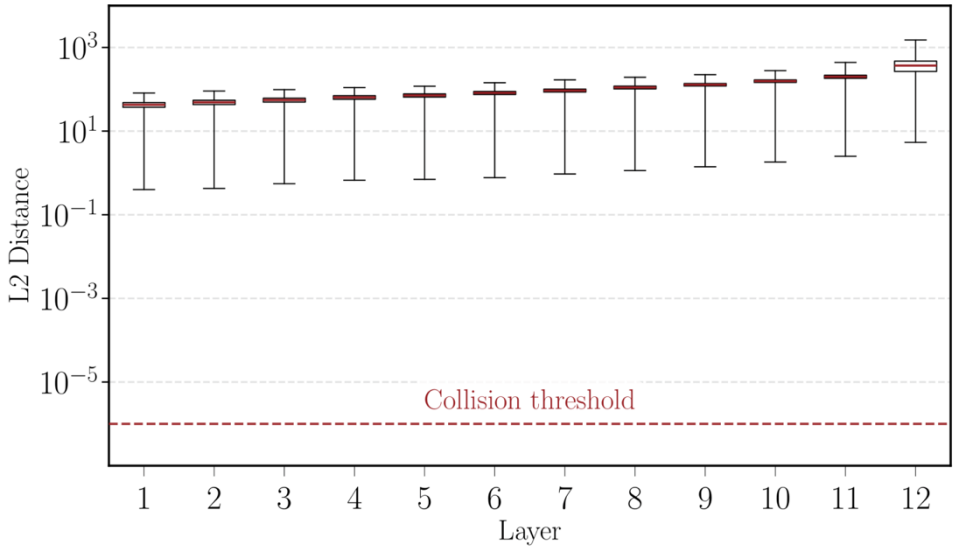
Across **billions to trillions** of tests, we observed **zero** collisions.



- The closest pairs still had surprisingly large distances, far from ambiguous.
- Results hold consistently across GPT, Gemma, Llama, and more model families!

Layerwise Minimum Distances

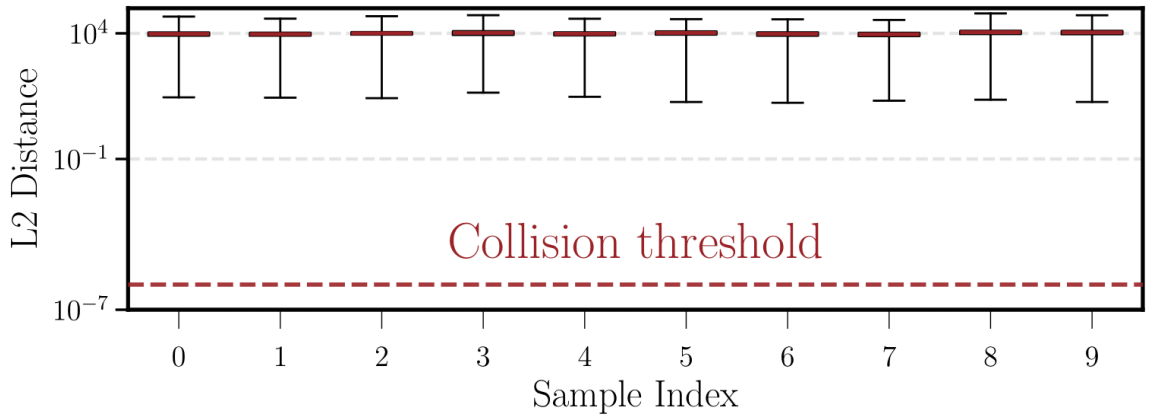
Separation **grows** with model depth; we observed a clear increasing trend.



Implication: Even tiny initial differences are amplified during the forward pass.

Stress-Testing Injectivity

Exhaustive search: Append every token to the same large prefix and compare outputs.



- Gemma alone: tested **400 billion pairs**.
- **Result:** No collisions, every prefix produced consistently large separations.

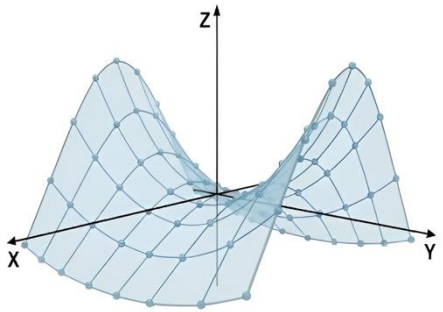
- Experiments strongly suggest injectivity, how could we prove it?
- **Prove that injectivity (1) holds at initialization and (2) is preserved by GD!**

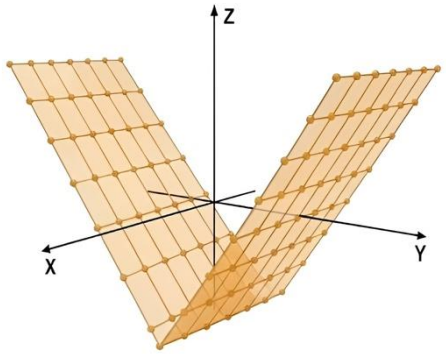
Real-Analytic Functions 101

Key Observation: modern decoder-only LLMs are **real-analytic functions!**

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k = \underbrace{a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \cdots}_{\text{“opened” form}}$$

Best behaved functions after polynomials; locally at every point they are equivalent to an ***infinite polynomial.***



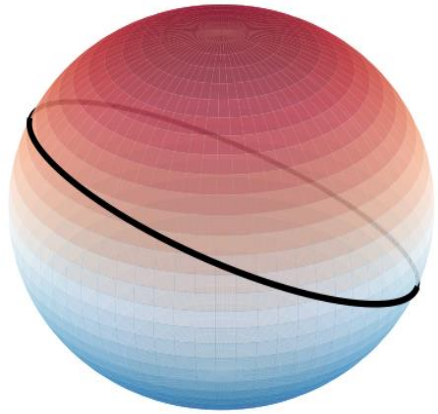


Why do we care about real-analyticity?

- The class of real-analytic function is **closed under composition**.
- The roots of a real-analytic function have **measure zero.**

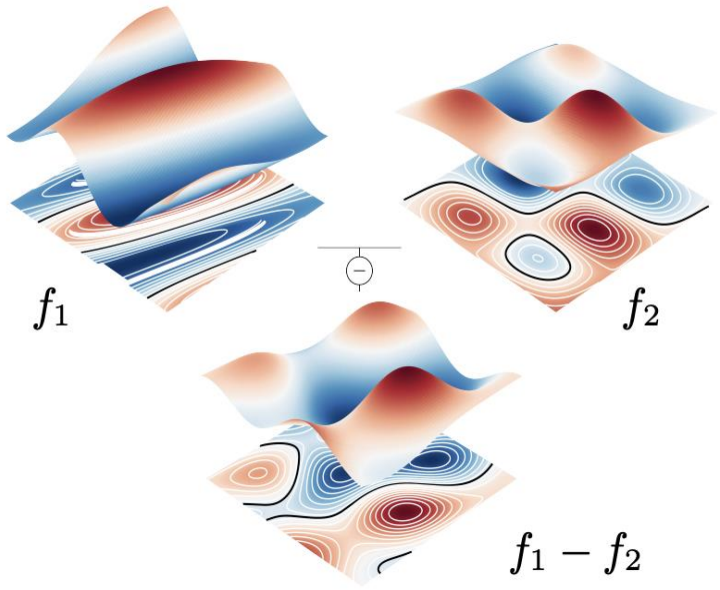
$$x \text{ is a root of } f \iff f(x) = 0$$

- Measure zero sets are **avoided** with probability one!



Measure zero translates to "mathematical exceptions".

Almost-Sure Injectivity at Initialization



We can create a real-analytic function whose
roots are non-injective parametrizations!

$$h(\boldsymbol{\theta}) = \|\mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta}) - \mathbf{r}(\mathrm{s}'\,;\,\boldsymbol{\theta})\|_2^2\,.$$

Theorem 2.2 (Almost-sure injectivity at initialization)**.** *Let $\boldsymbol{\theta}$ be drawn from any distribution with a density (e.g. Gaussian or uniform). Then for any two distinct prompts* $\mathrm{s}, \mathrm{s}' \in \mathcal{V}^{\leq K}$,

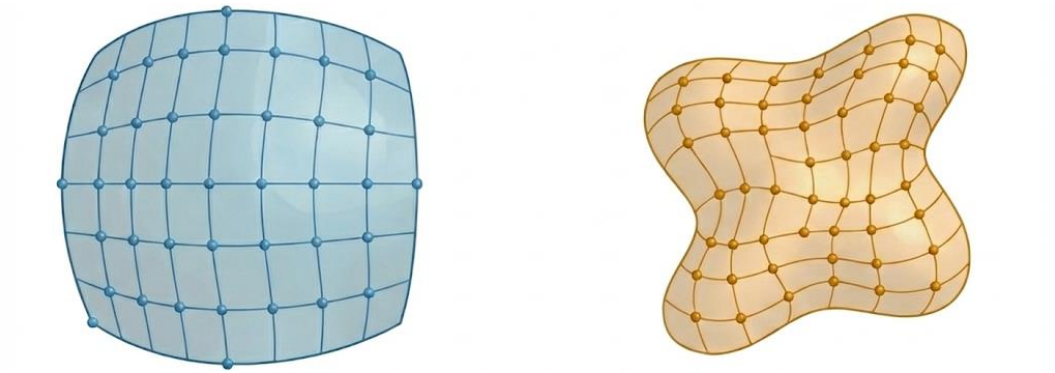
$$\Pr[\mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta}) = \mathbf{r}(\mathrm{s}'\,;\,\boldsymbol{\theta})] = 0\,.$$

Almost-Sure Injectivity under Gradient Descent

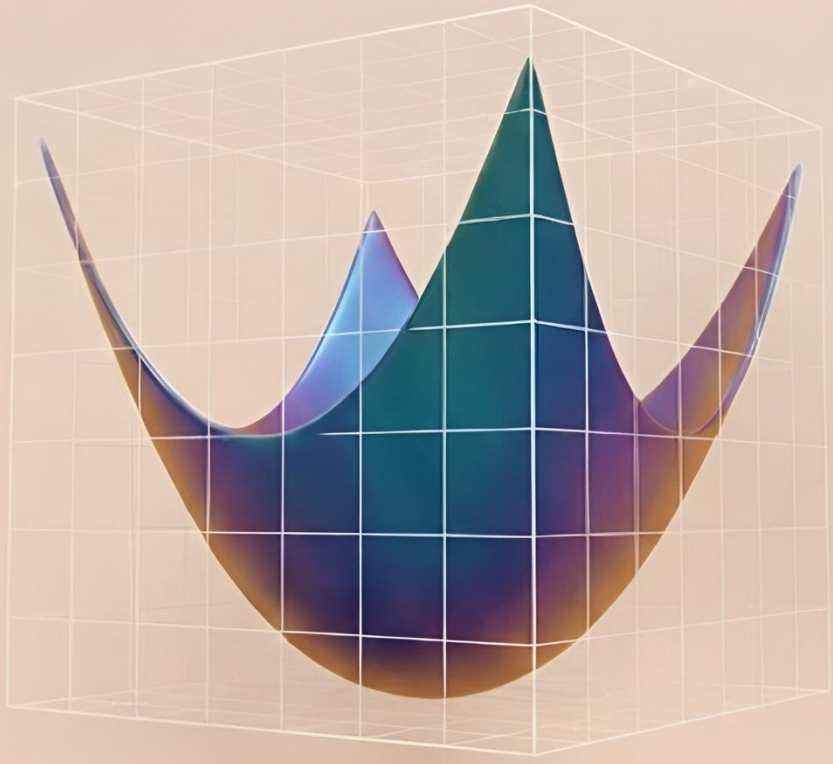
Key Theorem: In our setup, using the standard Cross-Entropy loss, Gradient Descent preserves absolute continuity of the parameters!

Theorem 2.3 (Injectivity preserved under training). *Let $\boldsymbol{\theta}_0$ be initialized from a distribution with a density, and let $\boldsymbol{\theta}_T$ be the parameters after T steps of gradient descent with step sizes in $(0, 1)$. Then with probability one,*

$$\mathrm{s} \neq \mathrm{s}' \quad \Longrightarrow \quad \mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta}_T) \neq \mathbf{r}(\mathrm{s}'\,;\,\boldsymbol{\theta}_T)\,,$$

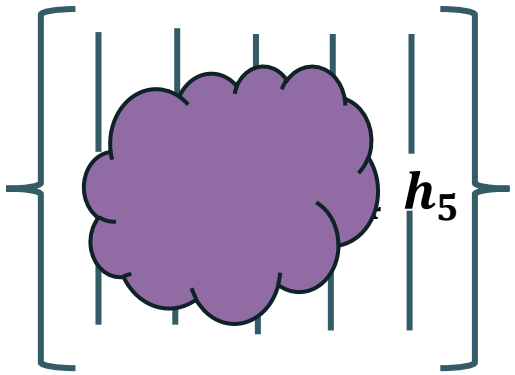


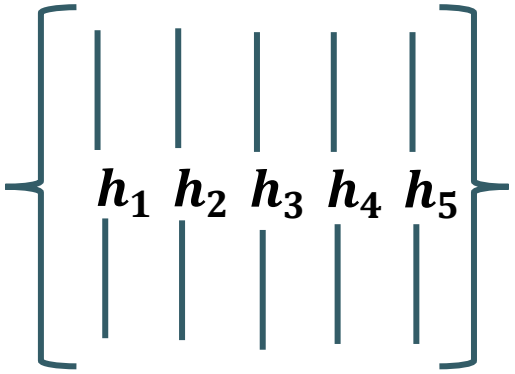
SIPIT: Inverting LLMs



Setting: Full Hidden State Matrix

Granularity: we assume full access to the hidden matrix H at layer ℓ.





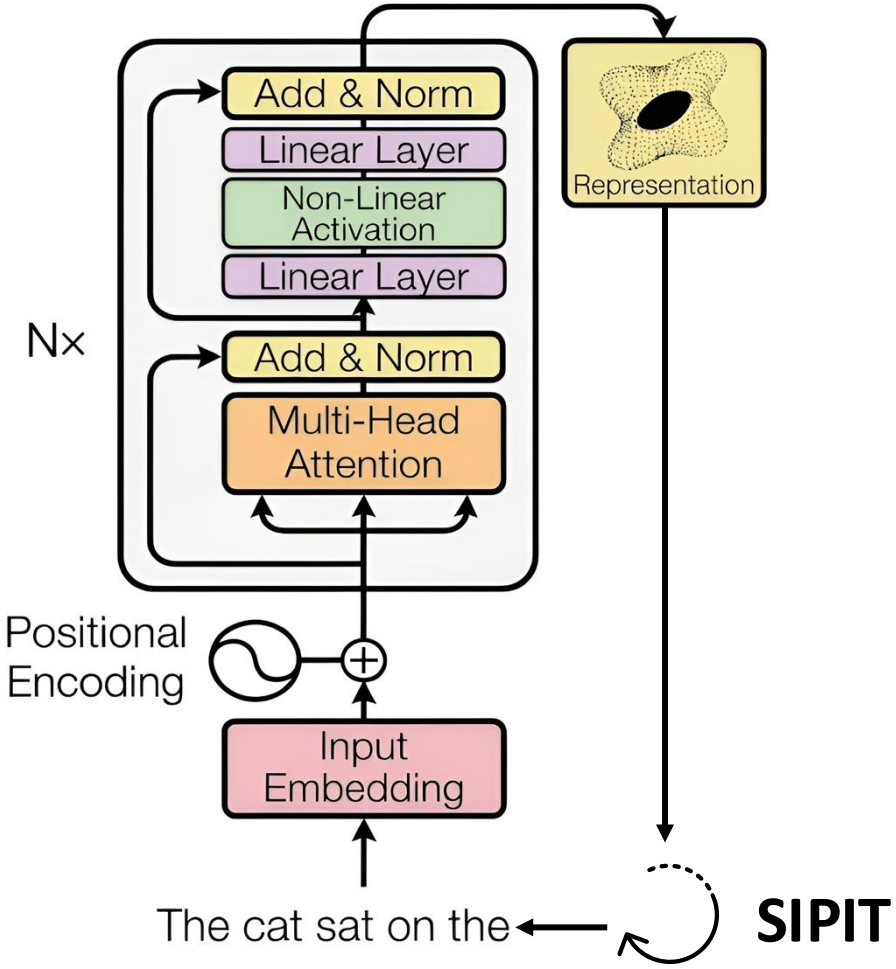
Causal Decoder: we assume that the model uses causal attention.

The cat is on	the	mat

$\mathbf{h}_{\text{on}}$ **does not** depend on $\mathbf{h}_{\text{the}}$

$\mathbf{h}_{\text{mat}}$ **does** depend on $\mathbf{h}_{\text{the}}$

If nothing is lost, how can we invert representations?



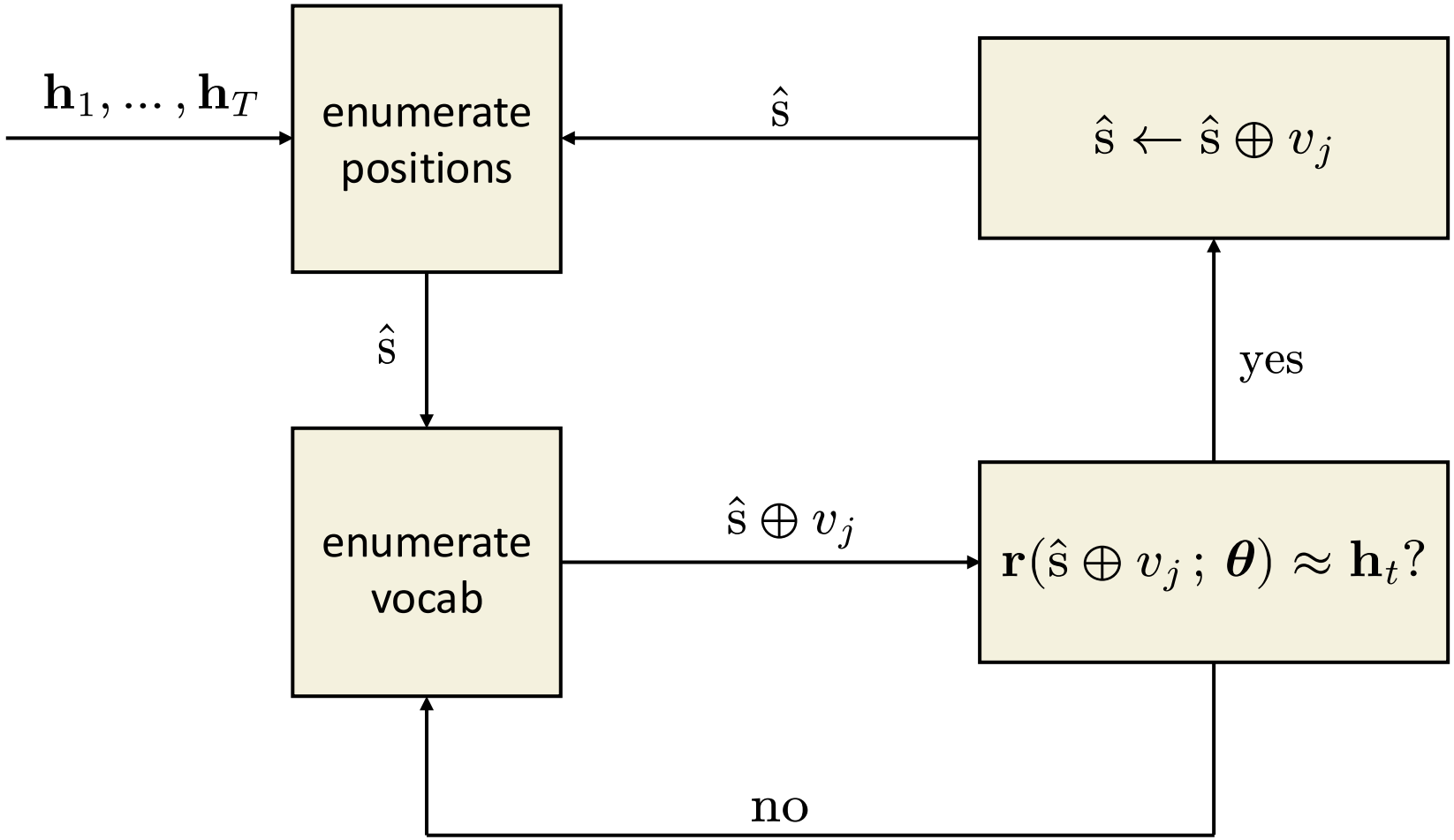
Injectivity: Implies that each representation corresponds to a **unique** prompt!

RQ: Given hidden states, can we **efficiently** recover the exact input text that produced them?

Answer: SIPIT does exactly that!

How does SIPIT work?

Key Insight: Causality enables Sequentiality; SIPIT exploits that!



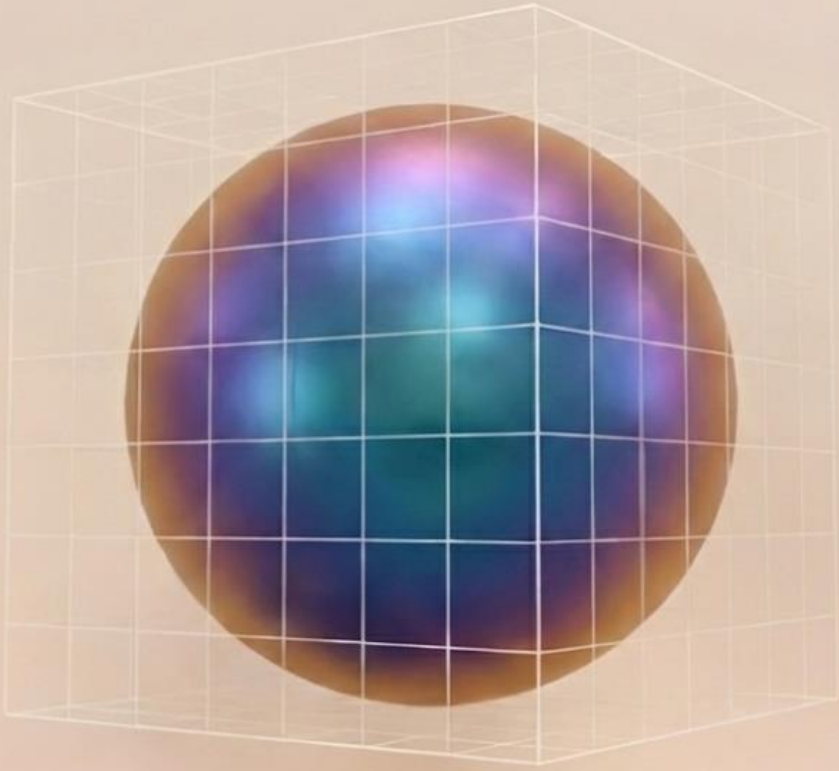
Accuracy and Efficiency of SIPIT

- **Theorem:** SIPIT recovers the true input sequence with probability one in at most $T|\mathcal{V}|$ steps!
- As claimed, SIPIT achieves **perfect reconstruction**, even on random sequences!
- Compared to baselines, the gradient heuristic makes SIPIT efficient!

Dataset	Inversion Time (s)	Accuracy
Train Data	146.48 ± 91.52	100%
Test Data	128.62 ± 83.40	100%
OOD	106.87 ± 39.10	100%

Method	Mean Time (s)	Accuracy
HARDPROMPTS	6132.59 ± 104.61	0.00
BRUTEFORCE (ours)	3889.61 ± 691.17	1.00
SIPIT (ours)	$\mathbf{28.01 \pm 35.87}$	**1.00**

The Math of Injectivity

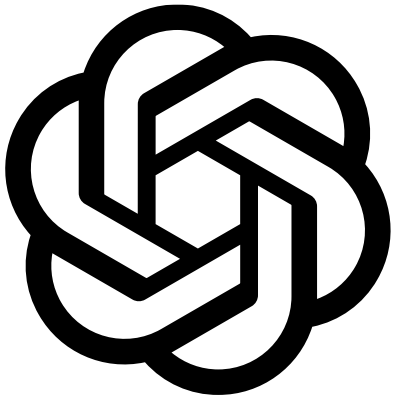


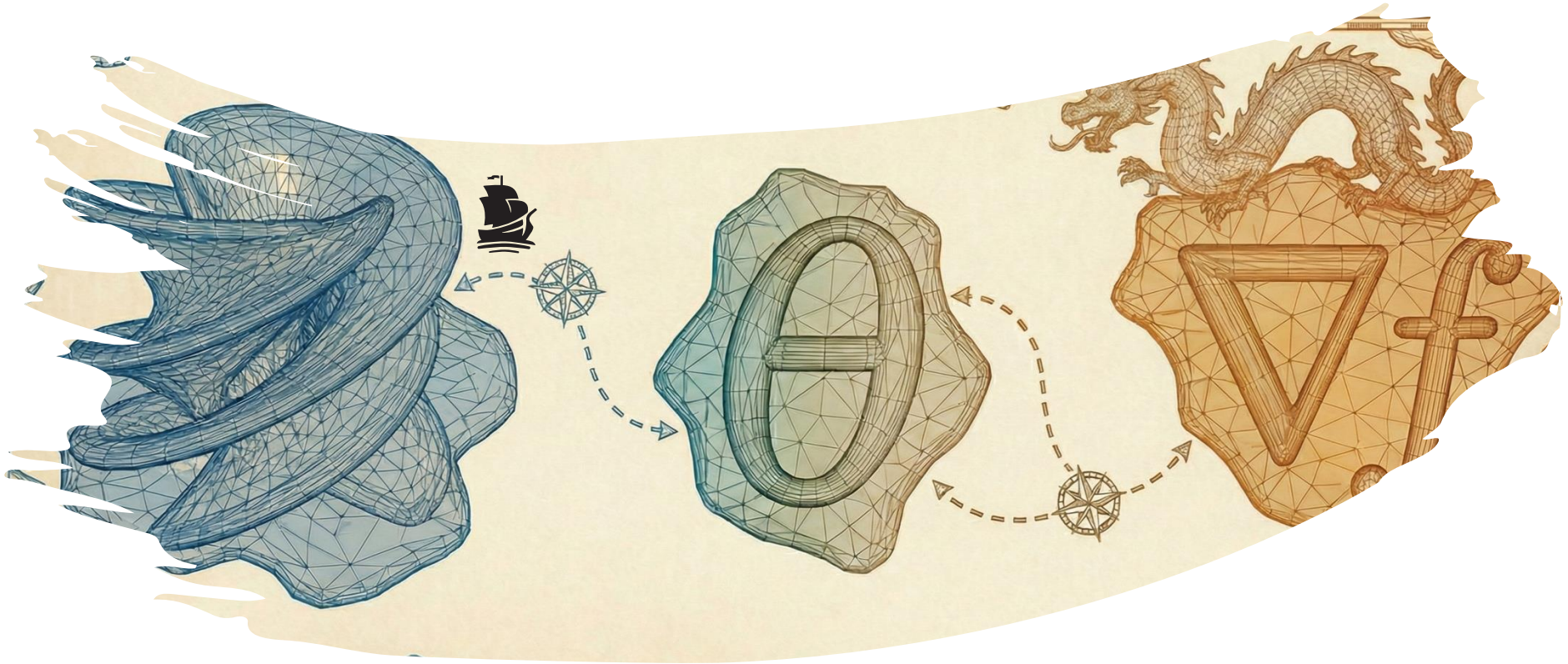
Key Components

1. Modern decoder-only Transformer Language Models are (mostly) real-analytic.

2. Parameters drawn from an absolutely continuous distribution almost surely create an injective model.

3. Gradient Descent (and variants) preserve absolute continuity of the parameters.





Part 1: Real-Analyticity

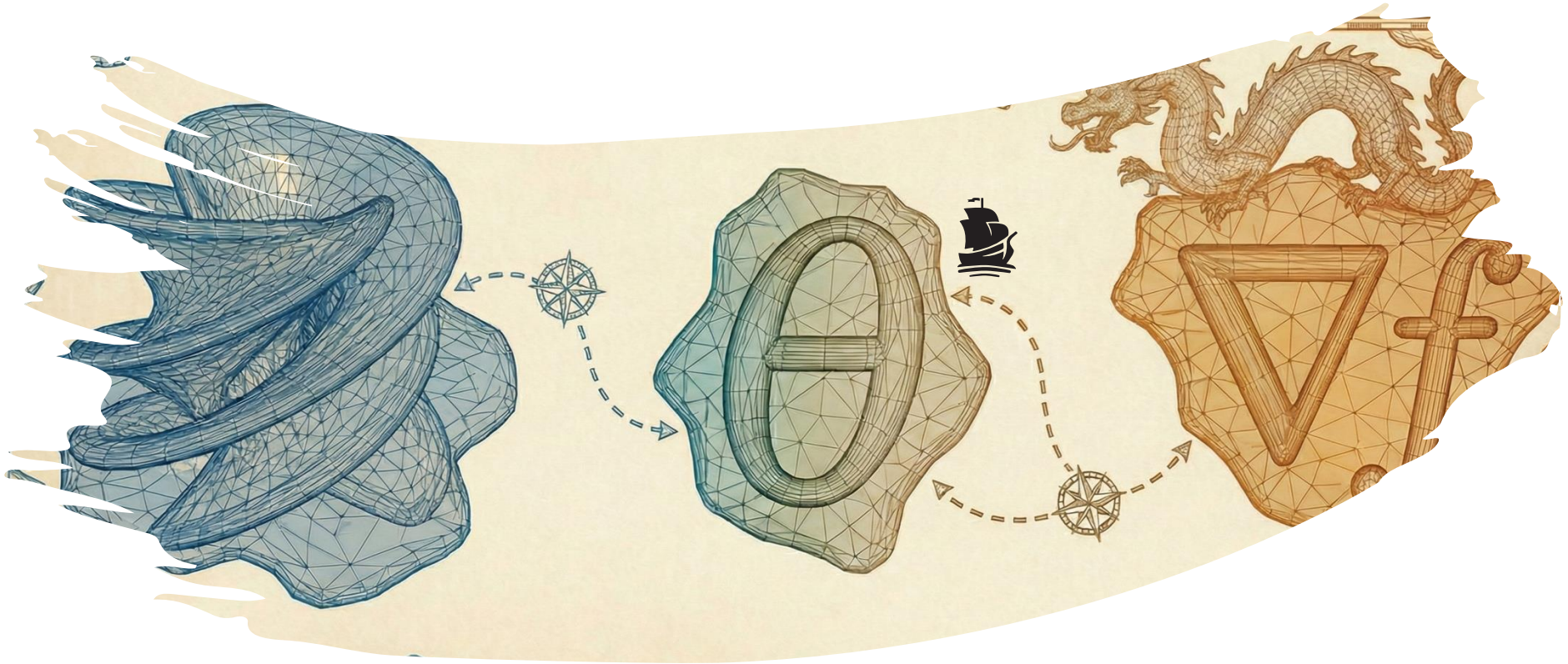
Real-Analyticity of Transformer Language Models

All sub-component are real-analytic: MLP (with real-analytic non-linearities), Multi-Head Self-Attention (causal or not), Normalization, Projection (…)

$$\boldsymbol{\eta}(\mathbf{X}) = \mathrm{softmax}\left(\frac{(\mathbf{XQ})\,(\mathbf{XK})^\top}{\sqrt{d_\eta}}\right)\mathbf{XV}$$

$$\mathrm{LN}(\mathbf{x}) = \boldsymbol{\gamma} \odot \frac{\mathbf{x} - \mu_{\mathbf{x}}\mathbf{1}_d}{\sqrt{\sigma_{\mathbf{x}}^2 + \varepsilon}} + \boldsymbol{\beta}$$

Key Property: Real-Analytic functions are closed under composition $\Longrightarrow$ LLMs are real-analytic!



Part 2: Injectivity at Initialization

Absolute Continuity implies Injectivity

Key Theorem: The zero-set of a <u>non-trivial</u> real-analytic map has measure zero.

For a given pair of sequences $\mathrm{s} \neq \mathrm{t} \in \mathcal{V}^{\leq K}$ consider the mapping:

$$h(\boldsymbol{\theta}) = \|\mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta}) - \mathbf{r}(\mathrm{t}\,;\,\boldsymbol{\theta})\|_2^2$$

This mapping is real-analytic by composition!

What is missing: We need to show that it is not identically zero!

Cont'd (1/2)

Witness Construction: It suffices to exhibit a single parameter setting $\boldsymbol{\theta}_\star$ where

$$\mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta}_\star) \neq \mathbf{r}(\mathrm{t}\,;\,\boldsymbol{\theta}_\star)$$

Two Cases:

1. **The sequences differ in the last position or in length:** freeze the network so that the last state reduces to embedding plus position, and choose distinct rows.

2. **The sequences differ earlier:** Using attention, transfer the difference to the last token.

Cont'd (2/2)

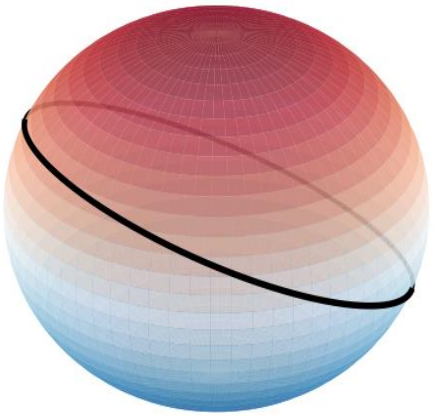
Key Fact: Absolutely continuous distribution $\Longrightarrow$ measure-zero sets have probability zero.

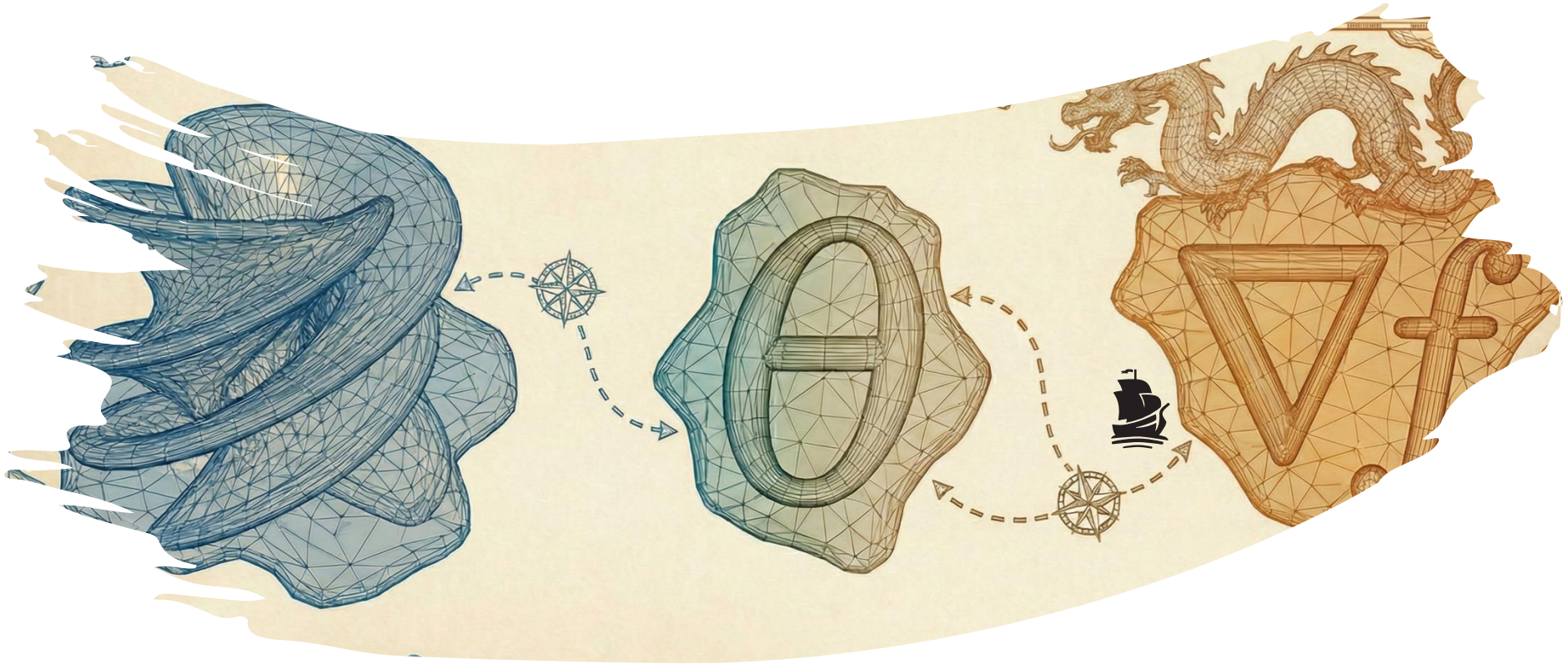
$$\lambda(A) = 0 \implies p_X(A) = 0$$

Theorem C.2 (Almost-sure pairwise distinctness of last-token representations)**.** *Let the parameter vector* $\boldsymbol{\theta} \in \mathbb{R}^p$ *be drawn from any distribution absolutely continuous with respect to Lebesgue measure. Then, for any fixed* $\mathrm{s} \neq \mathrm{t}$,

$$\Pr\left[\mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta}) = \mathbf{r}(\mathrm{t}\,;\,\boldsymbol{\theta})\right] = 0.$$

Finally, use a **union bound** over all possible distinct pairs!



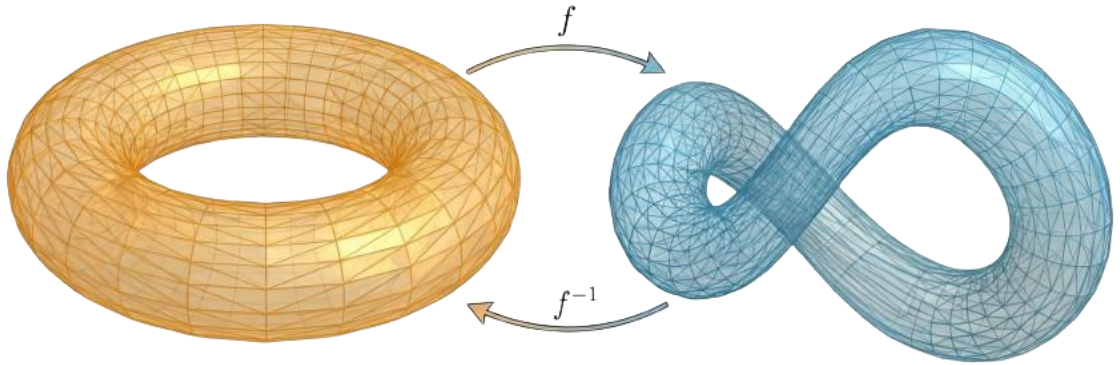


Part 3: Preservation of Injectivity

Prerequisites (1/2)

Diffeomorphism: A map $f : \mathcal{U} \to \mathcal{V}$ between open sets is a C^k diffeomorphism if:

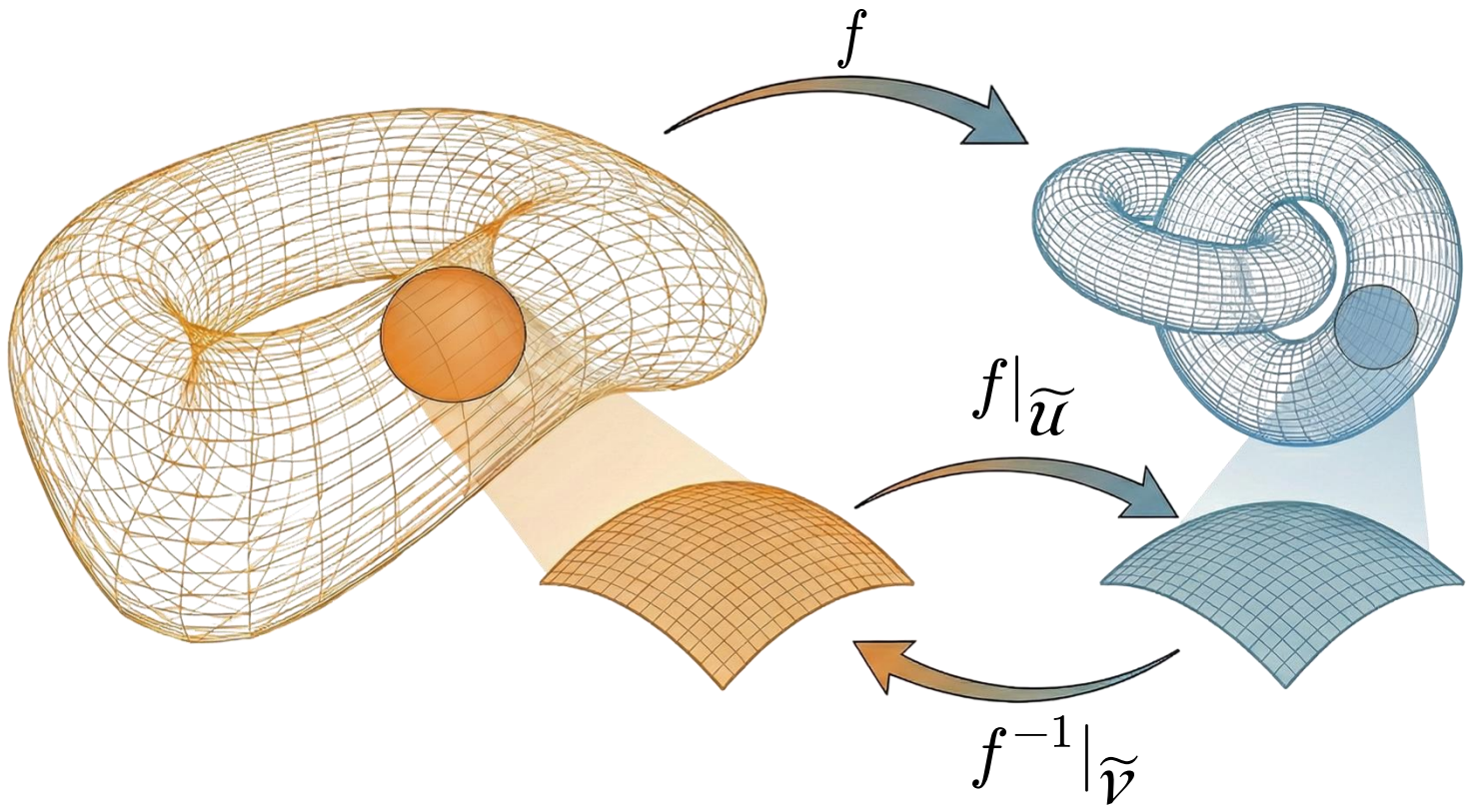
1. f is bijective
2. f is C^k (All partial derivatives of f up to order k exist and are continuous)
3. The inverse map f^{-1} is C^k



Intuition: *stretch, bend, or smoothly reshape* but do not *tear or glue parts together.*

Prerequisites (2/2)

Inverse Function Theorem: Given a map $f : \mathcal{U} \to \mathcal{V}$ between open sets, if the determinant of the Jacobian $\det Df(\boldsymbol{\theta}) \neq 0$, then f is *locally* a C^k diffeomorphism.



Key Intuition

Cool, what can we say about the GD mapping: $\phi(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta_t \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{\theta})$

Hope: Given an absolutely continuous random variable, applying the gradient mapping to it leads to a complex but absolutely continuous random variable.

Condition: The pre-image of every measure zero set must also be measure zero!

$$\lambda(A) = 0 \implies \lambda\big(T^{-1}(A)\big) = 0$$

Tool: Express the gradient descent mapping as a *local bijection almost everywhere*, and the result follows from standard real-analysis.

Cont'd (1/3)

Theorem: The critical set of the GD map has measure zero!

$$\mathcal{C} \;:=\; \{\boldsymbol{\theta} \in \mathbb{R}^p : \det D\phi(\boldsymbol{\theta}) = 0\}, \quad \mathrm{Leb}(\mathcal{C}) = 0$$

The proof although technical is straight forward:

1. The GD map is real-analytic, and so is the determinant of the jacobian.
2. Due to the nature of the architecture and the Cross-Entropy loss used, we can provide an explicit witness to show it is a non-trivial map.
3. Using the *by-now well-known* theorem, the result follows.

Cont'd (2/3)

Why does this matter: On the open set $\mathbb{R}^p \setminus \mathcal{C}$, the Inverse Function Theorem applies!

Theorem: The preimage of the GD map for a measure zero set is measure zero!

A very hand-wavey proof:

1. We pass from the uncountable cover provided by IFT to a countable subcover.
2. The preimage of the GD map for a measure-zero set is a subset of the countable images of the *inverse diffeomorphisms of IFT* plus the critical set.
3. Therefore, it is also measure-zero!

Cont'd (3/3)

Key Definition: Given a random variable $\boldsymbol{\theta} \sim \mu$, the pushforward $\phi_{\#}\mu$ refers to the distribution of the random variable $\phi(\boldsymbol{\theta})$, i.e. $\phi(\boldsymbol{\theta}) \sim \phi_{\#}\mu$.

Key Fact: The pushforward is absolutely continuous if applying ϕ doesn't create new probability on sets that were "negligible" (Lebesgue measure zero).

The theorem of the previous slide says that the pushforward of an absolutely continuous distribution throught the GD map, is absolutely continuous.

Combining: Gradient Descent preserves absolute continuity and therfore injectivity!

Theorem 2.3 (Injectivity preserved under training). *Let $\boldsymbol{\theta}_0$ be initialized from a distribution with a density, and let $\boldsymbol{\theta}_T$ be the parameters after T steps of gradient descent with step sizes in $(0, 1)$. Then with probability one,*

$$\mathrm{s} \neq \mathrm{s}' \quad \Longrightarrow \quad \mathbf{r}(\mathrm{s}\,;\,\boldsymbol{\theta}_T) \neq \mathbf{r}(\mathrm{s}'\,;\,\boldsymbol{\theta}_T)\,,$$

THANK YOU!

Donato Crisostomi

Andrea Santilli

Yannis Panagakis

Emanuele Rodolà

tommaso.mencattini@epfl.ch