Simplicial Sets: Designing the Geometric Transformer

Sridhar Mahadevan, Adobe Research and U.Mass, Amherst

Recap: Lectures 1-6

- Categories: collection of objects and arrows
- Functors: structured mapping between categories
- Natural transformations: structured mappings between functors
- Yoneda Lemma: Representable set-valued functors
- Limits and colimits: Turning diagrams into computational objectives
- Functor categories: Each object is a functor

Diagrammatic Backpropagation: Deep learning with (co)limits

- The textbook Categories for AGI has several chapters devoted to this week's topics
  - Chapters 5-10 (~100 pages!)
  - It is the longest section in the book!
- We will study these concepts further in the course
  - This week's lectures are a high-level summary

Applications: Diagrammatic Backpropagation with Geometric Transformers

- Language modeling:
  - Wiki-103: 100-million tokens dataset from Wikipedia articles
- Causal inference:
  - Democritus: construct causal models from documents
- RL:
  - Learn a policy from random trial-and-error exploration

GAIA: Generative AI Architecture

GAIA: Categorical Foundations of Generative AI*

A Preprint

Sridhar Mahadevan
Adobe Research and University of Massachusetts, Amherst
smahadev@adobe.com, mahadeva@umass.edu

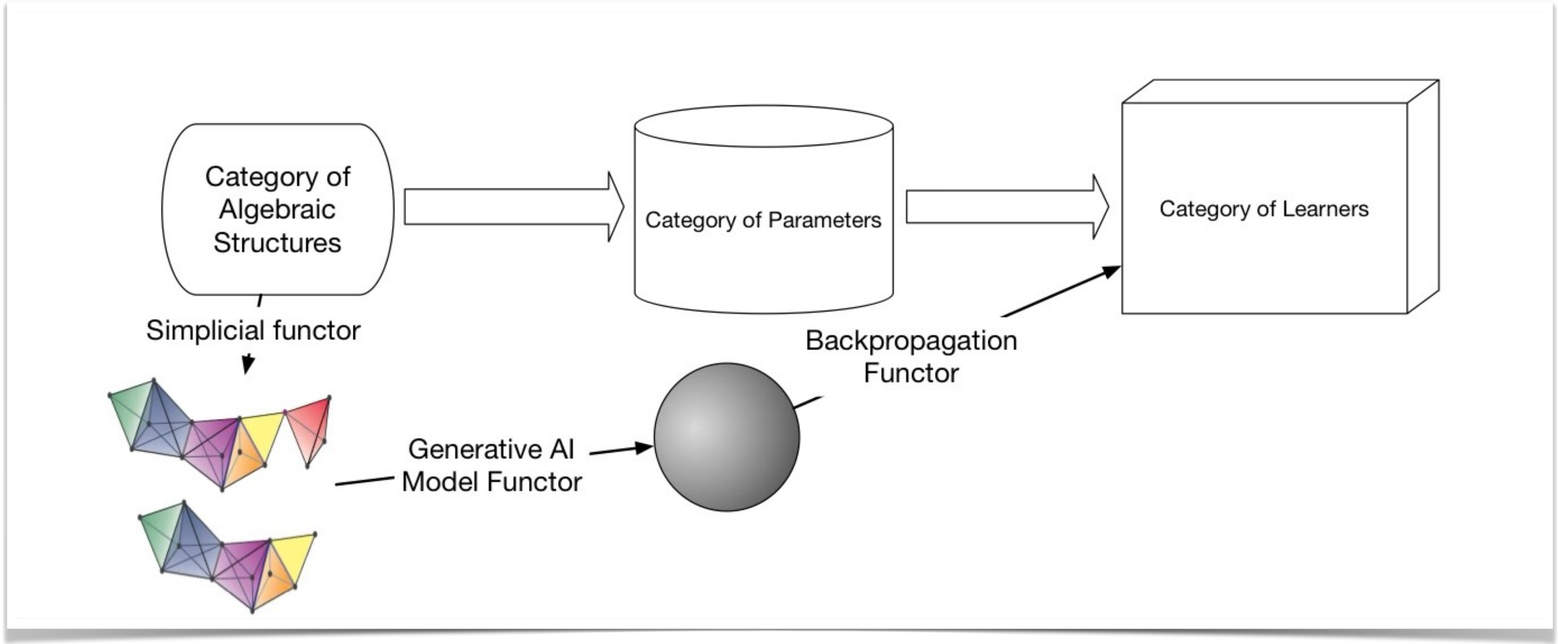
March 1, 2024

Abstract

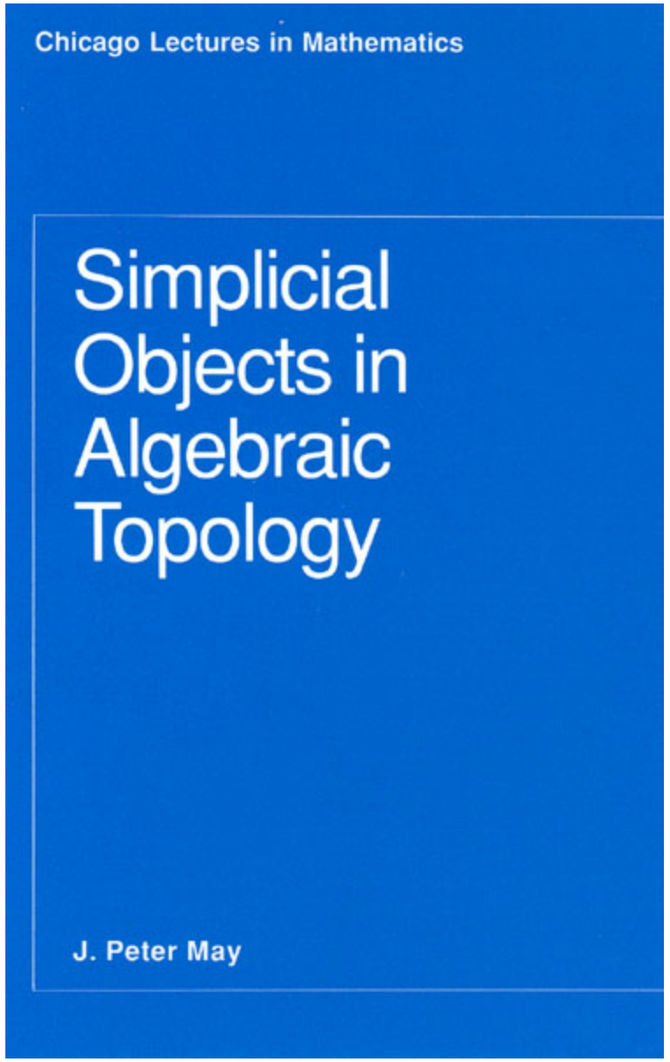
In this paper, we explore the categorical foundations of generative AI. Specifically, we investigate a Generative AI Architecture (GAIA) that lies beyond backpropagation, the longstanding algorithmic workhorse of deep learning. Backpropagation is at its core a compositional framework for (un)supervised learning: it can be conceptualized as a sequence of modules, where each module updates its parameters based on information it receives from downstream modules, and in turn, transmits information back to upstream modules to guide their updates. GAIA is based on a fundamentally different *hierarchical model*. Modules in GAIA are organized into a simplicial complex. Each n-simplicial complex acts like a manager of a business unit: it receives updates from its superiors and transmits information back to its $n+1$ subsimplicial complexes that are its subordinates. To ensure this simplicial generative AI organization behaves coherently, GAIA builds on the mathematics of the higher-order category theory of simplicial sets and objects. Computations in GAIA, from query answering to foundation model building, are posed in terms of lifting diagrams over simplicial objects. The problem of machine learning in GAIA is modeled as "horn" extensions of simplicial sets: each sub-simplicial complex tries to update its parameters in such a way that a lifting diagram is solved. Traditional approaches used in generative AI using backpropagation can be used to solve "inner" horn extension problems, but addressing "outer horn" extensions requires a more elaborate framework.

At the top level, GAIA uses the simplicial category of ordinal numbers with objects defined as $[n], n \geqslant 0$ and arrows defined as weakly order-preserving mappings $f : [n] \to [m]$, where $f(i) \leqslant f(j), i \leqslant j$. This top-level structure can be viewed as a combinatorial "factory" for constructing, manipulating, and destructing complex objects that can be built out of modular components defined over categories. The second layer of GAIA defines the building blocks of generative AI models as universal coalgebras over categories that can be defined using current generative AI approaches, including Transformers that define a category of permutation-equivariant functions on vector spaces, structured state-space models that define a category over linear dynamical systems, or image diffusion models that define a probabilistic coalgebra over ordinary differential equations. The third layer in GAIA is a category of elements over a (relational) database that defines the data over which foundation models are built. GAIA formulates the machine learning problem of building foundation models as extending functors over categories, rather than interpolating functions on sets or spaces, which yields canonical solutions called left and right Kan extensions. GAIA uses the metric Yoneda Lemma to construct universal representers of objects in non-symmetric generalized metric spaces. GAIA uses a categorical integral calculus of (co)ends to define two families of generative AI systems. GAIA models based on coends correspond to topological generative AI systems, whereas GAIA systems based on ends correspond to probabilistic generative AI systems.

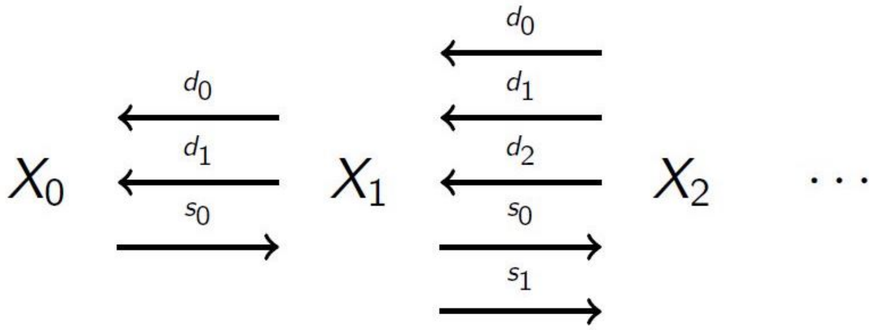
Keywords Generative AI · Foundation Models · Higher-Order Category Theory · Machine Learning

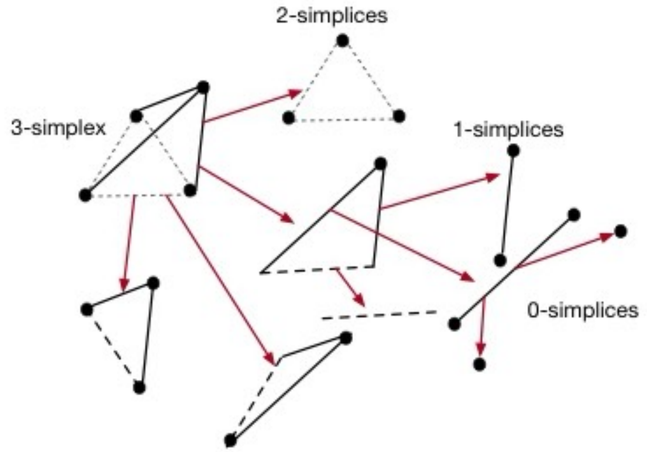


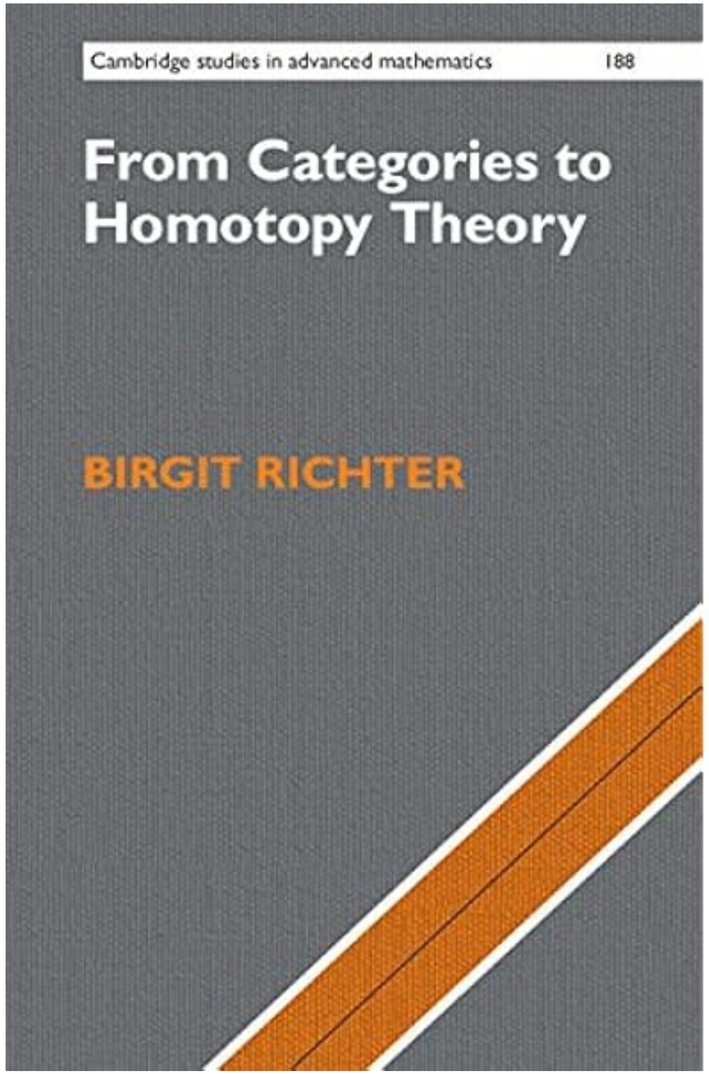
Simplicial Sets and Objects



Combinatorial model of spaces

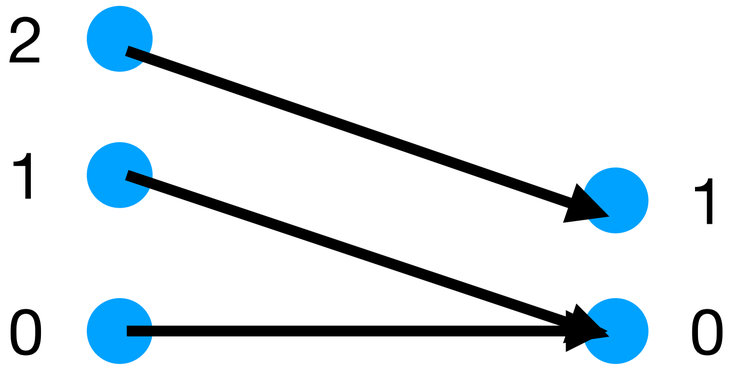






Simplicial Category Δ

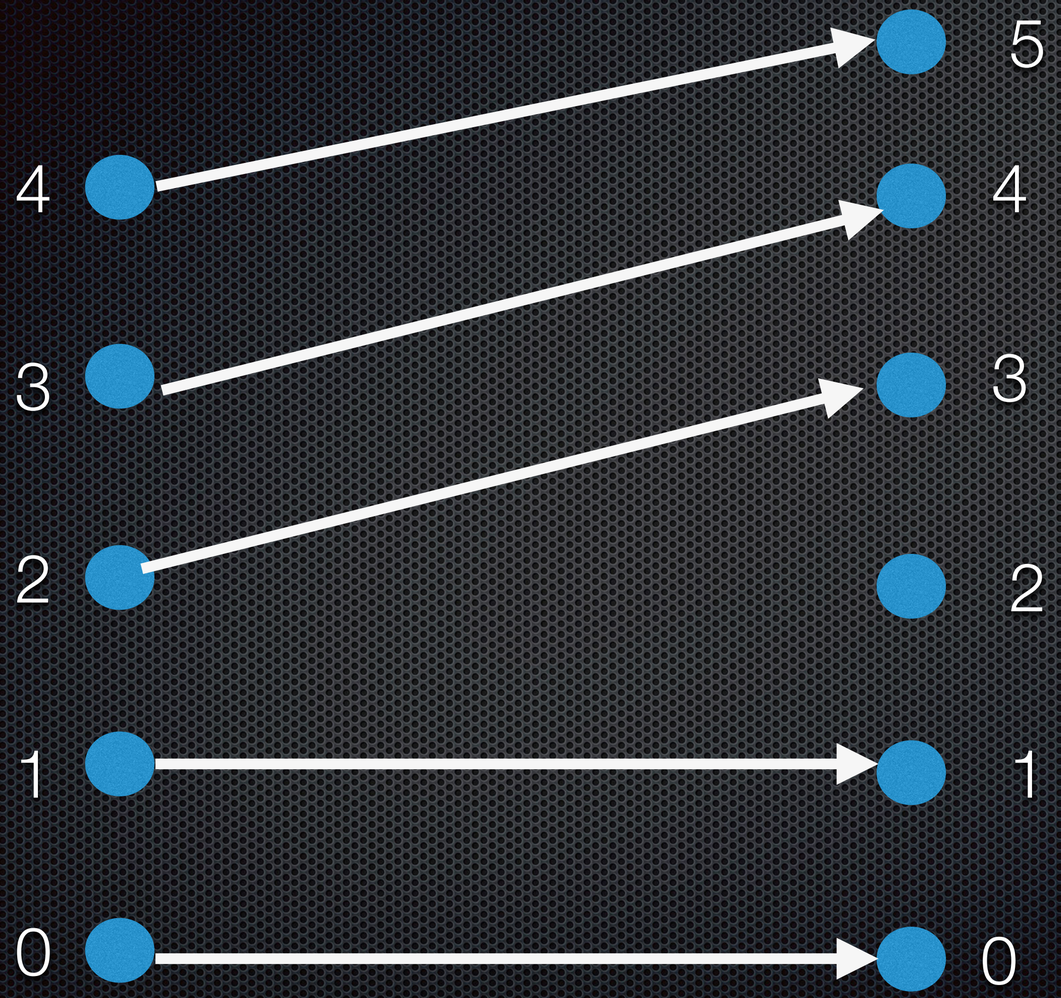
- **Objects**: ordinal numbers
  - $[n] = \{0,1,\ldots,n-1\}$
- **Arrows**:
  - $f : [m] \to [n]$
  - If $i \leq j$, then $f(i) \leq f(j)$



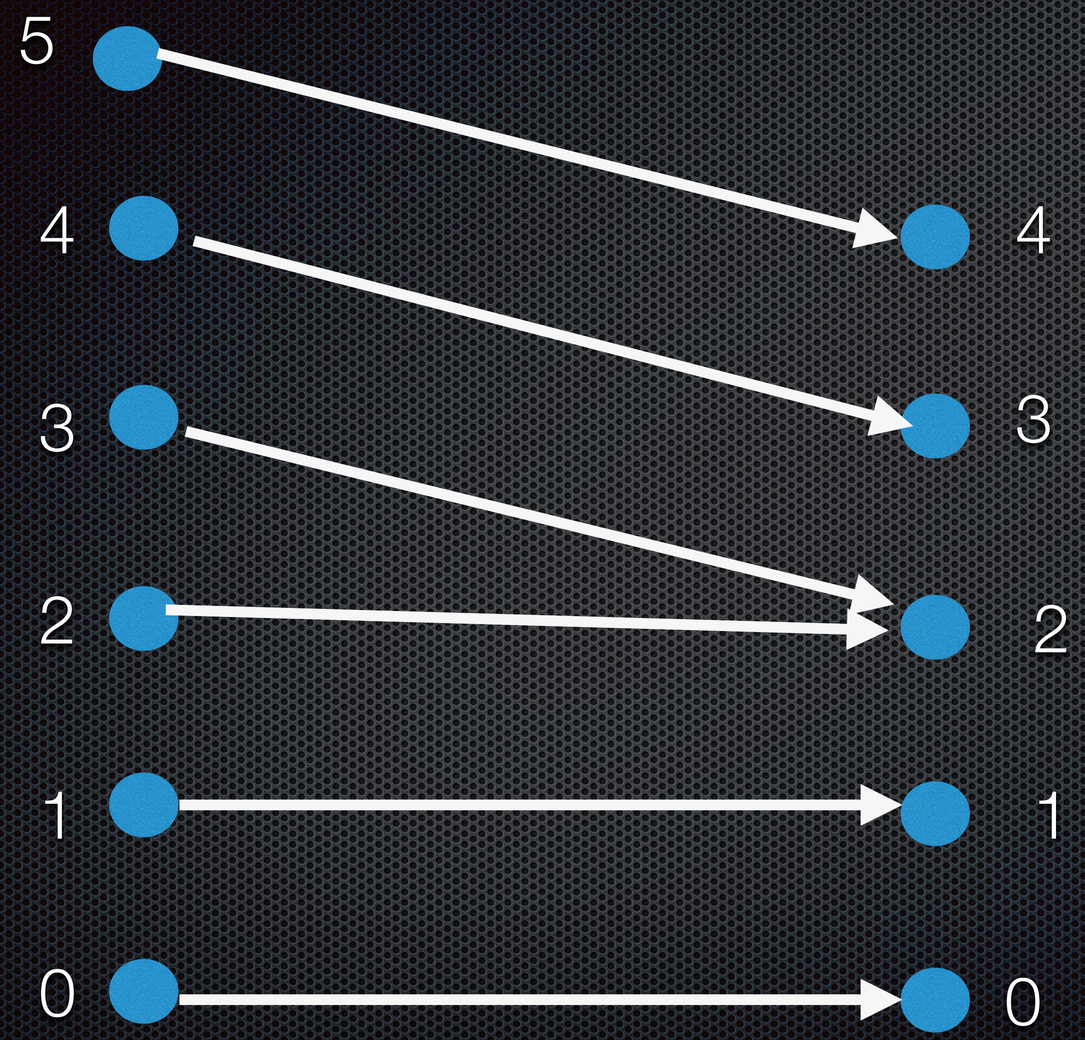
  - All morphisms can be built out of primitive injections/surjections
    - $\delta_i : [n] \to [n+1]$: injection skipping i
    - $\sigma_i : [n] \to [n-1]$, surjection repeating i

Primitive Injections

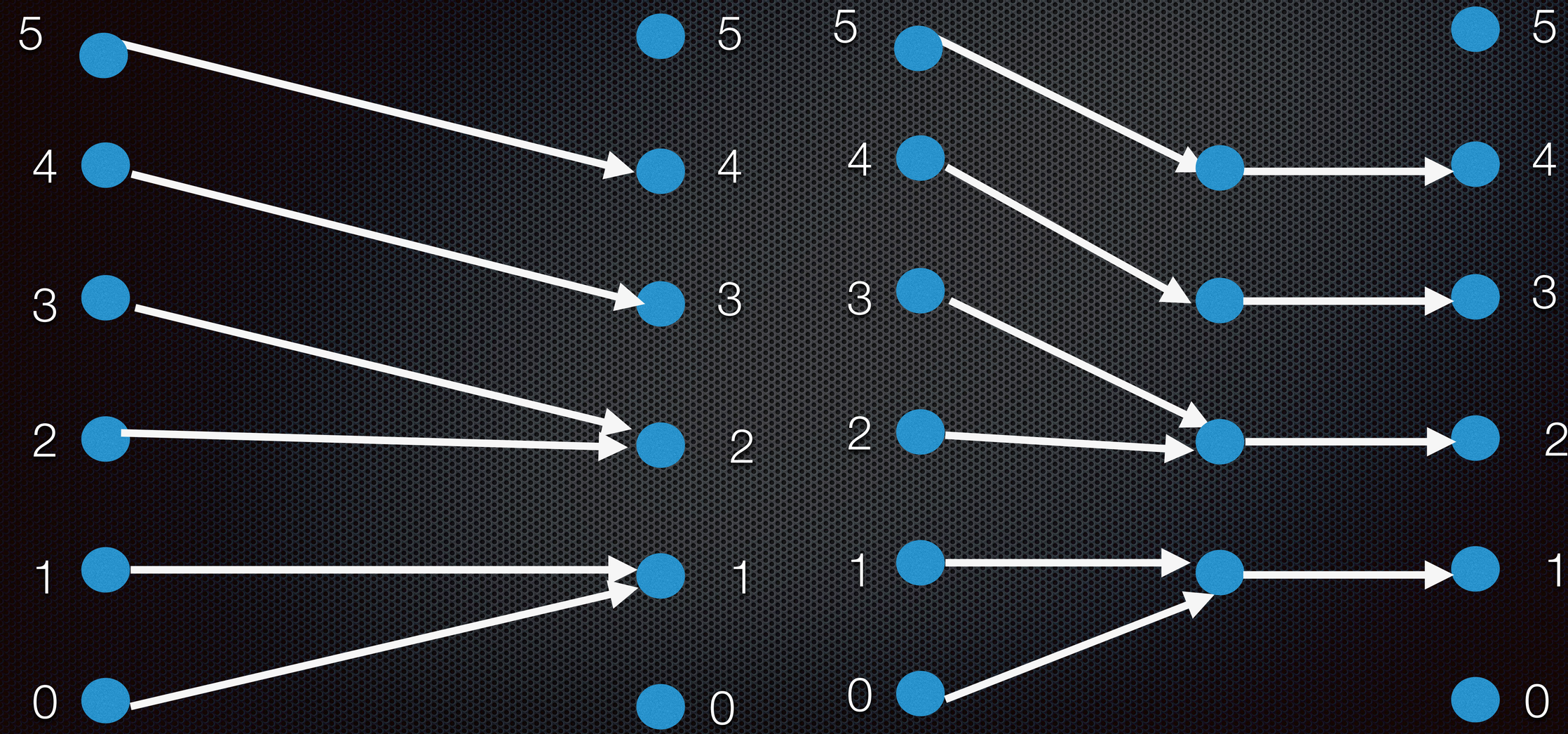
$\delta_2 : [n] \to [n+1]$



Primitive Surjections: $\sigma_2 : [n+1] \to [n]$



Every arrow is composition of primitive surjections + injections



Yoneda lemma for simplicial sets

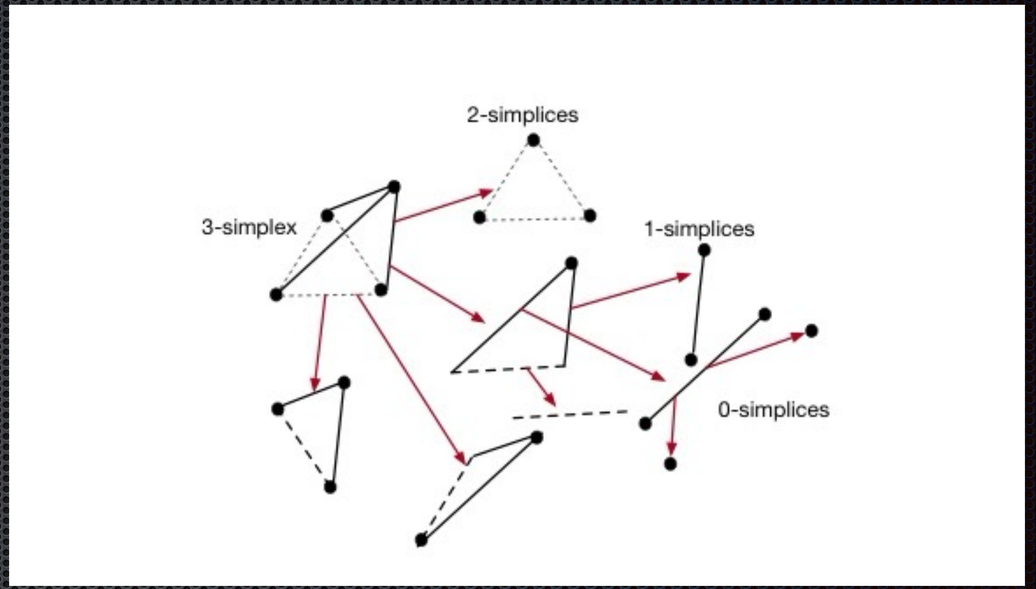
sSet: $\Delta \to \mathbf{Set}^{\Delta^{op}}$

$$\Delta^n = \Delta(-, [n])$$

$$\Delta^n_k = \Delta([k], [n])$$

$$d_i : f : [k] \to [n] \mapsto [k-1] \xrightarrow{\delta_i} [k] \xrightarrow{f} [n]$$

$$s_i : f : [k] \to [n] \mapsto [k+1] \xrightarrow{\sigma_i} [k] \xrightarrow{f} [n]$$



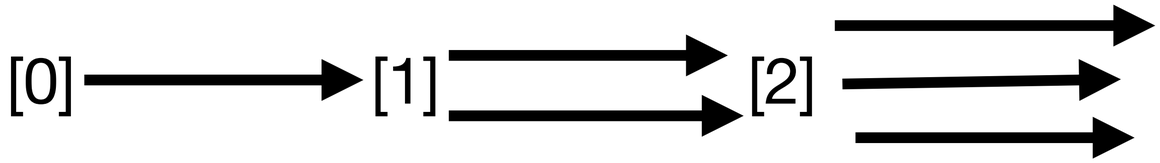
Nerve of a Category

- A simplicial set is a contravariant functor mapping the simplicial category to the category of sets

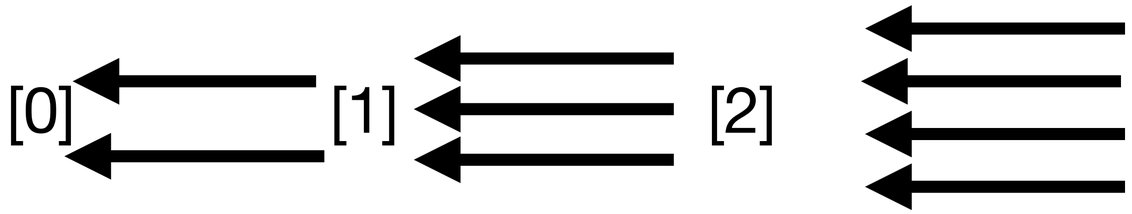
- Any category can be mapped onto a simplicial set by constructing its nerve

$$M_n(\mathscr{C}) = \{c_0 \xrightarrow{f_1} c_1 \xrightarrow{f_2} c_1 \cdots \xrightarrow{f_n} c_n \,|\, c_i \text{ is an object and } f_i \text{ is an arrow}\}$$

Simplicial Sets: Contravariant Functors



$$\delta_i^n : [n] \to [n+1]$$



$$\sigma_i^n : [n+1] \to [n]$$

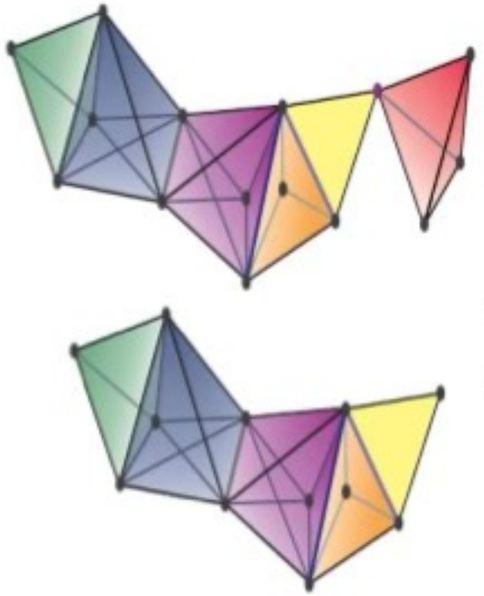
$$X_n : [n] \to X : \Delta^{op} \to X$$

Simplicial Sets vs. Categories

- Any category can be embedded faithfully into a simplicial set using its nerve
- The embedding is full and faithful (perfect reconstruction)
- Unfortunately, the converse is not possible
- Given a simplicial set, the left adjoint functor that maps it into a category is lossy!

Simplicial framework for generative AI

Simplicial learning is based on extension problems of inner and outer ``horns" of simplicial objects



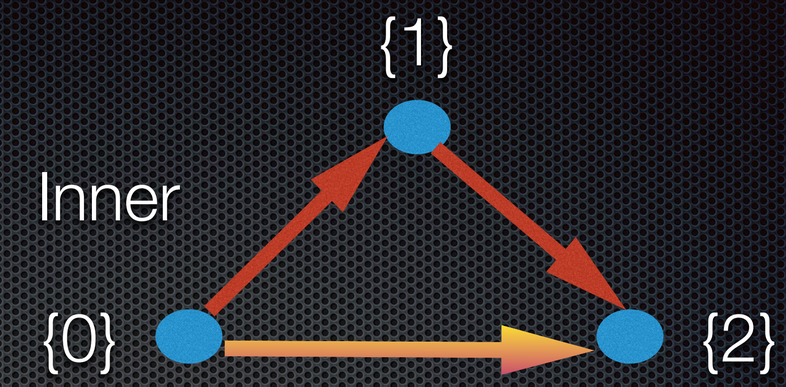
Each directed edge defines a morphism that represents a generative AI method

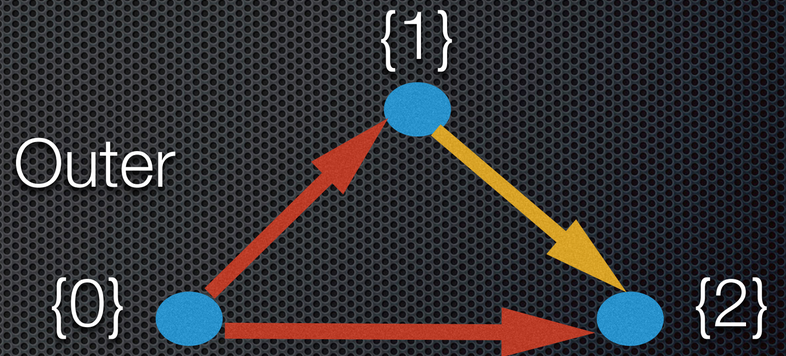
Each collection of simplices can be ``glued" on to compatible simplices through ``ports" that define the components of the simplex.

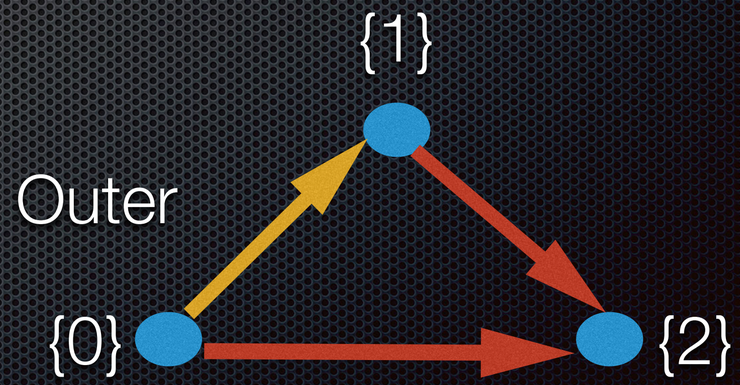
Horns of Simplicial Sets

A “horn” is a subset of a simplicial set

The “extension” problem is to “fill in” missing faces





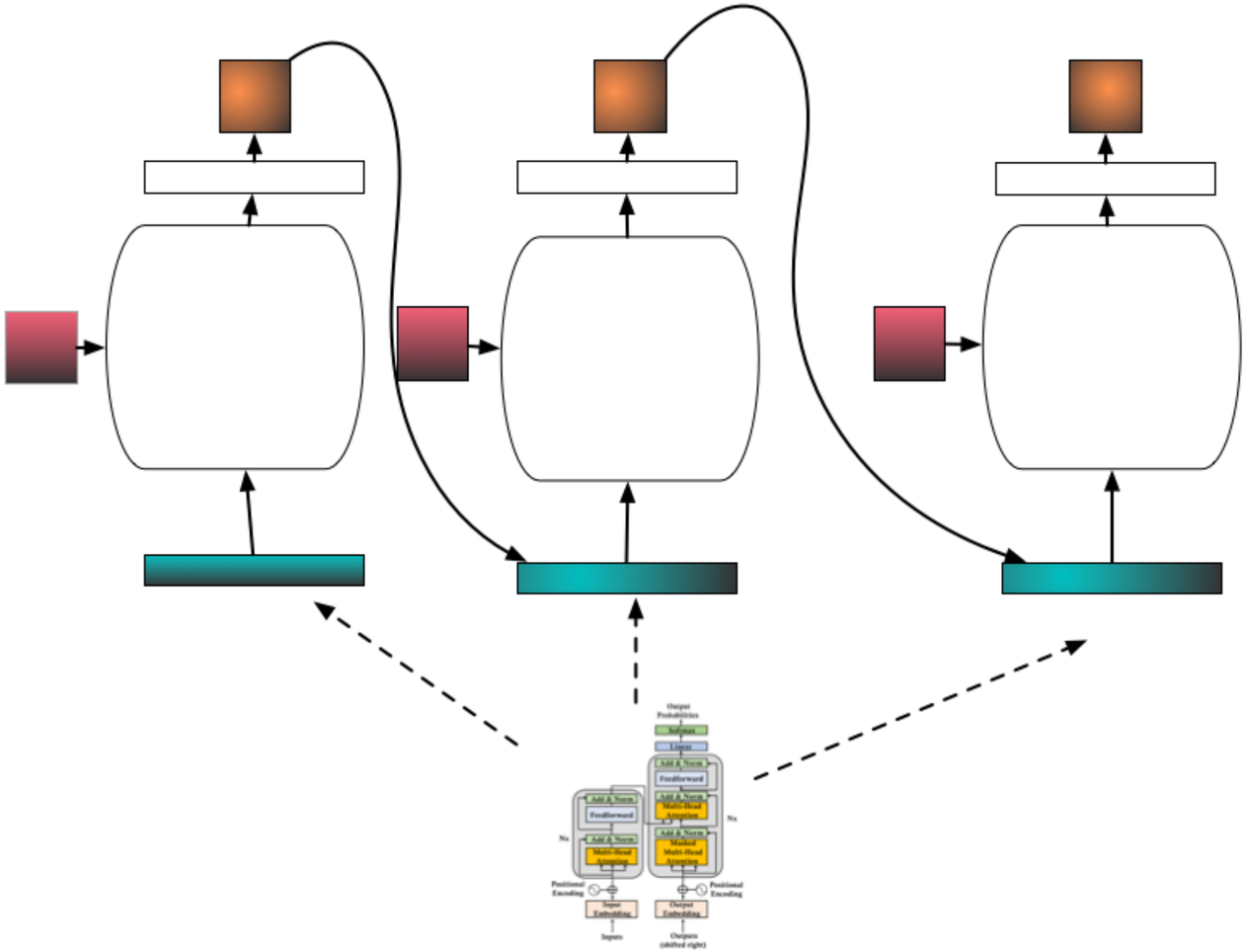


Lifting Diagrams for Simplicial Sets

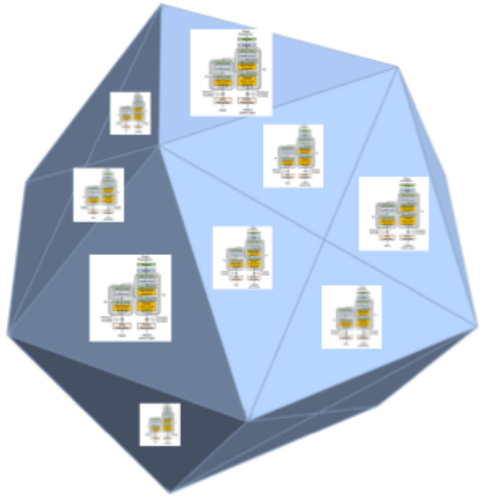
Recall that a lifting diagram was used as a way to model inference in databases.

Simplicial Generative AI

Sequential Generative AI Model



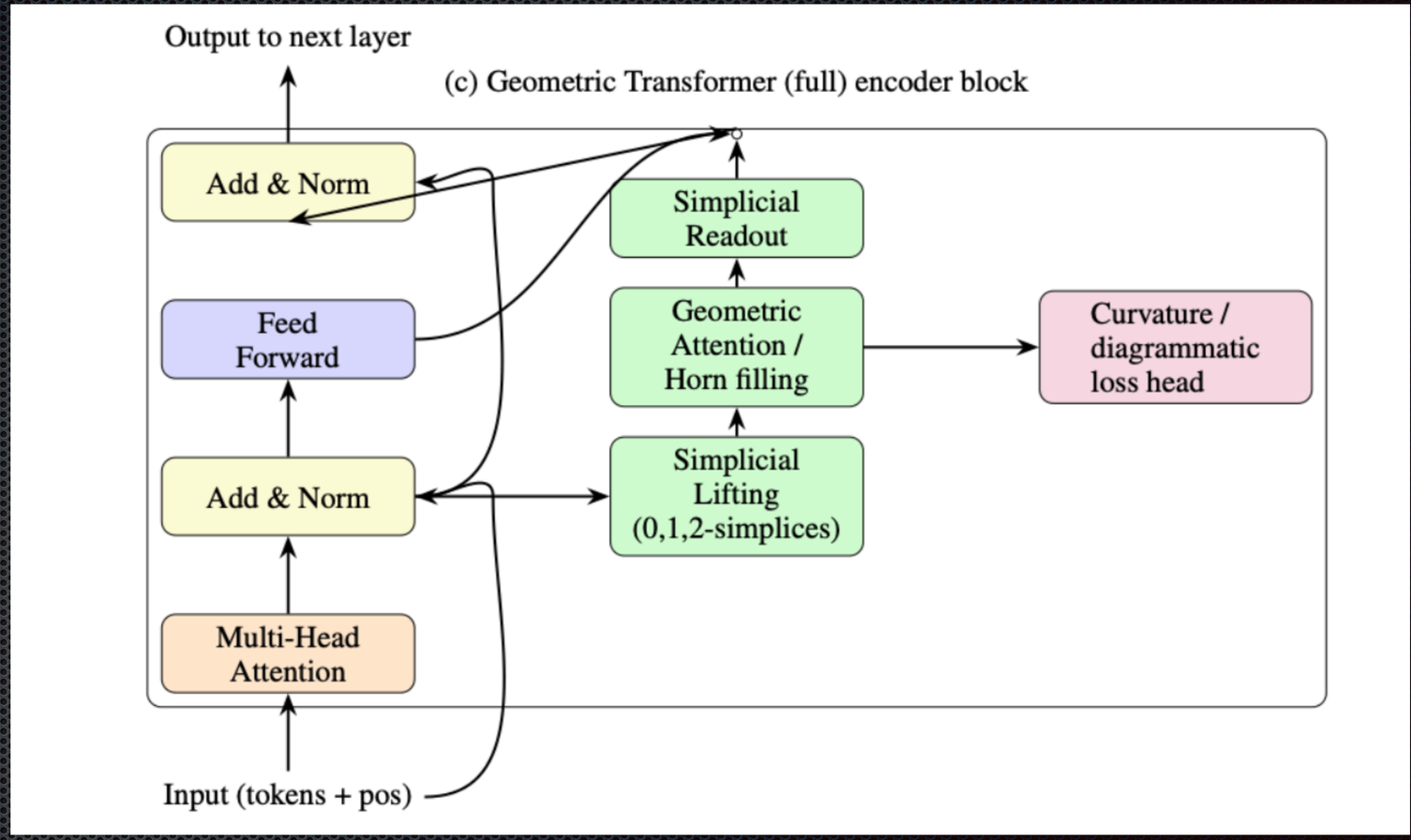
Simplicial Complex of Gen AI Models in GAIA



Geometric Transformers

GTs introduce explicit geometric inductive biases

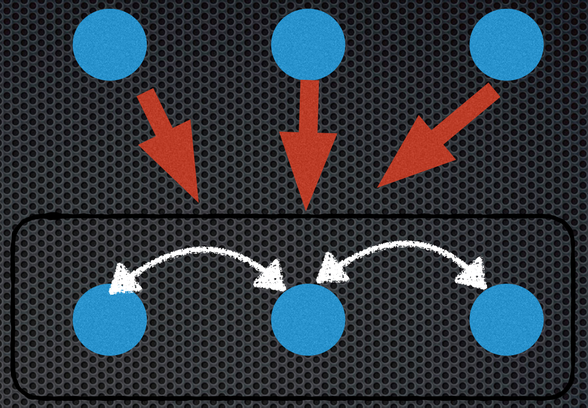
Token embedding is geometrically transported into a simplicial space



[Mahadevan, Categories for AGI, 2026]

GT-Lite

Uses a simple 1D convolution layer to “smooth” out token embeddings on adjacent tokens



1D convolution is the colimit

in a vector space embedding

Simplicial Message Passing: GT-Lite

Uses a discrete diffusion process

Attenuates local inconsistencies

Algorithm 9: GT-Lite Local Geometric Smoothing Operator C (1D Conv + Residual + LN)

Input: Token states $z \in \mathbb{R}^{B\times L\times d}$; convolution weights θ_C; smoothing scale α (e.g., 0.2); LayerNorm parameters θ_{LN}

Output: Smoothed token states $C(z) \in \mathbb{R}^{B\times L\times d}$

// 1) Convert to Conv1d layout and apply local neighborhood mixing

$z^\top \leftarrow \text{transpose}(z,\ (B,L,d) \to (B,d,L))$

$u^\top \leftarrow \text{Conv1D}_{\theta_C}(z^\top)$ ▷ kernel size 3, padding 1, output shape (B,d,L)

$u \leftarrow \text{transpose}(u^\top,\ (B,d,L) \to (B,L,d))$

// 2) Residual update and normalization

$z' \leftarrow z + \alpha \cdot u$

$C(z) \leftarrow \text{LayerNorm}_{\theta_{\text{LN}}}(z')$

return $C(z)$

[Mahadevan, Categories for AGI, 2026]

Diagrammatic Backpropagation

- Central idea: penalize deviations from compositionality
- Propagate gradients around loops!
- By explicitly measuring the lack of compositionality, we get a novel error signal!

[Mahadevan, Categories for AGI, 2026]

Algorithm 7: Diagrammatic Backpropagation Training Loop (code-faithful)

Input: Model M_θ (Transformer or GT variant); dataset iterator $\mathcal{B}$ yielding minibatches (s, y); optimizer AdamW; steps T; warmup steps T_w; weights λ_{db}, λ_{geom}; clipping norm c; small constant ε

Output: Trained parameters θ

Initialize: set global step $t \leftarrow 0$
for $t = 1$ **to** T **do**
  Sample minibatch $(s, y) \sim \mathcal{B}$ ▷ s: tokens (or inputs), y: labels/targets
  // 1) Forward pass for the primary task
  $\hat{y} \leftarrow M_\theta(s)$
  $\mathcal{L}_{\text{task}} \leftarrow \text{LossTask}(\hat{y}, y)$ ▷ e.g., cross-entropy for LM
  // 2) Optional DB control loss (commutator-energy terms)
  $\mathcal{L}_{\text{db}} \leftarrow 0$
  if $\lambda_{\text{db}} > 0$ **then**
    $(\text{Obs}, \{\text{Obs}_\ell\}) \leftarrow \text{CommProbe}(M_\theta, s; \varepsilon)$ ▷ Algorithm 6
    $\mathcal{L}_{\text{db}} \leftarrow \text{Obs}$ ▷ scalar average over layers; can also weight layers/heads
  end if
  // 3) Optional diagrammatic curvature loss (triangle/square residuals)
  $\mathcal{L}_{\text{geom}} \leftarrow 0$
  if $\lambda_{\text{geom}} > 0$ **and** (s, y) includes a compiled diagram/simplicial instance **then**
    ▷ For diagram benchmarks, s contains $(\mathcal{D}, x)$ with constraints.
    $\mathcal{L}_{\text{geom}} \leftarrow \text{CurvatureEnergy}(M_\theta, s)$ ▷ triangle/square residual energy; cf. v1 Alg. (learned energy)
  end if
  // 4) Warmup schedule for auxiliary terms (optional but common)
  $\lambda_t^{\text{db}} \leftarrow \lambda_{\text{db}} \cdot \min(1, t/T_w)$
  $\lambda_t^{\text{geom}} \leftarrow \lambda_{\text{geom}} \cdot \min(1, t/T_w)$
  // 5) Total loss and parameter update
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} + \lambda_t^{\text{db}} \mathcal{L}_{\text{db}} + \lambda_t^{\text{geom}} \mathcal{L}_{\text{geom}}$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ ▷ implemented via AdamW step
  Clip gradients: $\|\nabla_\theta \mathcal{L}\| \leq c$
  Optional: normalize selected parameter groups (e.g., edge transforms / relation embeddings)
end for
return θ

Simplicial Message Passing: GT-Full

Uses a discrete diffusion process

Attenuates local inconsistencies

Algorithm 8: GT-Full Geometric Transport Operator G (Simplicial/Graph Message Passing)

Input: Token states $z \in \mathbb{R}^{B\times L\times d}$; fixed adjacency `edge_index` $\in \mathbb{N}^{2\times E_{\max}}$ over positions $0,\ldots,L_{\max}-1$; relation ids `rel_ids` $\in \mathbb{N}^{E_{\max}}$; domain ids `dom_ids` $\in \mathbb{N}^{E_{\max}}$; depth K; learnable parameters θ_G

Output: Transported token states $G(z) \in \mathbb{R}^{B\times L\times d}$

// 1) Flatten batch and restrict edges to current sequence length
$h \leftarrow \text{reshape}(z,\ (B\cdot L)\times d)$
$m \leftarrow \{e \in [1..E_{\max}] : \texttt{edge\_index}[0,e] < L\}$ ▷ mask endpoints within L
$E \leftarrow \texttt{edge\_index}[:,m]$; $r \leftarrow \texttt{rel\_ids}[m]$; $dID \leftarrow \texttt{dom\_ids}[m]$

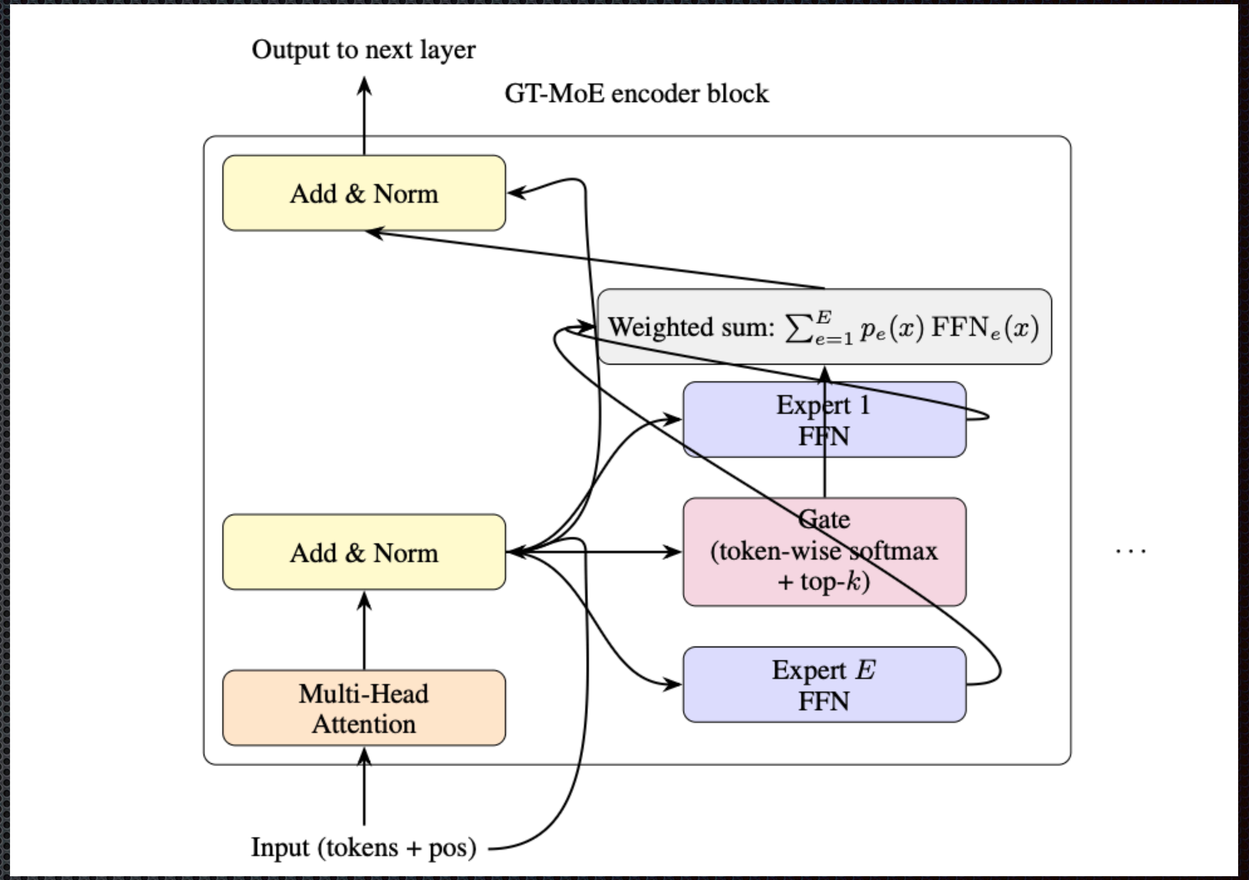
// 2) Iterate K rounds of geometric message passing
for $k = 1$ **to** K **do**
  Initialize message accumulator $M \leftarrow 0 \in \mathbb{R}^{(BL)\times d}$
  for each directed edge $(u \leftarrow v)$ in E **do**
    $e \leftarrow (u,v)$
    Edge embedding: $e_{\text{emb}} \leftarrow \text{Emb}_{rel}(r_e) + \text{Emb}_{dom}(dID_e)$
    Compute message: $m_{v\to u} \leftarrow \phi_\theta(h_v, h_u, e_{\text{emb}})$ ▷ e.g., MLP on $(h_v, h_u, e_{\text{emb}})$ or attention-style score
    $M_u \leftarrow M_u + m_{v\to u}$
  end for
  Degree normalize: $M \leftarrow \text{DegNorm}(M, E)$
  Update rule: $h \leftarrow \text{Update}_\theta(h, M)$ ▷ e.g., GRUCell/residual MLP + LayerNorm
end for

// 3) Reshape back to token grid
$G(z) \leftarrow \text{reshape}(h,\ B\times L\times d)$
return $G(z)$

[Mahadevan, Categories for AGI, 2026]

Mixture of Experts GT

GT-MoE combines explicit geometric inductive biases with gated experts



[Mahadevan, Categories for AGI, 2026]

Simplicial Message Passing: GT-MoE

Uses a discrete diffusion process

Attenuates local inconsistencies

Algorithm 10: GT-MoE Mixing Operator M (Gating + Top-k + Expert Aggregation)

Input: Token states $z \in \mathbb{R}^{B\times L\times d}$; experts $\{E_e\}_{e=1}^{E}$; gating network $g_\theta : \mathbb{R}^d \to \mathbb{R}^E$; optional top-$k$ value k; LayerNorm parameters θ_{LN}

Output: Mixed token states $M(z) \in \mathbb{R}^{B\times L\times d}$

// 1) Compute gating probabilities per token

$\ell \leftarrow g_\theta(z)$ ▷ $\ell \in \mathbb{R}^{B\times L\times E}$ (gate logits)

$p \leftarrow \mathrm{softmax}(\ell, \text{ over experts})$ ▷ $p \in \mathbb{R}^{B\times L\times E}$

// 2) Optional sparsification (top-k routing) and renormalization

if $k < E$ **then**

  $(v, I) \leftarrow \mathrm{TopK}(p, k)$ ▷ $v, I \in \mathbb{R}^{B\times L\times k}$

  $\tilde{p} \leftarrow 0$ ▷ initialize $\tilde{p} \in \mathbb{R}^{B\times L\times E}$

  $\tilde{p}[\cdot,\cdot,I] \leftarrow v$ ▷ scatter top-k weights into expert slots

  $p \leftarrow \tilde{p} / \left(\sum_{e=1}^{E} \tilde{p}_e + \epsilon\right)$ ▷ renormalize per token

end if

// 3) Apply experts and aggregate weighted outputs

$u \leftarrow 0 \in \mathbb{R}^{B\times L\times d}$

for $e = 1$ **to** E **do**

  $u \leftarrow u + p_e \odot E_e(z)$ ▷ p_e broadcast to shape (B, L, d); $\odot$ elementwise

end for

// 4) Residual update and normalization

$z' \leftarrow z + u$

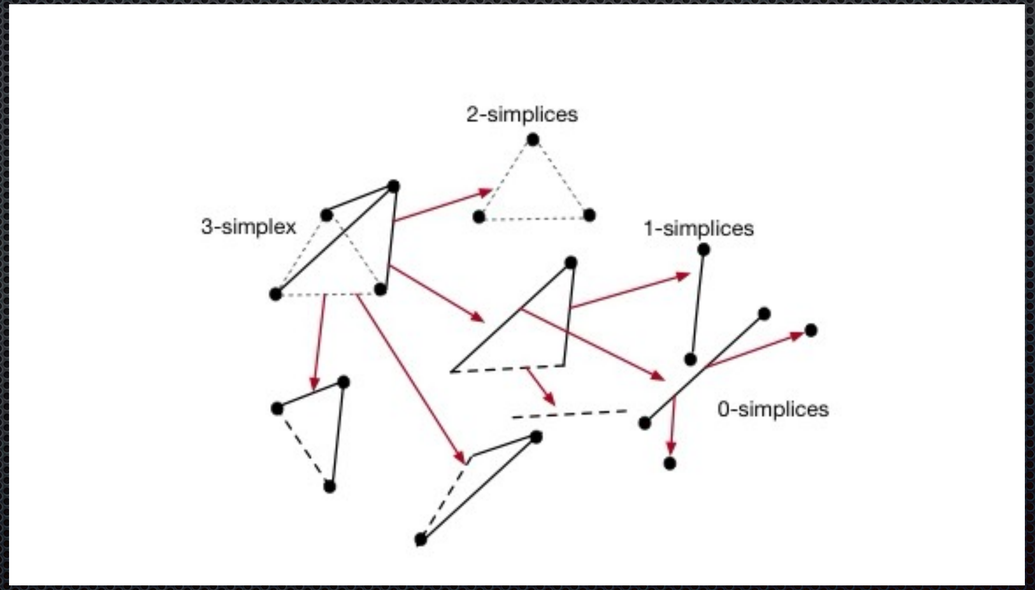
$M(z) \leftarrow \mathrm{LayerNorm}_{\theta_{\mathrm{LN}}}(z')$

return $M(z)$

[Mahadevan, Categories for AGI, 2026]

Experimental Results

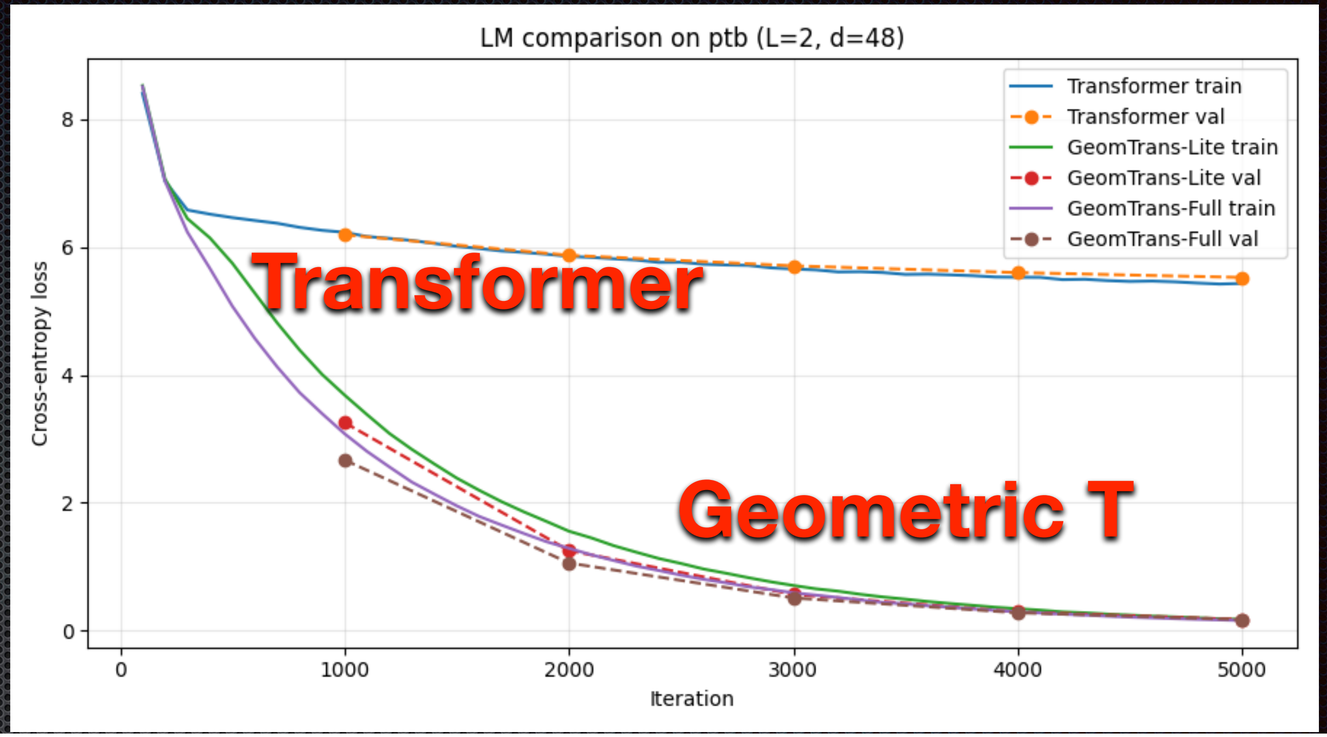
Domain / Task	Base complex (0/1/2-simplices)	Features on 0-simplices	Objective
PTB, WikiText-103 LM	Tokens / adj. pairs / (none)	Word + positional embeddings	Next-token CE loss
Triangle / DB synthetic	Nodes / edges / triangles	Node features (structural)	Triangle / motif classification
FATE causal motifs	Vars / causal edges / motifs	Variable embeddings	Causal motif / FATE score
Large causal manifolds	Vars / causal edges / small motifs	Text embeddings of variables	Unsupervised GT refinement
MNIST diffusion	Pixels / grid edges / 2×2 patches	Pixel or latent codes	Denoising / diffusion loss

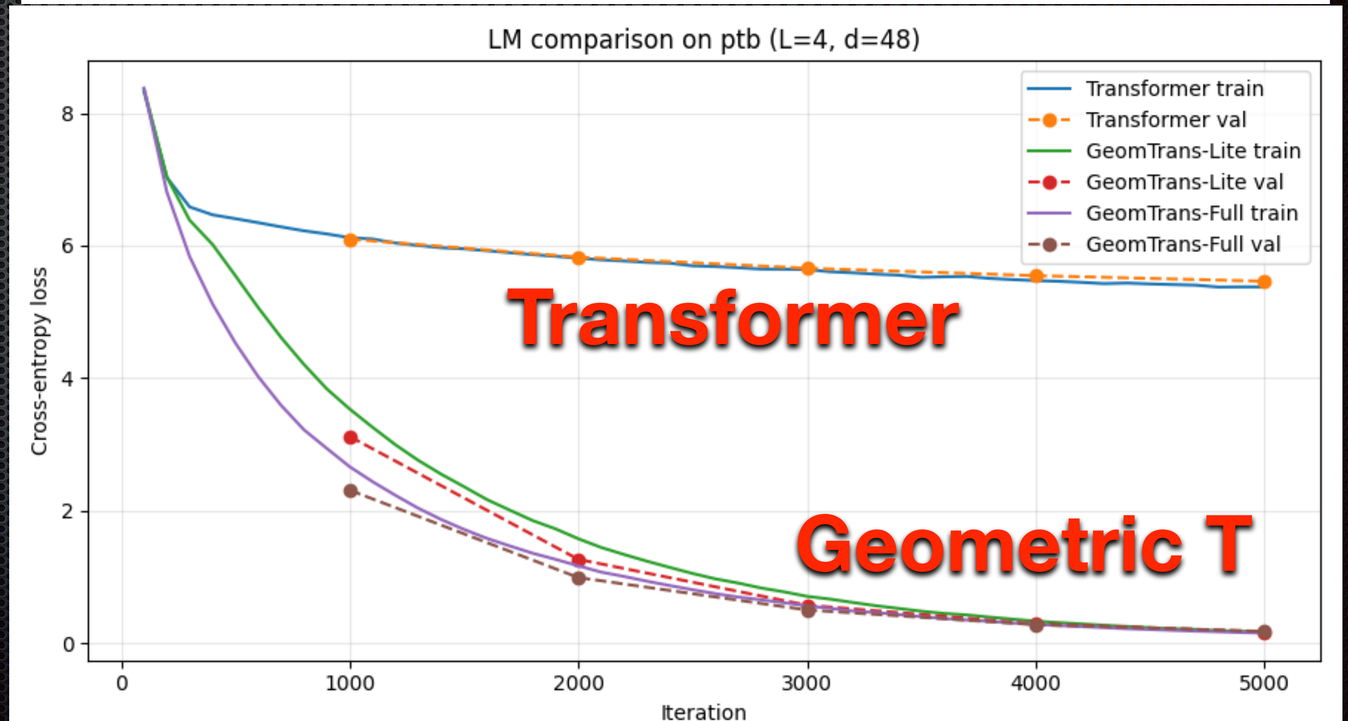


Language Modeling

Penn Tree Bank (PTB)

The Penn Treebank (PTB) project selected 2,499 stories from a three year Wall Street Journal (WSJ) collection

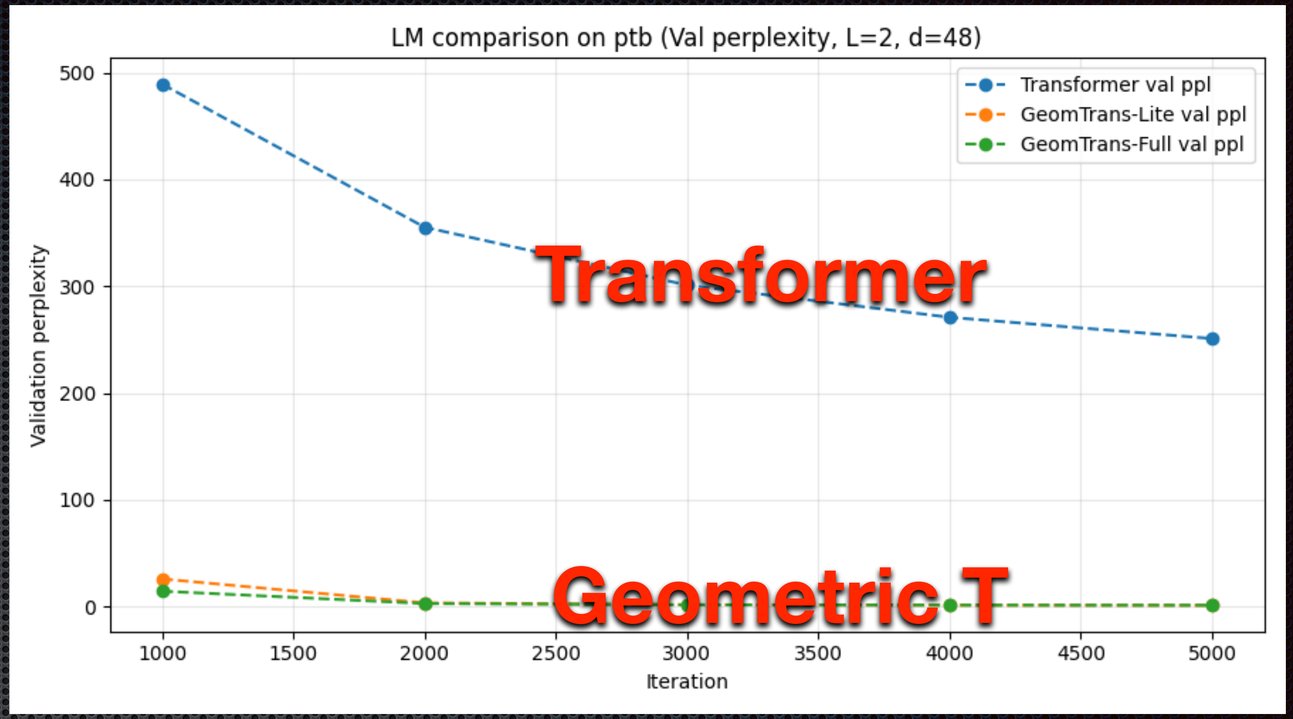


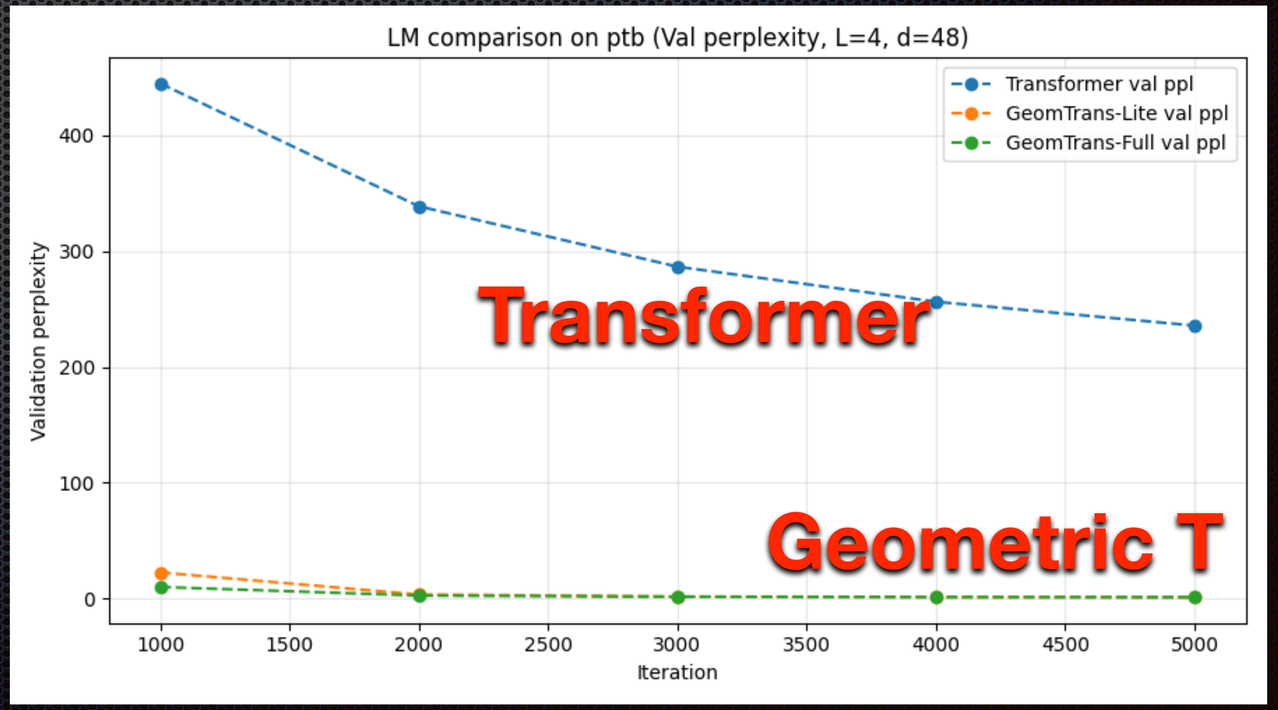


Language Modeling

Penn Tree Bank (PTB)

The Penn Treebank (PTB) project selected 2,499 stories from a three year Wall Street Journal (WSJ) collection



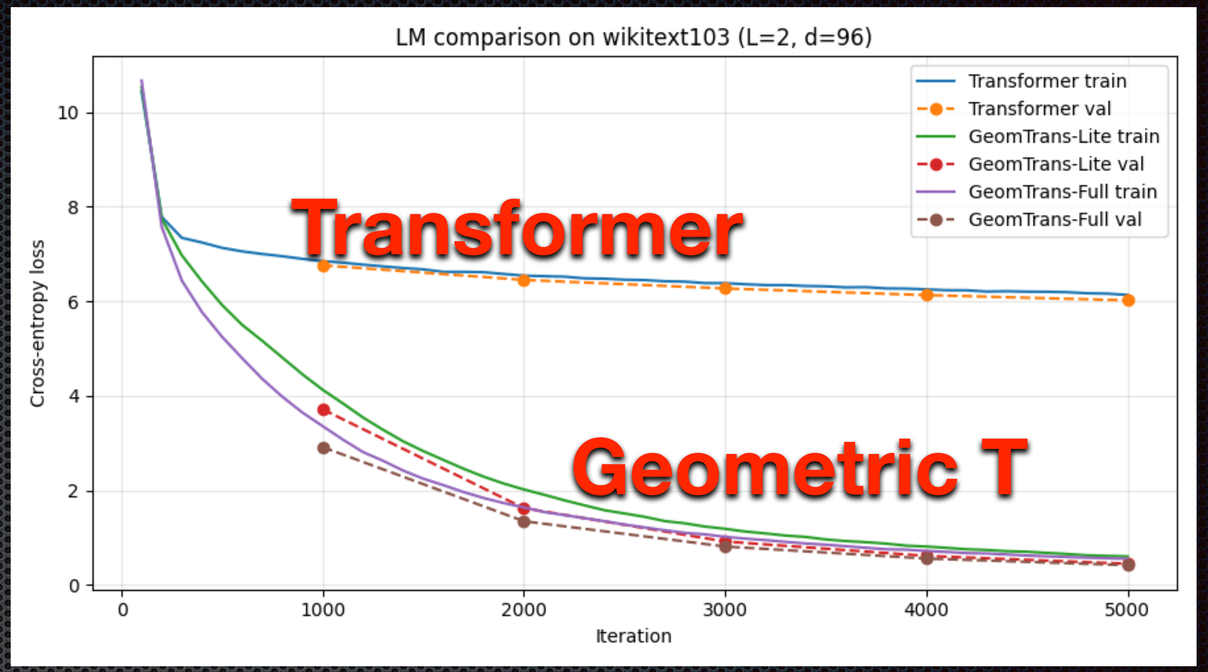


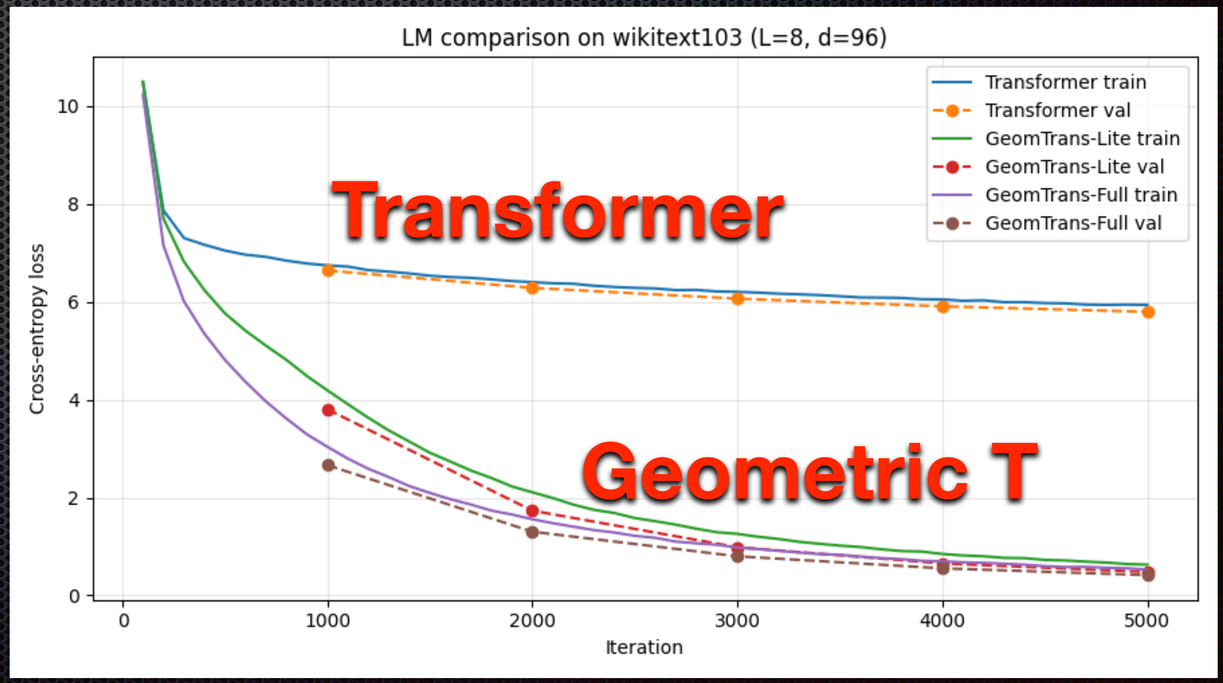
Language Modeling

Wiki-103: 100-million token dataset of carefully curated Wikipedia articles

The WikiText-103 dataset include

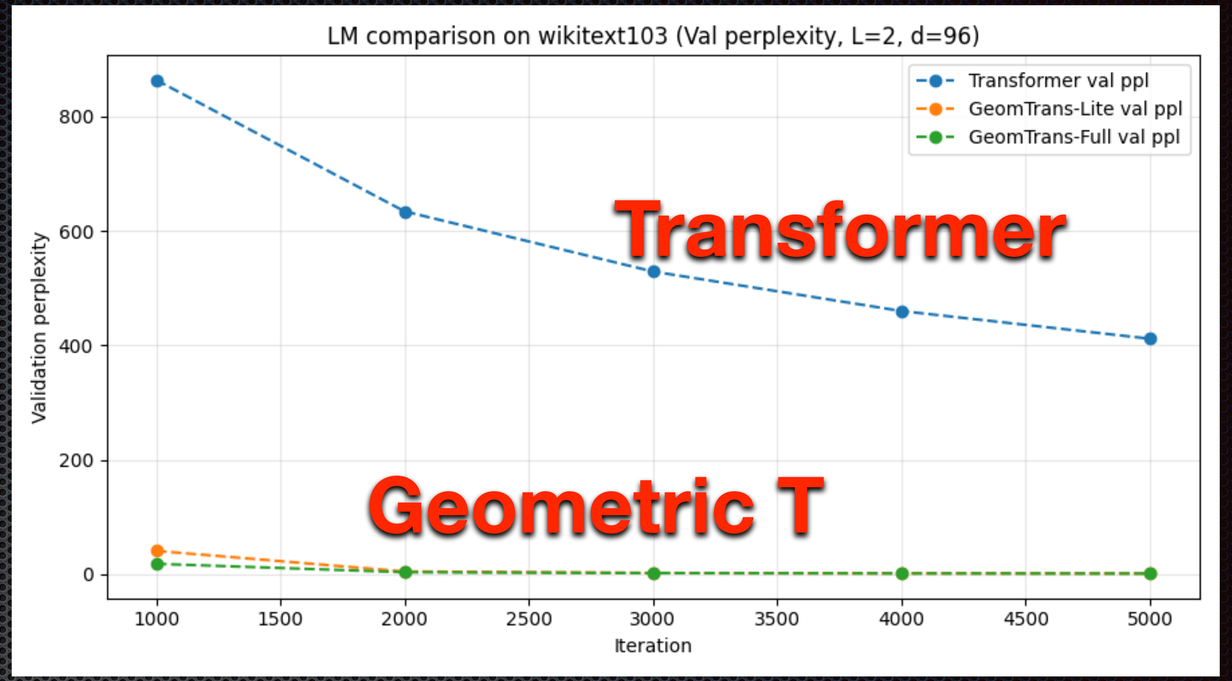
- 28,475 Wikipedia articles
- Over 100 million words in English
- Complete, untruncated and low-noise texts
- A raw format (TXT), suitable for training autoregressive or bidirectional models

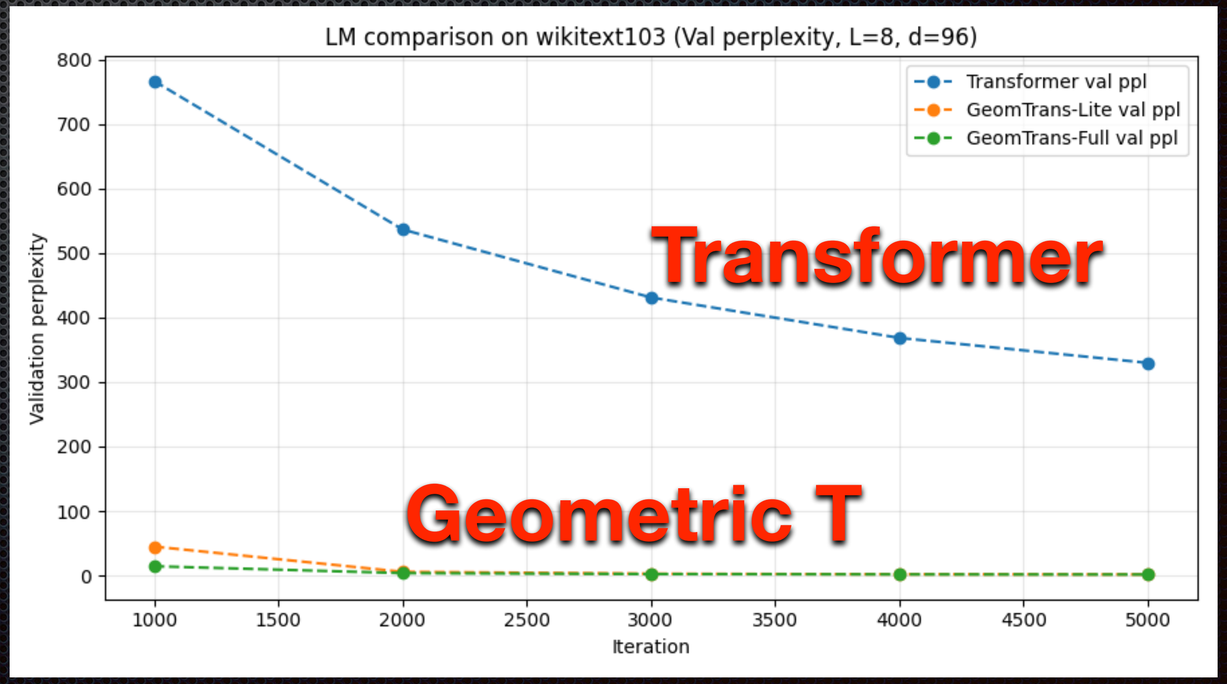




Language Modeling

Wiki-103: 100-million token dataset of carefully curated Wikipedia articles





DEMOCRITUS: Causality from Language

[Mahadevan, Large Causal Models from Large Language Models, Arxiv 2025]

The Washington Post
Democracy Dies in Darkness

When did humanity take its first step? Scientists say they now know.

A new analysis of fossils uncovered in Central Africa offers additional evidence that a human ancestor walked upright 7 million years ago.

January 2, 2026

A Sahelanthropus tchadensis skull found in Chad. (Philippe Psaila/Science Source)

By Dino Grandoni

More than two decades ago, scientists digging in Central Africa unearthed the 7-million-year-old remains of what may be one of the earliest known human ancestors.

AI Overview

Summary is AI-generated, newsroom-reviewed.

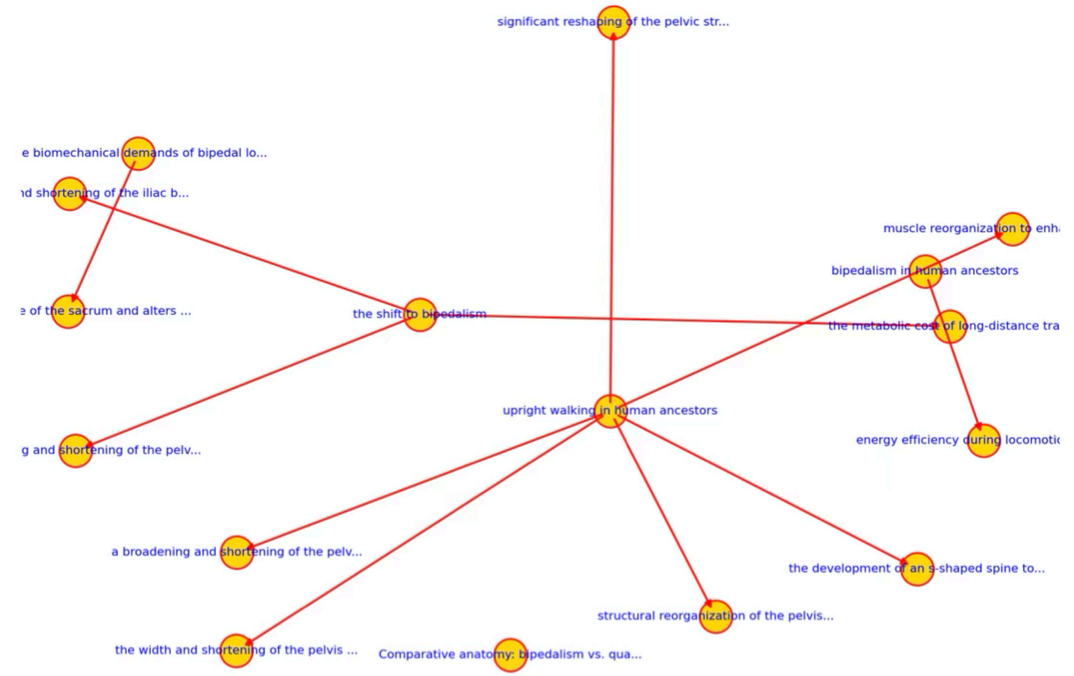
A study published in Science Advances suggests Sahelanthropus tchadensis, a 7-million-year-old ancestor, walked upright, indicating early bipedalism in human evolution. The analysis of limb bones, particularly the femoral tubercle, supports this claim. However, the debate continues as some scientists argue the fossils are too damaged to confirm bipedalism. Further discoveries are needed to resolve the controversy.

Read the full article for more on:

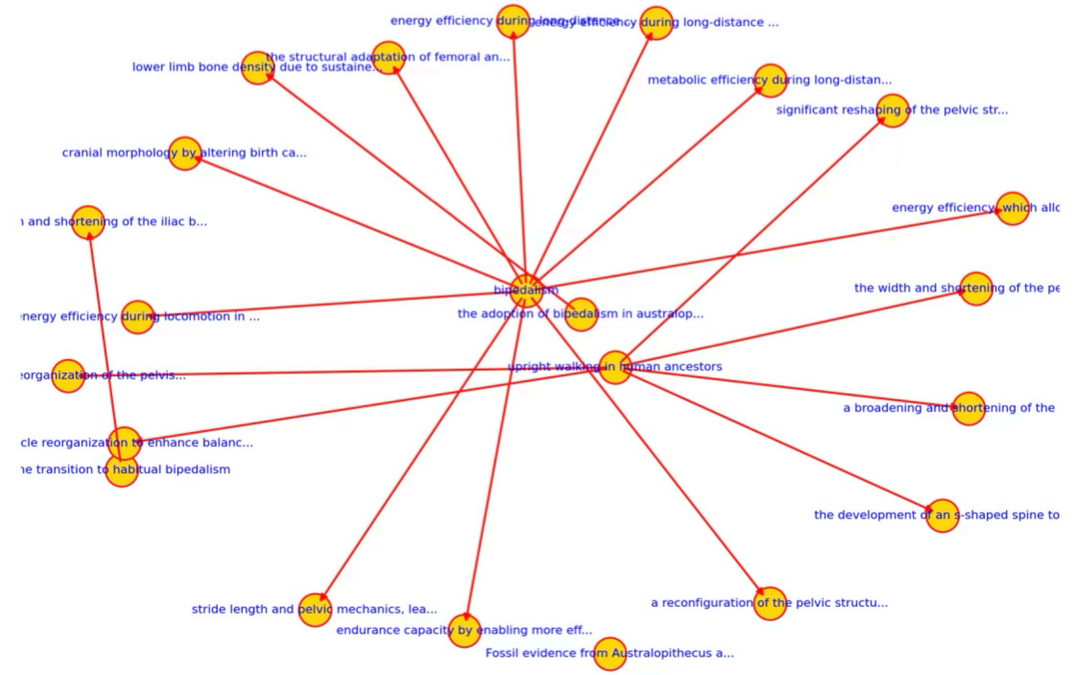
- The significance of the femoral tubercle in determining bipedalism.
- Why some scientists remain skeptical about the study's conclusions.
- Future plans for fossil hunting in Chad's Djurab Desert.

Did our AI help? Share your thoughts.

LLM summary



umanity_take_its_first_step_scientists_say_they_no_70ccf7f26a62 | rank 1 | score=0.155 | Fossil evidence from Australopithecus afa



DEMOCRITUS

Causal DeepDive

DEMOCRITUS explains the causal background to any story

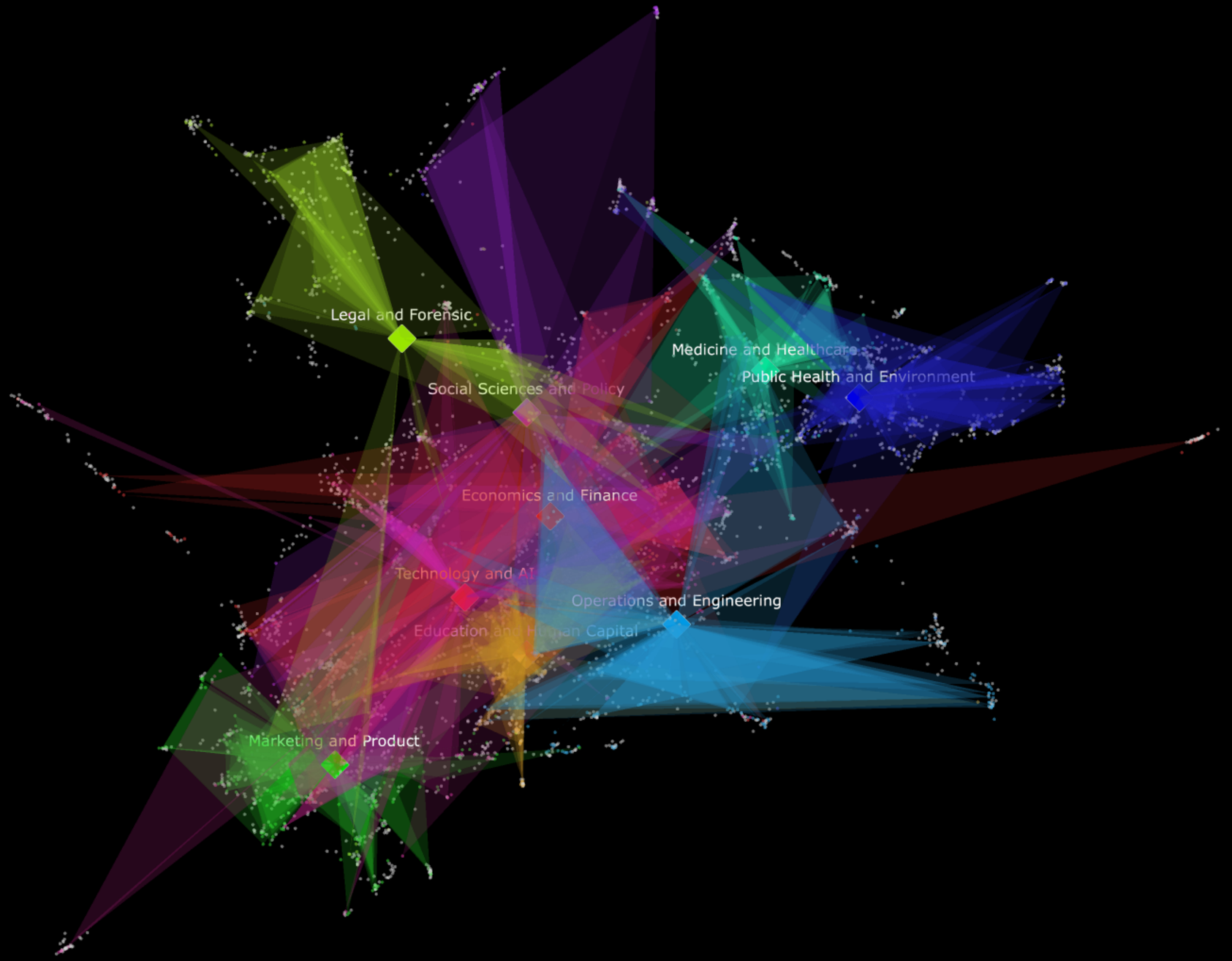
Why, not what!

Democritus Atlas Summary (v0)

- Nodes: 501
- Edges (unique): 65
- Edge-support rows: 717
- SCC modules (size>1): 0

Top stable bonds (by support_docs, then score_sum)

rank	support_docs	support_lcms	score_sum	controversy	rel_type	src	dst
1	1	15	1.462	0.000	INFLUENCES	`bipedalism`	`metabolic efficiency during long-distance travel by optimizing stride mechanics and muscle utilization`
2	1	13	1.328	0.000	INCREASES	`bipedalism`	`energy efficiency during locomotion in early hominins by reducing the metabolic cost of walking over long distances`
3	1	13	1.264	0.000	INCREASES	`bipedalism`	`energy efficiency during long-distance locomotion by reducing the metabolic cost of walking compared to quadrupedal gaits`
4	1	13	1.256	0.000	INFLUENCES	`bipedalism`	`endurance capacity by enabling more effective heat dissipation and sustained pacing over extended distances`
5	1	12	1.209	0.000	INCREASES	`bipedalism`	`energy efficiency which allows for greater metabolic resources to be allocated to brain development`
6	1	11	1.201	0.000	INFLUENCES	`bipedalism`	`structural adaptation of femoral and tibial bones by altering stress distribution patterns over evolutionary time`
7	1	11	1.103	0.000	INFLUENCES	`bipedalism`	`stride length and pelvic mechanics leading to more sustained and efficient terrestrial movement`
8	1	6	0.464	0.000	INFLUENCES	`reduced forest cover`	`selection for upright walking by favoring energy-efficient movement over long distances`
9	1	4	0.403	0.000	INCREASES	`changes in environmental conditions that favored energy-efficient locomotion over long distances`	`likelihood of bipedalism in early hominins`
10	1	6	0.156	0.000	INFLUENCES	`abnormal femoral tubercle position`	`knee joint alignment during locomotion by altering the line of pull on the patellar tendon`
11	1	2	0.136	0.000	INFLUENCES	`fossil structure of ardipithecus`	`emergence of efficient terrestrial walking in early hominins by demonstrating adaptations in the pelvis and foot that support bipedal locomotion`
12	1	6	0.129	0.000	CAUSES	`abnormal femoral tubercle position`	`increased postoperative patellar instability after tibial tubercle osteotomy`
13	1	2	0.128	0.000	INCREASES	`muscle reorganization for balance and propulsion in bipedalism`	`efficiency of upright walking in human ancestors`
14	1	6	0.081	0.000	CAUSES	`larger brain size in primates`	`enhanced problem-solving abilities by enabling greater neural complexity and cognitive flexibility`
15	1	4	0.076	0.000	INCREASES	`lateral displacement of the femoral tubercle`	`patellar maltracking leading to altered patellofemoral alignment`

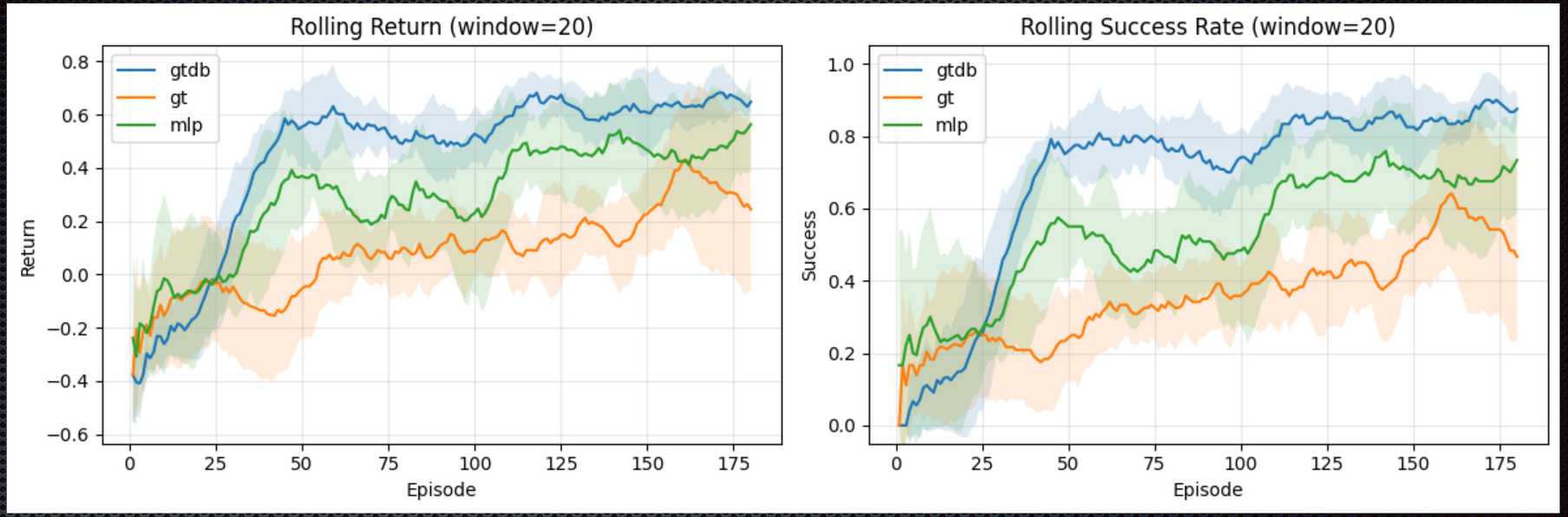


100,000 causal claims

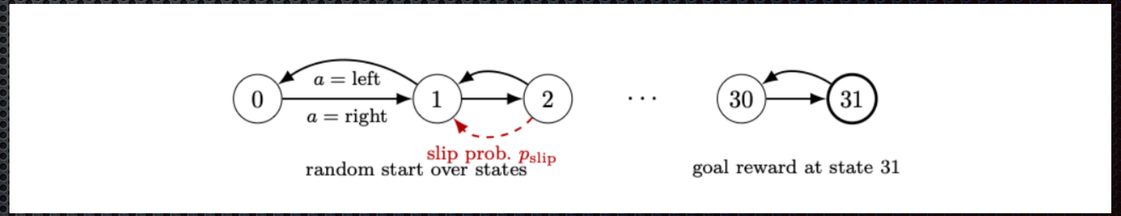
Domain	# Triangles
Economics and Finance	1140
Education and Human Capital	950
Legal and Forensic	553
Marketing and Product	1059
Medicine and Healthcare	932
Operations and Engineering	1281
Public Health and Environment	1336
Social Sciences and Policy	1015
Technology and AI	782

Table 1. Simplicial complex induced by local causal neighborhoods in the 100K-claim dataset.

Reinforcement Learning with Diagrammatic Backpropagation



MDPs can be modeled as universal coalgebras



The GitHub repo has a simple demo of RL with GT and DB

Summary

- The ordinal category Δ has as objects [n], and arrows weakly order-preserving functions on [n]
- Simplicial objects are contravariant functors $X[n] : \Delta \to C$
- GAIA: Generative AI Architecture based on simplicial objects
- Geometric Transformer is a GAIA architecture
- Diagrammatic Backpropagation is used to train a GT model