Deep sequence models tend to memorize geometrically (it's unclear why!)

Vaishnavh Nagarajan

[recent preprint]

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

Shahriar Noroozizadeh*, CMU

Elan Rosenfeld Google Deepmind

[ICML 2024]

The Pitfalls of Next-Token Prediction

Gregor Bachmann [*1] Vaishnavh Nagarajan [*2]

Gregor Bachmann*, Apple

Consider a dataset of "atomic facts"

Input → Output:

a→b

b→c

c→d

d→e

No high-level rules etc.,
Just arbitrary associations!

Not an arithmetic/reasoning task.

Pure memorization task: nothing to "generalize" or "learn" or "represent" here.

Question: How would a deep sequence model store these associations in its parameters?

The popular intuition: *associative* memory

Birth of a Transformer: A Memory Viewpoint

Alberto Bietti* Flatiron Institute
Vivien Cabannes FAIR, Meta
Diane Bouchacourt FAIR, Meta
Hervé Jégou FAIR, Meta
Léon Bot FAIR, M

Abstract

and the slower development of an "induction head" mechanism for the in-context bigrams. We highlight the role of weight matrices as associative memories, provide theoretical insights on how gradients enable their learning during training, and study the role of data-distributional properties.

Transformer Feed-Forward Layers Are Key-Value Memories

Mor Geva[1,2] **Roei Schuster**[1,3] **Jonathan Berant**[1,2] **Omer Levy**[1]
[1]Blavatnik School of Computer Science, Tel-Aviv University
[2]Allen Institute for Artificial Intelligence
[3]C

Figure 1: An illustration of how a feed-forward layer ulates a key-value memory. Input vectors (here, x

Understanding Factual Recall in Transformers via Associative Memories

Eshaan Nichani* Princeton University
Jason D. Lee Princeton University
Alberto Bietti Flatiron Institute

December 10, 2024

The popular intuition: *associative memory*

- Birth of a Transformer: A Memory Viewpoint, Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, Leon Bottou, NeurIPS 2023
- Scaling Laws for Associative Memories, Vivien Cabannes, Elvis Dohmatob, Alberto Bietti, ICLR 2024
- Understanding Finetuning for Factual Knowledge Extraction, Gaurav Ghosal, Tatsunori Hashimoto, Aditi Raghunathan, ICML 2024
- Towards a Theoretical Understanding of the 'Reversal Curse' via Training Dynamics, Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael Jordan, Jiantao Jiao, Yuandong Tian, Stuart Russell, NeurIPS 2024,
- Learning Associative Memories with Gradient Descent, Vivien Cabannes, Berfin Simsek, Alberto Bietti, 2024
- Understanding Factual Recall in Transformers via Associative Memories, Eshaan Nichani, Jason D. Lee, Alberto Bietti, 2024
- Do LLMs dream of elephants (when told not to)? Latent concept association and associative memory in transformers, Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, NeurIPS 2024

Associative memory:
atomic facts are stored as a lookup table

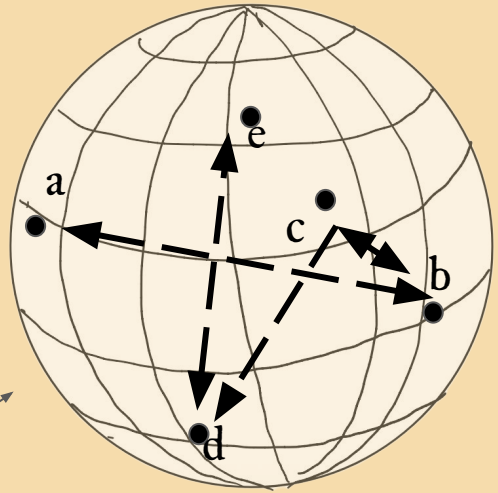
Input → Output:

a→b

b→c

c→d

d→e



Embedding "keys" **Φ**

is a matrix lookup with W_{assoc}

Associative memory:
atomic facts are stored as a lookup table

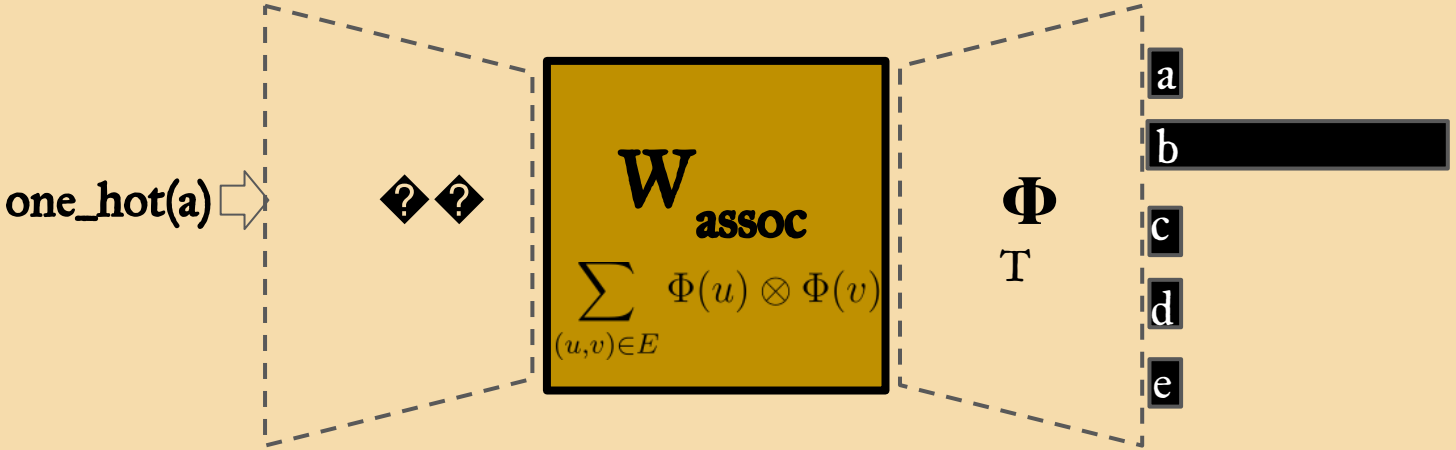
Input → Output:

a→b

b→c

c→d

d→e



embedder:
A matrix of
orthogonal/random
"keys"

lookup table:
The (rotated)
adjacency
matrix

unembedder

Outline

- Background: associative memory
- Geometric memory
- Implications and open questions
- Closing remarks

Finding 1: Deep sequence models (Transformers, Mamba, neural networks) tend to memorize *geometrically*, not *associatively*.

Geometric memory: an *alternative* way to fit our data

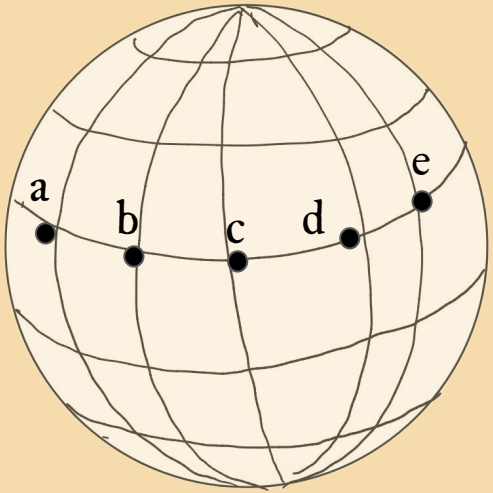
Input → Output:

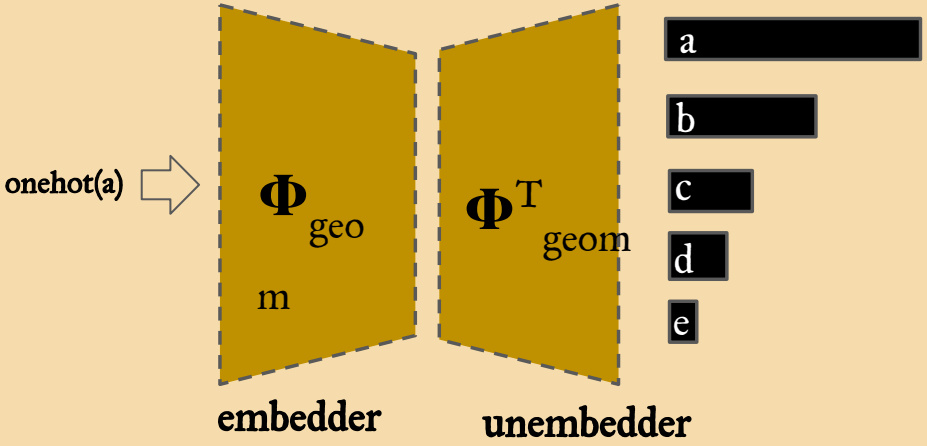
a→b

b→c

c→d

d→e





Highly organized **embeddings** Φ_{geom}

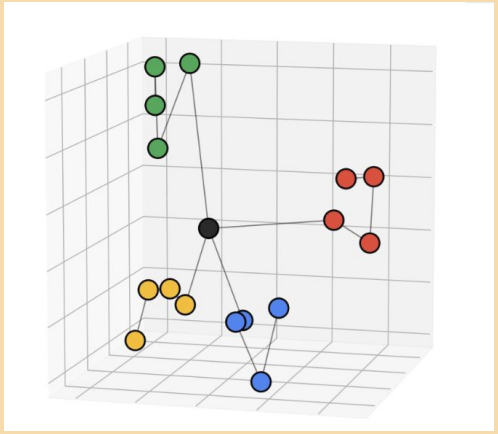
Think: (low-rank) factorization of adjacency matrix / graph spectrum

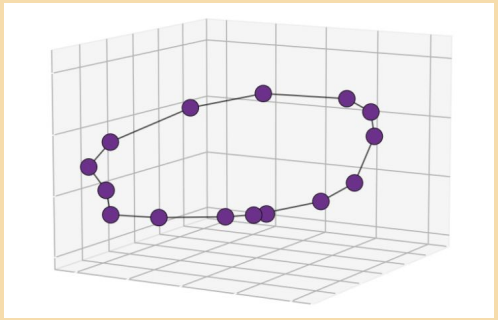
Rather than just storing "Canada →US", "US →Mexico", the model has inferred a map.

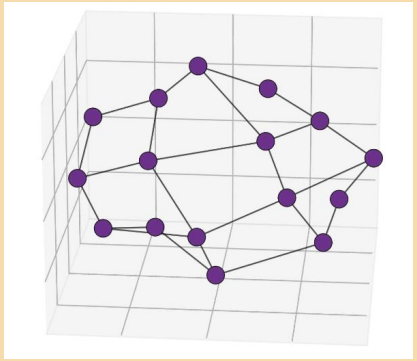
Some toy empirical evidence of geometric memory

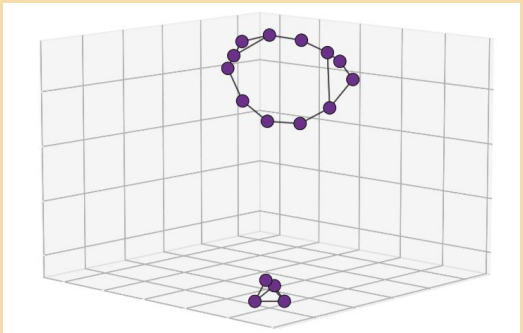
Token embeddings of a Transformer on various small graph datasets

(similarly true of Mamba and Neural network models – it's not the magic of attention!)









Geometries (of *concepts*) are well-known & familiar!

Linguistic Regularities in Continuous Space Word Representations

Tomas Mikolov*, Wen-tau Yih, Geoffrey Zweig
Microsoft Research
Redmond, WA 98052

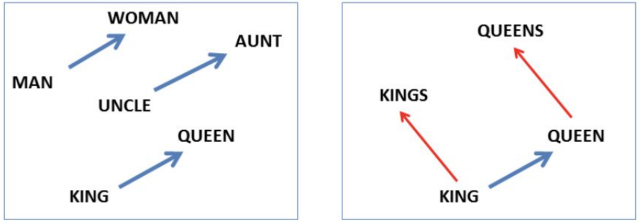


Figure 2: Left panel shows vector offsets for three word pairs illustrating the gender relation. Right panel shows a different projection, and the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word.

The Linear Representation Hypothesis and the Geometry of Large Language Models

Kiho Park [1] Yo Joong Choe [1] Victor Veitch [1]

Toy Models of Superposition

AUTHORS

Nelson Elhage*, Tristan Hume*, Catherine Olsson*, Nicholas Schiefer*, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg*, Christopher Olah‡

AFFILIATIONS

Anthropic, Harvard

PUBLISHED

Sept 14, 2022

* Core Research Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

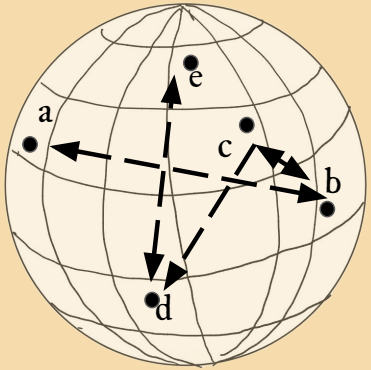
So what's new here/what's the big deal?

These embeddings are a form of parametric memory of atomic facts &

we must contrast it against associative memory.

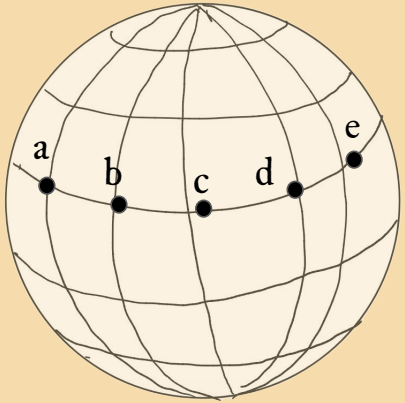
Finding 2: Unlike associative memory, geometric memory has synthesized novel *global* information *not* explicit in training data.

Associative memory: what you get is what you give (local associations)



one_hot(a) ⇨ ?? W_{assoc} $\sum_{(u,v)\in E} \Phi(u) \otimes \Phi(v)$ Φ^T → a, b, c, d, e

Geometric memory: what you get is *more* than is what you give



one_hot(a) → Φ_{geo} m, Φ^T_{geom} → a, b, c, d, e

Novel, global multi-hop information

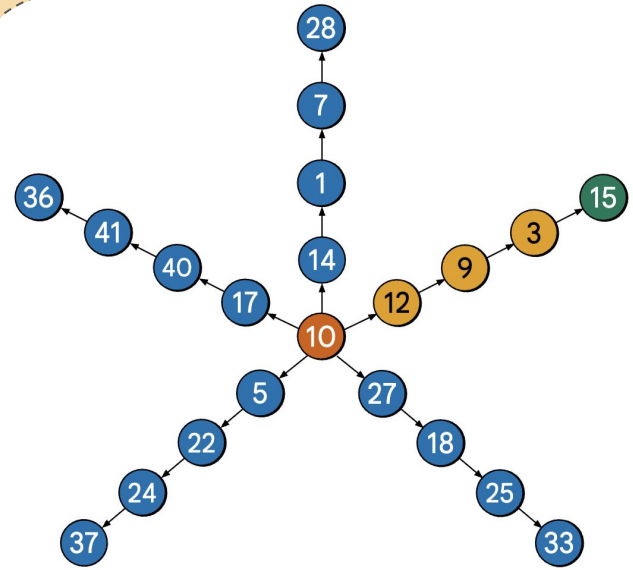
Outline

- Background: associative memory
- Geometric memory
- Implications & open questions
- Closing remarks

Implication 1: Geometric memory in a *Seq2Seq* model makes some hard forms of *implicit reasoning* learnable

Implicit reasoning = The next-token target requires a chain of thought but this isn't spelled out

A simple yet notoriously hard implicit reasoning task




```
Prefix:
(10,14) | ( 9, 3) |
(22,24) | (10,12) |
(25,33) | (17,40) |
 ...

Start: 10
Goal : 15

Target: 10,12,9,3,15
```

Context: randomized path-star graph's adjacency list

Target: start →goal path

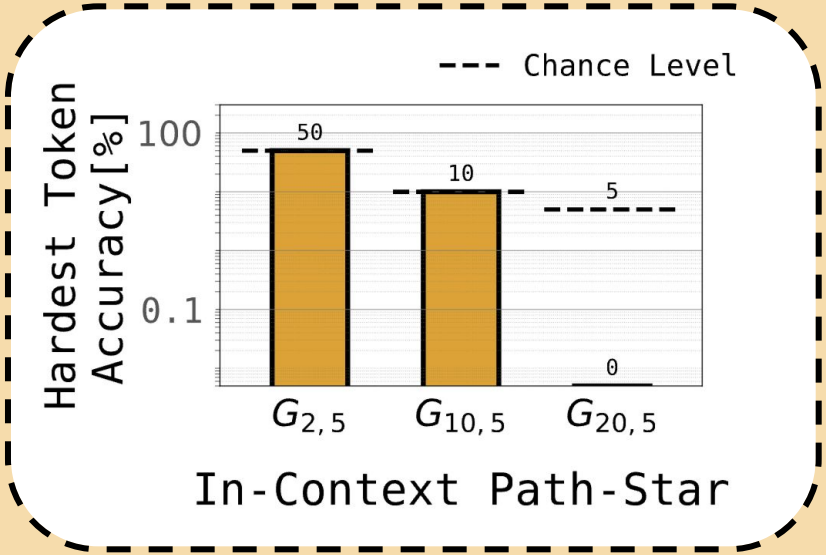
A clean, simple task: answer is the *reverse* of *unique* path from the goal.

The Pitfalls of Next-Token Prediction

Gregor Bachmann [*1] Vaishnavh Nagarajan [*2]

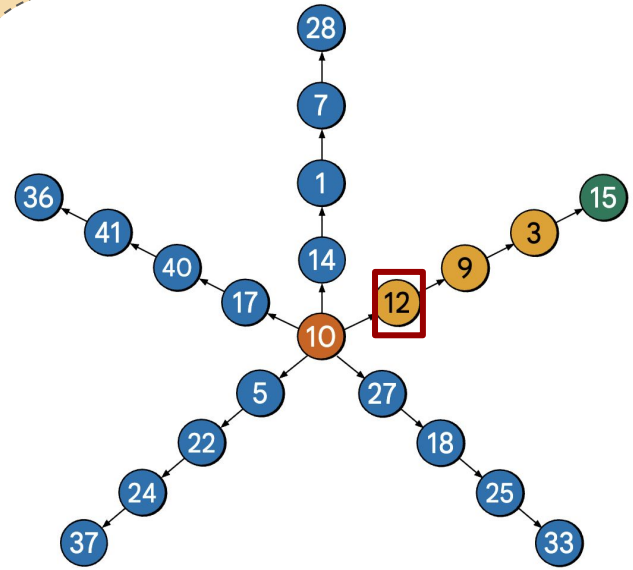
[ICML'2024

A simple yet notoriously hard implicit reasoning task



Next-token-learners fail *in-distribution* on even degree=2, length=5 graphs!

Intuition for failure




```
Prefix:
(10,14) | ( 9, 3) |
(22,24) | (10,12) |
(25,33) | (17,40) |
 ...

Start: 10
Goal : 15

Target: 10,12,9,3,15
```

(learned via shortcuts)

First token is an implicit *compositional* reasoning task.

First token =
Parent(Parent(Parent.... Parent(leaf)))

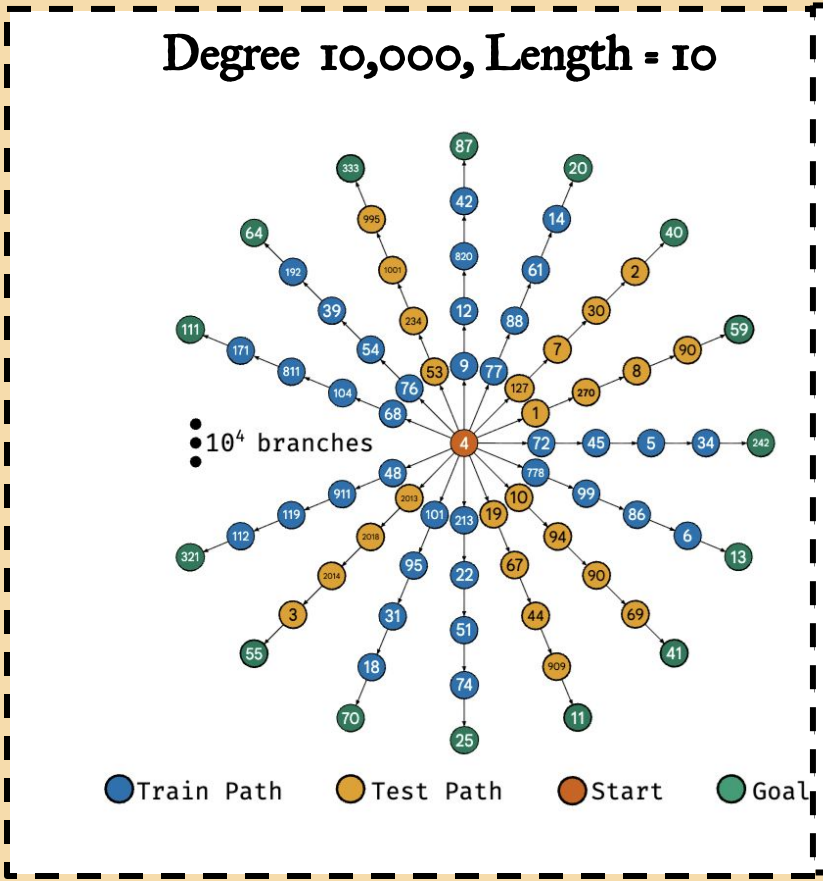
This is an (exponentially) hard task to learn — without explicit chain-of-thought or curriculum!

The Pitfalls of Next-Token Prediction

Gregor Bachmann [*1] Vaishnavh Nagarajan [*2]

I'm going to repurpose this "in-context" task as an "in-weights" task to understand memorization.

An in-weights path-star task—on a much harder graph



Edge Memorization Examples

(Edge Bigrams)

Prompt: 4	Prompt: 820
Target: 127	Target: 12
Prompt: 101	Prompt: 31
Target: 95	Target: 18
Prompt: 18	Prompt: 12
Target: 31	Target: 820

...

Path Finding Examples

(Training Paths)

Prompt: 20
Target: 4,77,88,61,14,20

Prompt: 87
Target: 4,9,12,820,42,87

Prompt: 242
Target: 4,72,45,5,34,242

Prompt: 13
Target: 4,778,99,86,6,13

(Unseen Test Paths)

Prompt: 40
Target: 4,127,7,30,2,40

Prompt: 59
Target: 4,1,270,8,90,59

...

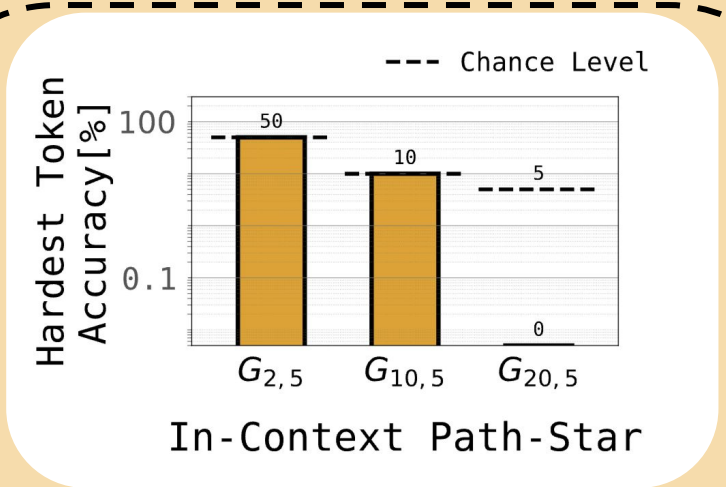
Memorize the edges of a *fixed* large graph into the weights

Train for path-finding on 50% of paths

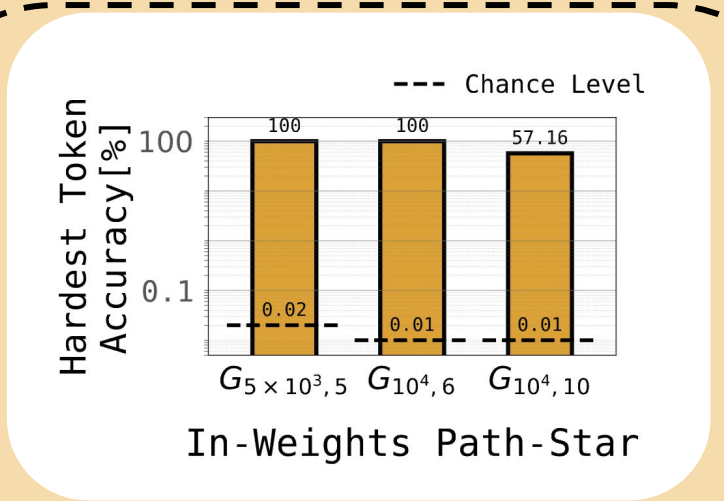
Test on *unseen* paths (only individual edges seen)

Even harder: mask the full path; only learn the first token

The key surprise:



The implicit compositional reasoning that was exponentially hard in-context for next-token prediction...



becomes trivial in-weights

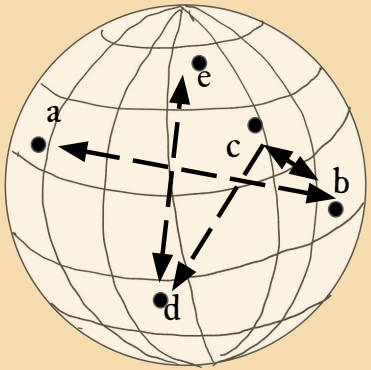
without

- explicit chain-of-thought,
- curriculum (all path lengths same)
- test-train overlap (paths are disjoint)!

Why?

Recall: *Geometric memory has synthesized global information not explicit in training data!*

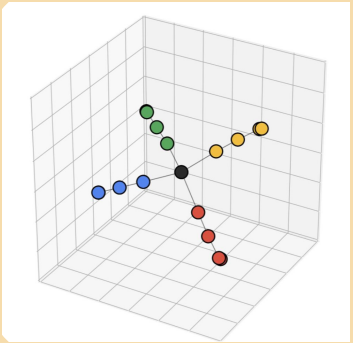
My default intuition was associative memory, where



the first token must be a hard-to-learn compositional task

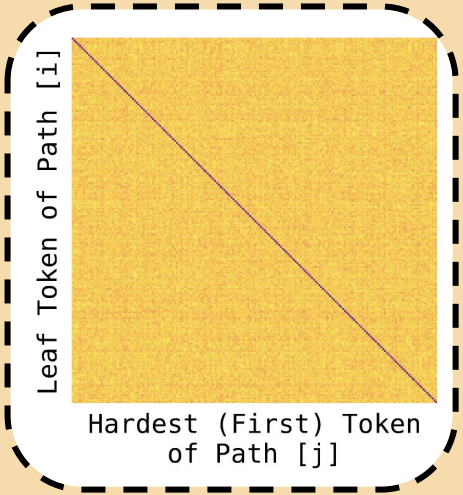
First token = $W_{assoc}(W_{assoc}(..W_{assoc}...W_{assoc}(leaf))$

But what is learned is a *global* geometry,

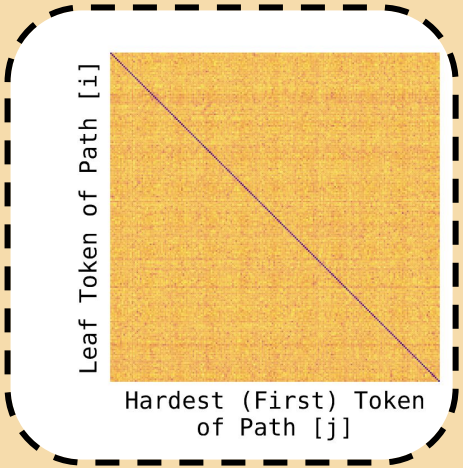


it is approximated into a learnable *"vector-based map navigation"* task

First token ≈ $[\Phi(\text{Leaf goal}) - \Phi(\text{Start})] / \text{length}$



Transformer



Mamba (so the magic isn't in the self-attention !)

Evidence of geometry: Cosine similarity of first-token vs leaf-token embeddings

Model has geometrically "visualized" a 50000 node graph!

Fragmented glimpses of (what we interpret as) "geometric memorization aids reasoning"

"Embeddings reflect some notion of distance"

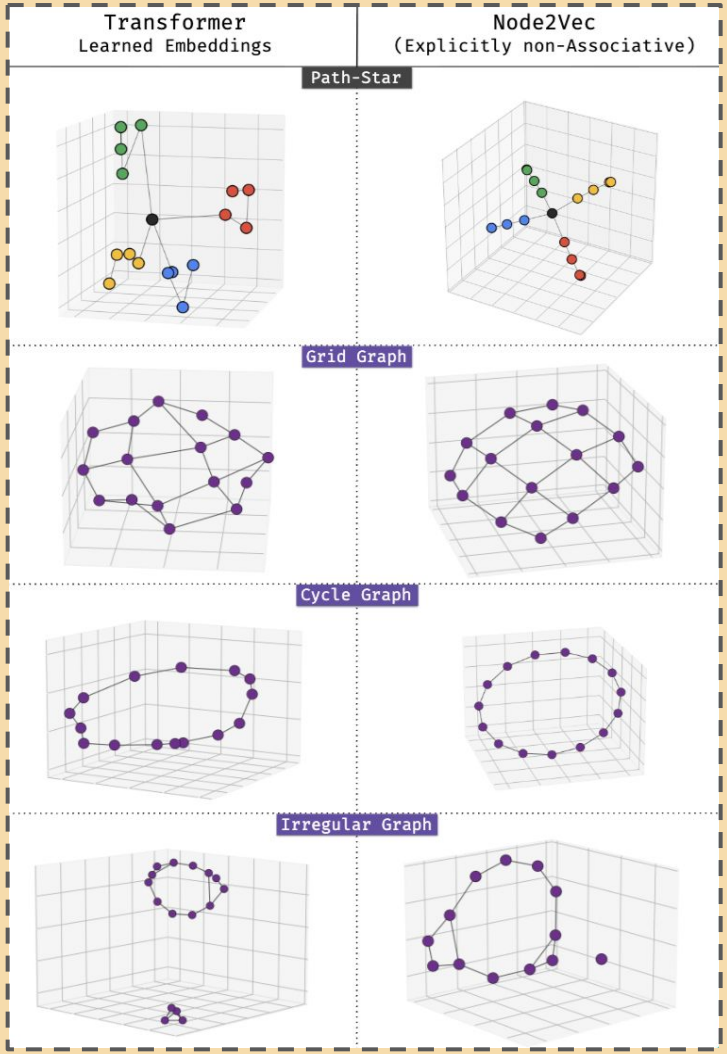
- **Towards an Understanding of Stepwise Inference in Transformers: A Synthetic Graph Navigation Model,** Mikail Khona, Maya Okawa, Jan Hula, Rahul Ramesh, Kento Nishi, Robert Dick, Ekdeep Singh Lubana, Hidenori Tanaka, 2024
- **Do LLMs dream of elephants (when told not to)? Latent concept association and associative memory in transformers,** Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, NeurIPS 2024
- **How do Transformers Learn Implicit Reasoning?** Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Weichuan Liu, Xiaoyin Che, Lei Hou, Juanzi Li, NeurIPS 2025
- **Representation Shattering in Transformers: A Synthetic Study with Knowledge Editing,** Kento Nishi, Rahul Ramesh, Maya Okawa, Mikail Khona, Hidenori Tanaka, Ekdeep Singh Lubana, ICML 2025

"Factorized storage can help 2-hop inference"

- **A mathematical theory of semantic development in deep neural networks,** Andrew M. Saxe, James L. McClelland, Surya Ganguli, 2018
- **Generalization or Hallucination? Understanding Out-of-Context Reasoning in Transformers,** Yixiao Huang, Hanlin Zhu, Tianyu Guo, Jiantao Jiao, Somayeh Sojoudi, Michael I. Jordan, Stuart Russell, Song Mei, NeurIPS 2025

Outline

- Background: associative memory
- Geometric memory
- Implications & three open questions
- Closing remarks



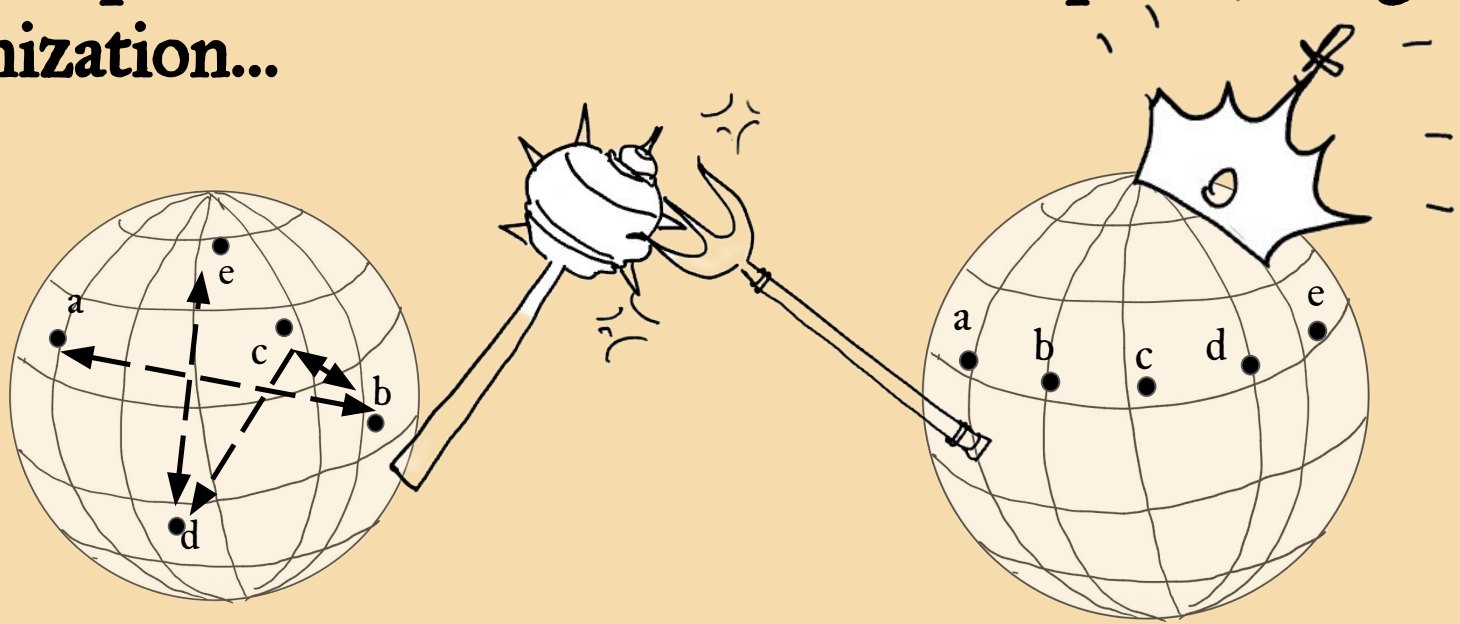
The non-associative Node2Vec model $\Phi^{T}(u)\,\Phi(v)$ gives us a much cleaner geometry.

Transformer geometry is "adulterated" with associative memory.

Has implications on Generative retrieval vs. dual encoder models

Open (practical) question 1: make Transformer memory more geometric?

The two parametric memories must compete during optimization...

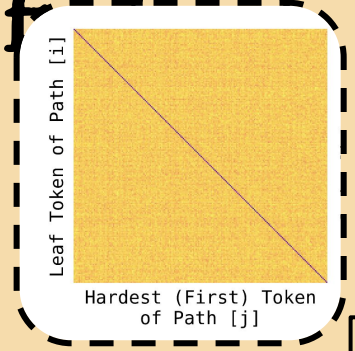


Why does the geometric prevail over the associative?

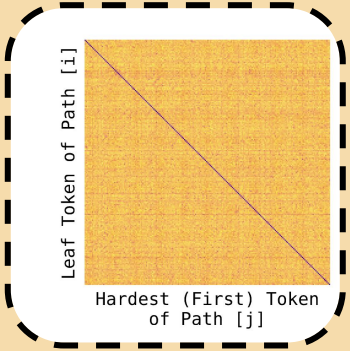
Is there some training-time pressure to prefer geometric over associative memory?

1. Maybe, the *supervision* prefers a geometry since it involves global path-f[illegible]

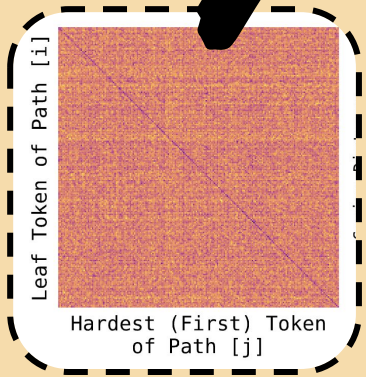
Recall: Embedding cosine similarities for Transformer

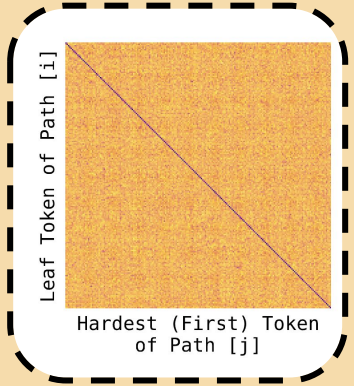


And Mamba

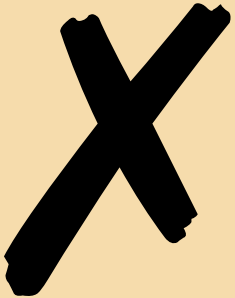


Geometry persists even for pure edge-memorization (no path-finding)





2. Maybe, *architecture/regularization explicitly* prevents associative memory on our large graphs?



Understanding Factual Recall in Transformers via Associative Memories

Eshaan Nichani*
Princeton University

Jason D. Lee
Princeton University

Alberto Bietti
Flatiron Institute

December 10, 2024

3.1 Linear Associative Memories

We first consider the case where F is a linear map $F(\boldsymbol{e}_x) = \boldsymbol{W}\boldsymbol{e}_x$.

Theorem 1. *Assume that f^* is injective. If $d^2 \gtrsim N \operatorname{poly}\log N$, then with high probability over the draw of the embeddings, there exists a $\boldsymbol{W}$ such that*

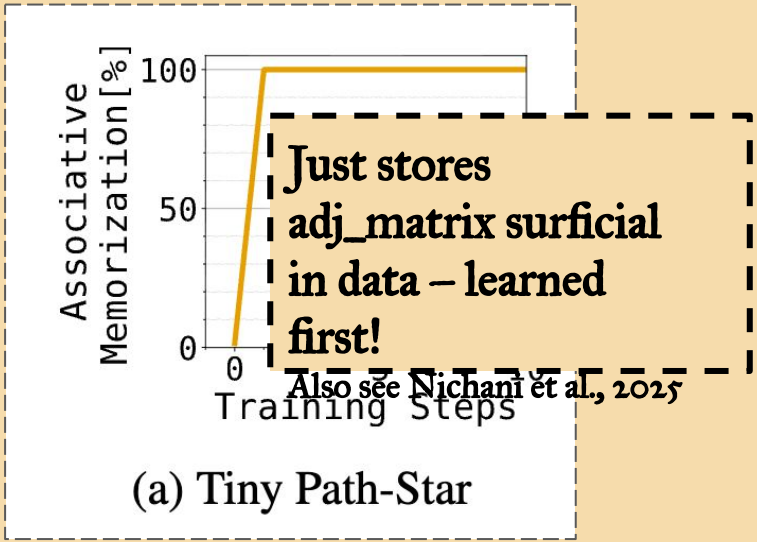
$$\arg\max_{y\in[M]} \boldsymbol{u}_y^\top \boldsymbol{W}\boldsymbol{e}_x = f^*(x) \quad \textit{for all } x \in [N]. \tag{1}$$

This capacity is obtained by the construction $\boldsymbol{W} = \sum_{x\in[N]} \boldsymbol{u}_{f^*(x)}\boldsymbol{e}_x^\top$. Furthermore, if $\boldsymbol{W}$ is restricted to be a rank m matrix, then such a $\boldsymbol{W}$ exists when $md \gtrsim N \operatorname{poly}\log N$; this construction is $\boldsymbol{W} = \sum_{x\in[N]} \boldsymbol{u}_{f^*(x)}\boldsymbol{e}_x^\top \sum_{i=1}^m \boldsymbol{v}_i\boldsymbol{v}_i^\top$, where $\boldsymbol{v}_i \in \mathbb{R}^d$ are drawn i.i.d from the standard Gaussian.

Matrix can store (width)^2 associations,
which suffices for associative memorization in our tasks! +

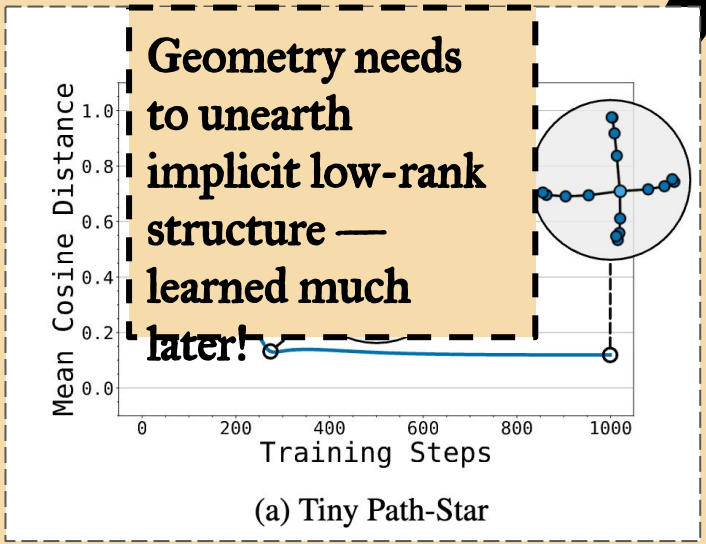
Freezing embedding layer can still fit our tasks

3. Maybe, associative lookup is *farther away* in loss landscape?



Just stores adj_matrix surficial in data – learned first!

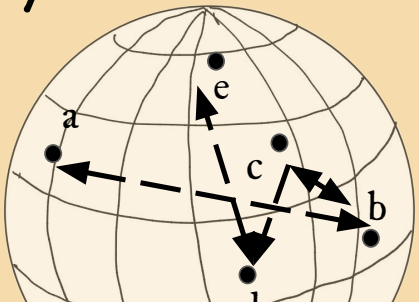
Also see Nichani et al., 2025



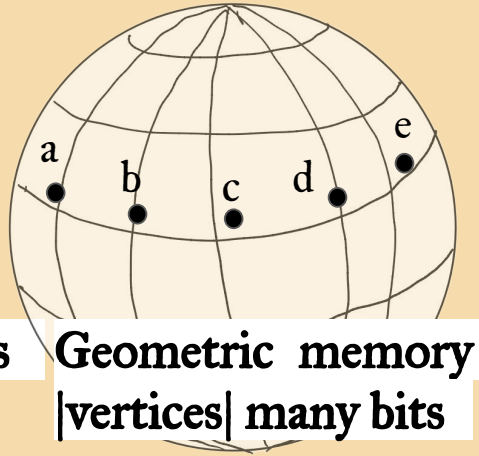
Geometry needs to unearth implicit low-rank structure — learned much later!

1. A minimal version of grokking (with just local supervision).
2. "Models memorize, then generalize/reason" is under-the-hood "models memorize associatively, then geometrically"

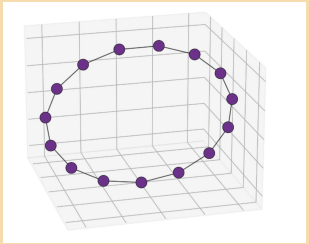
Maybe, *gradient descent implicitly* prefers geometry since it is more succinct in bit/norms?

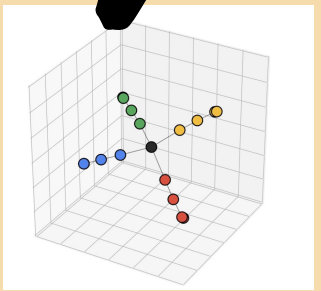


Associative memory needs |edges| many bits



Geometric memory needs |vertices| many bits





Graphs like path-star & cycle have |edges| = |vertices|

Geometric memory is not necessarily more succinct!

Is there training-time pressure to prefer geometric over associative memory?

Pressure from "global" supervision? No. Geometry is learned even on local edge-memorization

architectural pressure? No. There *is* sufficient capacity to memorize associatively!

Pressure from optimization? No. In fact: Associative memory is just 2 steps away! Geometry is farther away.

Pressure from compression? No. Counterintuitively: both mechanisms have similar bit/norm complexity for cycle graphs, path-star graphs...

Is there training-time pressure to prefer geometric over associative memory?

explicit architectural pressure?

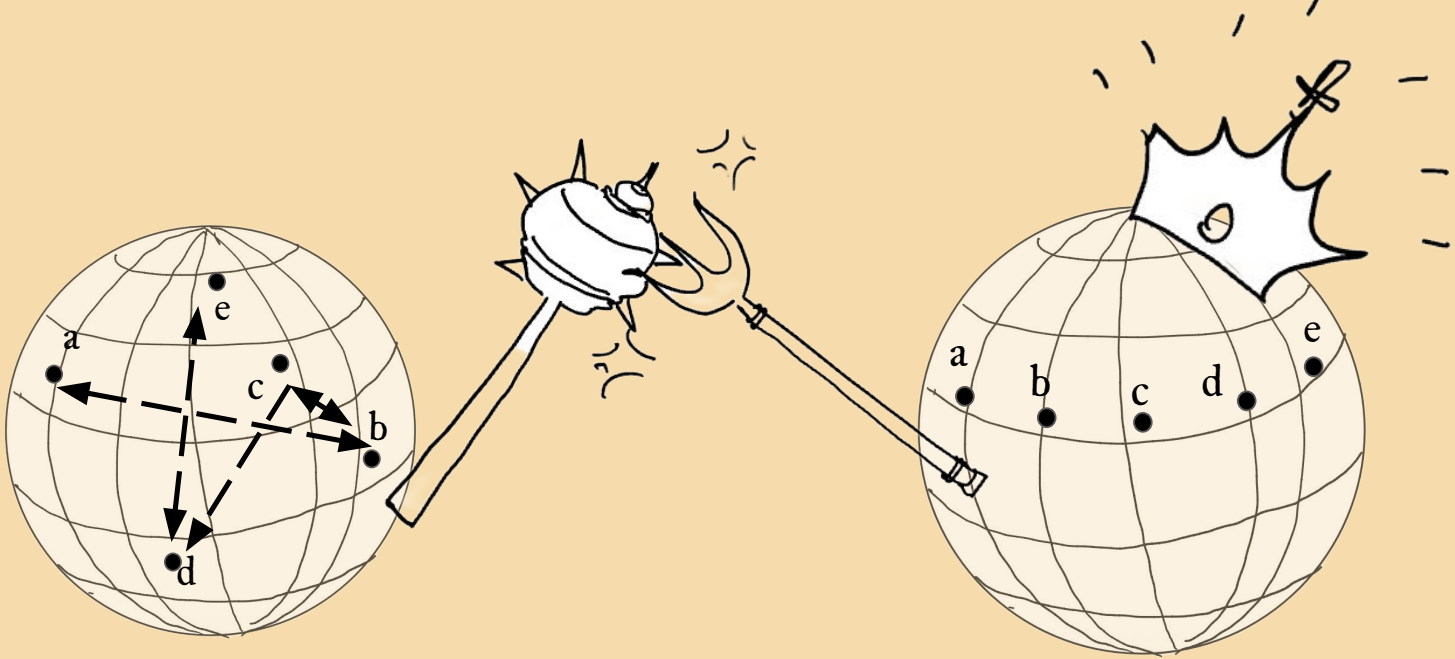
Matrix of width = 384 can store 384^2 associations, which suffices for associative memorization in our tasks.

implicit capacity pressure from GD?

Description length of path-star & cycle graphs are equal in geometric and associative memories since |vertices| = |edges|

supervisory pressure?

A global geometry arises even without path-finding supervision. *Local* edge memorization suffices.



Open (theoretical) question II:

Why does a deep sequence model trained on local edges, synthesize a global geometry without any pressure to do so?

Sequence Modeling

UNDERSTANDING DEEP ~~LEARNING~~ REQUIRES RE-THINKING ~~GENERALIZATION~~

Memorization

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

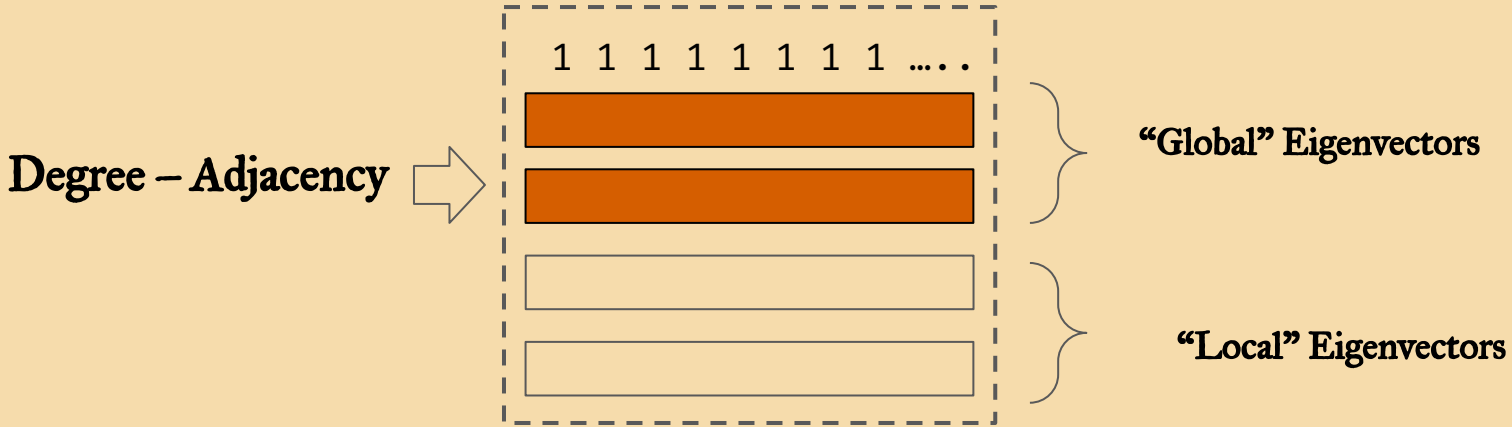
Oriol Vinyals
Google DeepMind
vinyals@google.com

(this was our originally proposed title)

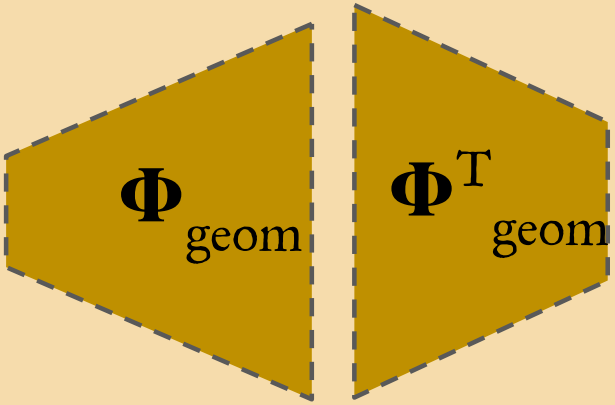
Open theoretical question: The memorization puzzle

Well-known fact: Gradient descent on Node2Vec induces a s *pectral bias*

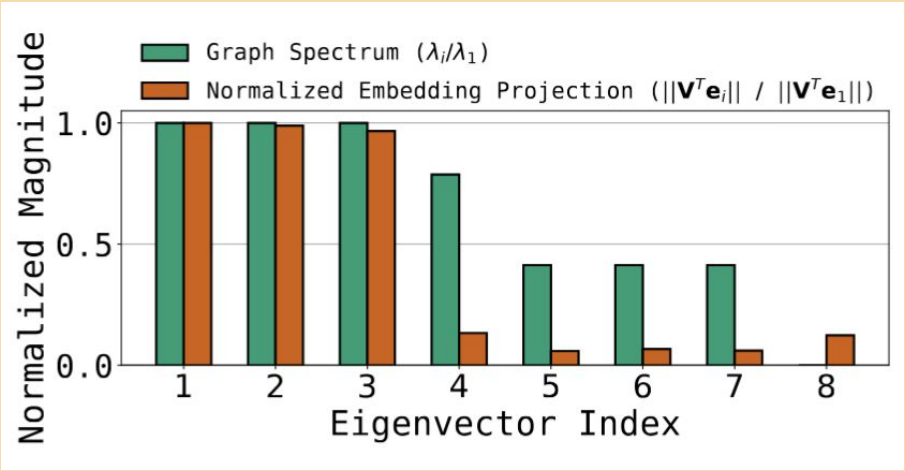
Insight: The “ *global* ” geometry comes from eigenvectors of the Graph laplacian.



Open question: Extend this argument to softmax loss + Seq2Seq models



embedder unembedder



A simpler open question III: *Wide* two-layer models *under cross-entropy loss* implicitly learn the low-rank factorization/geometry.

Why? What is the precise solution? (We believe it finds a zero-gradient solution.)

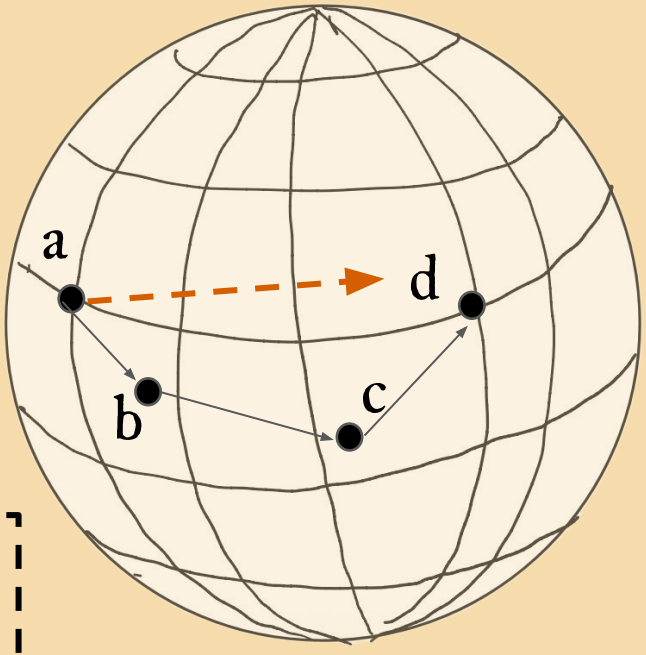
Outline

- Background: associative memory
- Geometric memory
- Implications & open questions
- Closing remarks

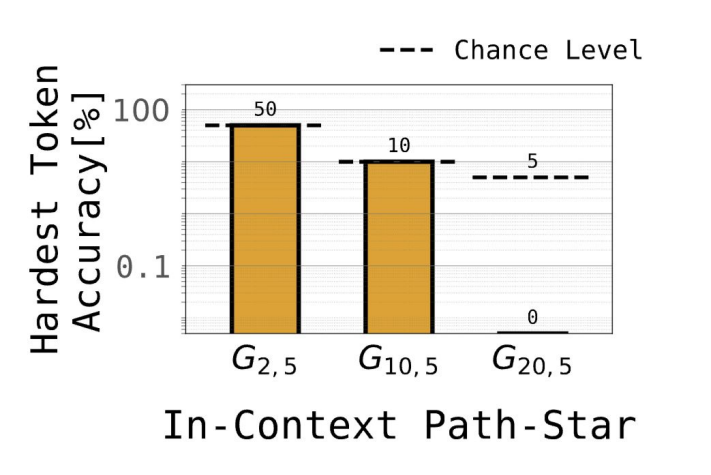
Roll the dice & look before you leap:
Going beyond the creative limits of next-token prediction

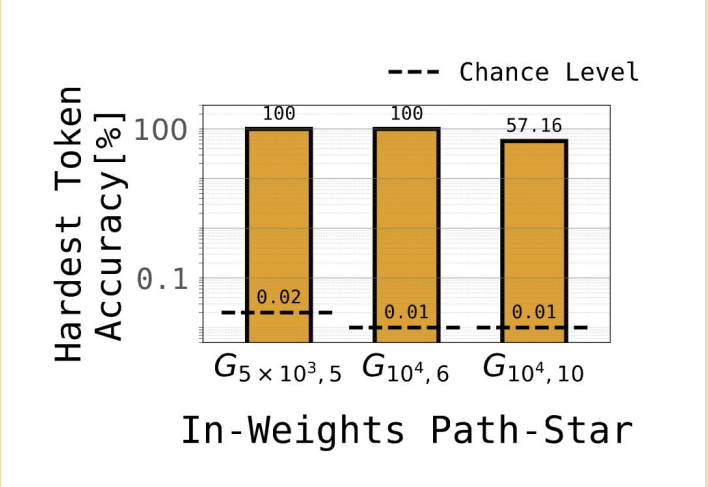
Vaishnavh Nagarajan [*1] Chen Henry Wu [*2] Charles Ding [2] Aditi Raghunathan [2]

The first class of tasks involves *combinational creativity*: drawing novel connections in knowledge, like in research, wordplay or drawing analogies (see Fig 1 for task descrip-



Implication: Discovery/creativity over knowledge becomes possible with geometric memory.





Implication: Parametric memory *integrates* knowledge in a way that contextual memory does not.

Caveats, remarks, limitations…

1. The jump from symbolic → language tasks can be murky (bridging this is hard work)
   a. E.g., Transformer implicit reasoning fails in natural language likely because memory is not geometric
   b. Open question: How do you make it more geometric in natural language?

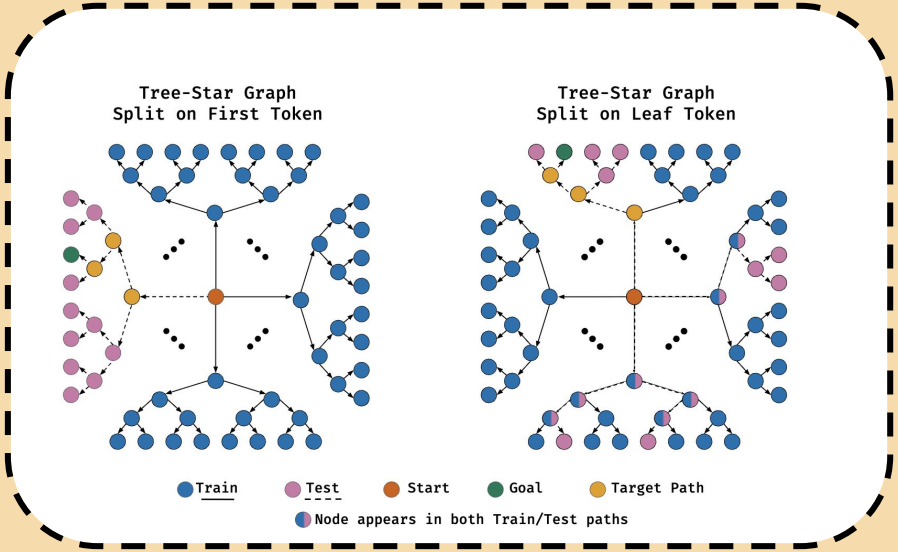
2. Path-star is not a benchmark. Must only be used as a *quick preliminary* litmus test & inspire algorithms.

Caveats, remarks, limitations...

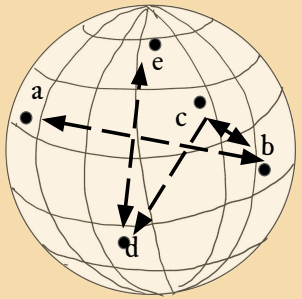
3. Other graph topologies may be much harder

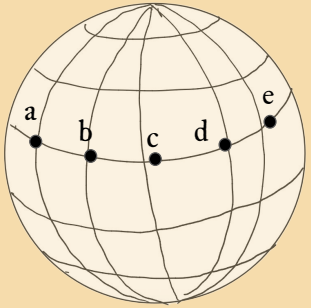
4. Associative memory can be desirable for accurate retrieval.

5. Less of implicit “reasoning” and more so, approximate pattern matching?



Revisiting (and challenging) unstated "associative" assumptions





I. Storage capacity / expressivity

[From "On the optimal memorization capacity of Transformers" by Kajitsuka and Sato]

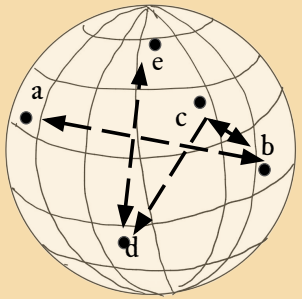
Table 1: Comparisons between our results and related work regarding the memorization capacity of Transformers. The variable ω in the bounds presented by Madden et al. (2024) represents the vocabulary size, or the number of distinct word vectors that appear in input sequences.

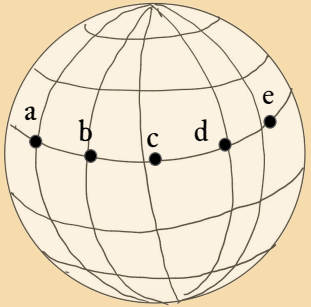
Paper	Setting	Input	#layers	Upper bound	Lower bound
Kim et al. (2023)	seq-to-seq	token-wise (r, δ)-separated	$\tilde{O}(n + \sqrt{nN})$	$\tilde{O}(n + \sqrt{nN})$	-
Mahdavi et al. (2023)	next-token	linearly independent	1	$O(d^2N/n)$	-
Kajitsuka & Sato (2023)	seq-to-seq	token-wise (r, δ)-separated	1		
Madden et al. (2024)	next-token	with positional encoding	1		
Ours	next-token	token-wise (r, δ)-separated	$\tilde{O}(\sqrt{N})$		
	seq-to-seq	token-wise (r, δ)-separated	$\tilde{O}(\sqrt{nN})$		

3.3 Bit complexity

In this paper, we consider not only the number of parameters but also the number of bits required to represent the model. Specifically, we adopt the definition of **bit complexity** proposed by Vardi et al. (2022). According to this definition, the bit complexity of a parameter is defined as the number of bits needed to represent that parameter. The bit complexity of a model is then defined as the maximum bit complexity among its individual parameters. It is important to note that by multiplying the bit complexity of the model by the number of parameters, we can estimate the total number of bits required to represent the entire model.

Revisiting (and challenging) unstated "associative" assumptions





2. Scaling laws

SCALING LAWS FOR ASSOCIATIVE MEMORIES

Vivien Cabannes FAIR, Meta
Elvis Dohmatob FAIR, Meta
Alberto Bietti Flatiron Institute

3. Other theoretical analyses

Towards a Theoretical Understanding of the 'Reversal Curse' via Training Dynamics

E.2.2 Near orthogonal embeddings

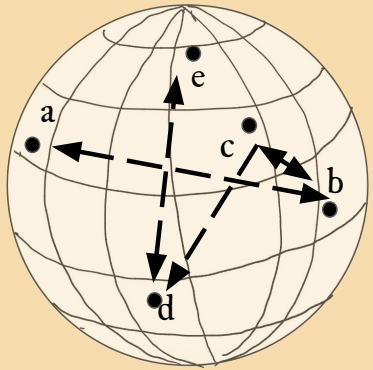
Note that in Section 3, our analysis relies on the fact that embeddings of different tokens are nearly orthogonal, and in Section 4, the embeddings are effectively one-hot. In Figure 11, We show that in practice, even if the embedding dimension is much smaller than the vocabulary size, the near orthogonality condition still holds.

Understanding Finetuning for Factual Knowledge Extraction

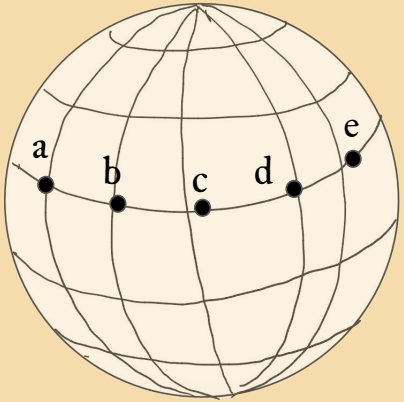
Gaurav Ghosal [1] Tatsunori Hashimoto [2] Aditi Raghunathan [1]

Simplified Model We analyze a one-layer, single-headed transformer model with fixed, orthogonal token embeddings (denoted as $\phi(t)$ for token $t \in \mathcal{T}$). Our model has two learnable parameters: the value matrix, W_V, and the key-query

Revisiting (and challenging) unstated "associative" assumptions



3. Unlearning / editing
    a. Much harder with geometric memory



4. New notions of "geometric" learning-theoretic complexity?

4. AI systems are so complex.

As a complement to large-scale black-box empirics and "hard" theory,

we need simple examples and "soft" theory with vivid intuition

- to guide and inspire new solutions &
- formulate new questions.

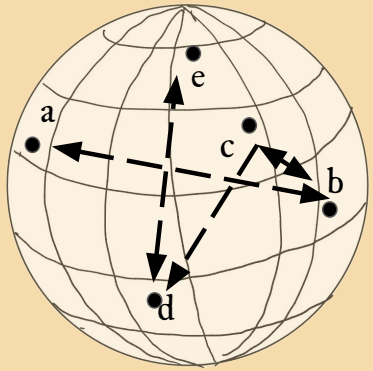
Summary

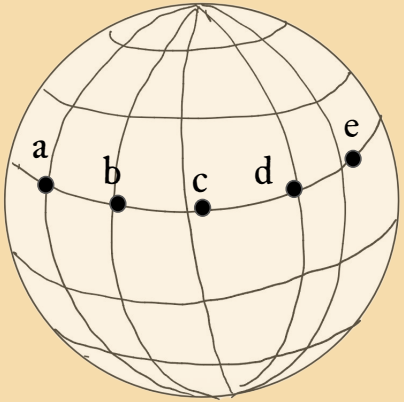
1. Deep sequence models memorize geometrically, not associatively.

2. We don't understand why.

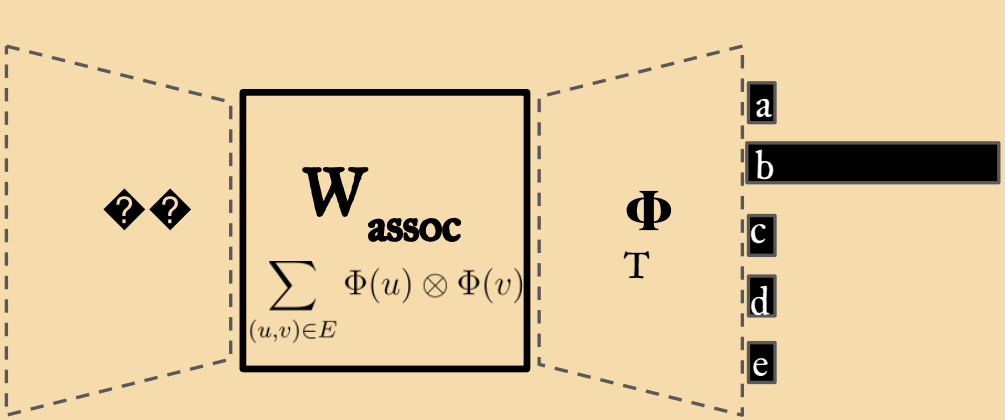
3. Enhancing this geometry can open up new forms of reasoning and discovery.

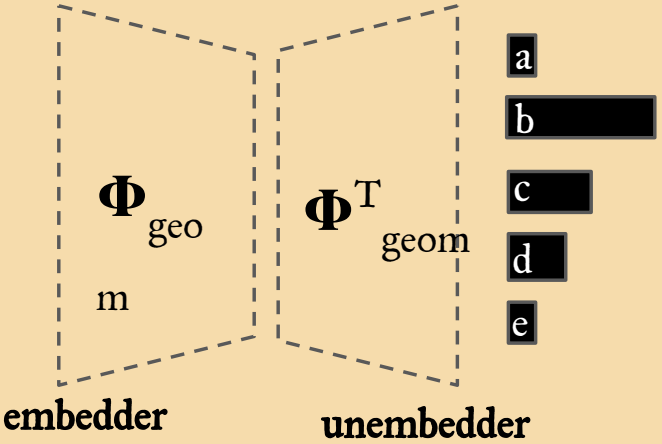
4. Go back and rethink many analyses!

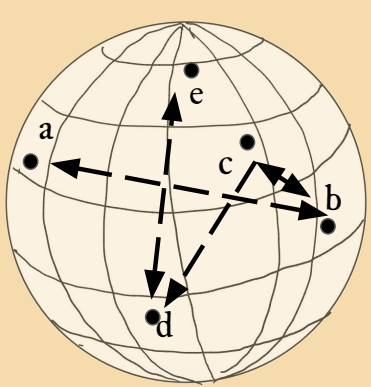


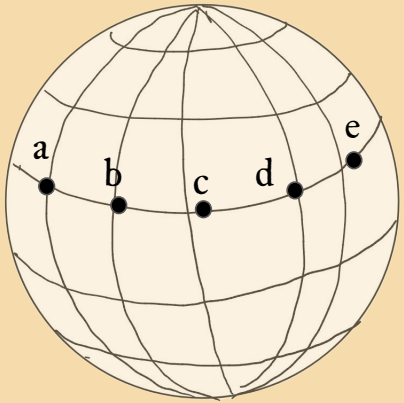


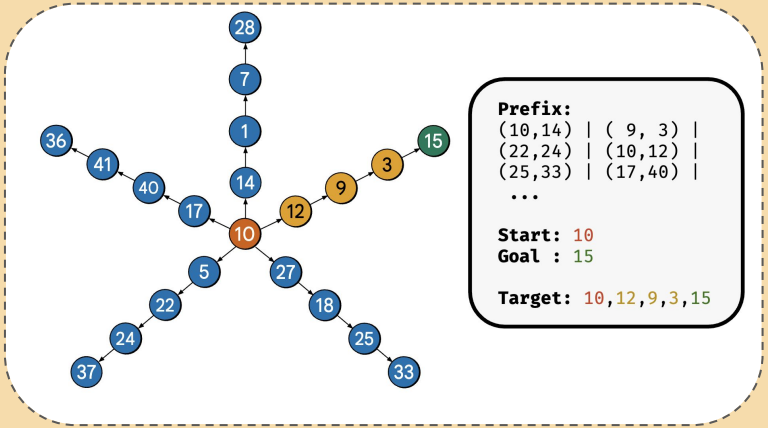
Thank you!


W assoc
$\sum_{(u,v)\in E} \Phi(u) \otimes \Phi(v)$
Φ
T
a
b
c
d
e


Φ geo
m
Φ T geom
a
b
c
d
e
embedder
unembedder

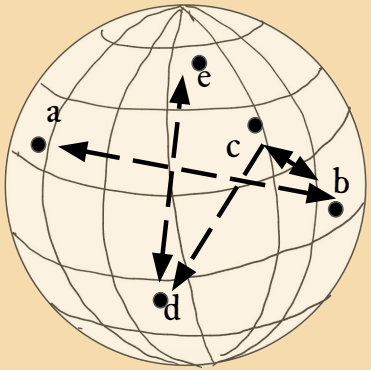

a
b
c
d
e


a
b
c
d
e


28
7
1
14
10
36
41
40
17
5
22
24
37
12
9
3
15
27
18
25
33
Prefix:
(10,14) | (9, 3) |
(22,24) | (10,12) |
(25,33) | (17,40) |
...
Start: 10
Goal : 15
Target: 10,12,9,3,15

Recall: *Geometric memory has synthesized* global *information not explicit in training data!*

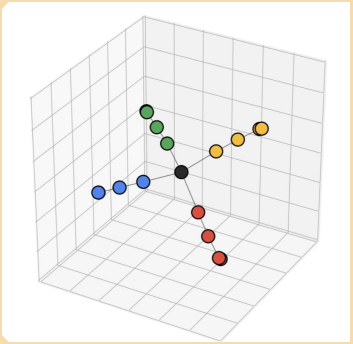
In default intuition of associative memory,



the first token must be a hard-to-learn compositional task

First token = $W_{assoc}(W_{assoc}(..W_{assoc}...W_{assoc}(leaf)))$

But what is learned is a global geometry,



where it is approximated into a learnable navigation task

First token ≈

$$[\Phi(\text{Leaf goal}) - \Phi(\text{Start})] / \text{length}$$

Geometric memory: an *alternative* way to fit our data

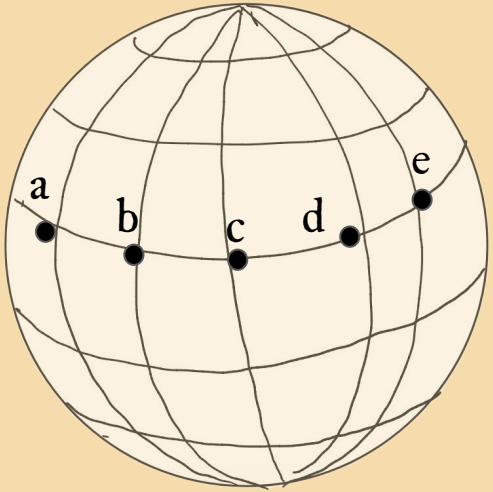
Input → Output:

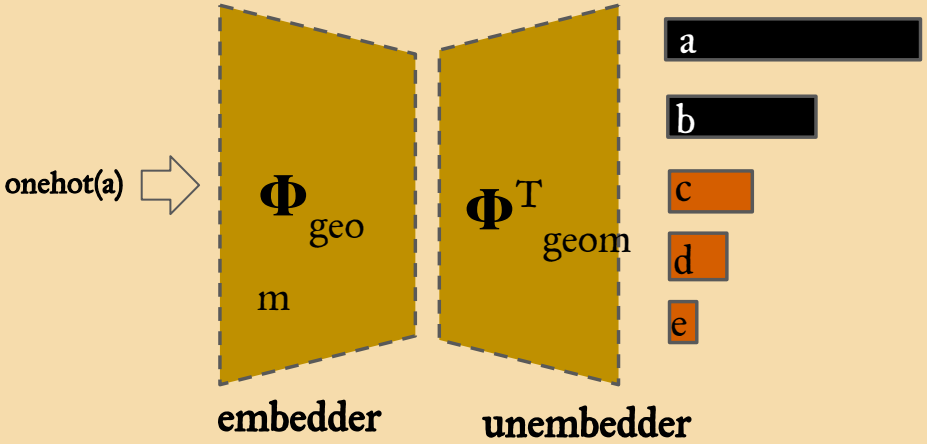
a→b, b→a

b→c, c→b

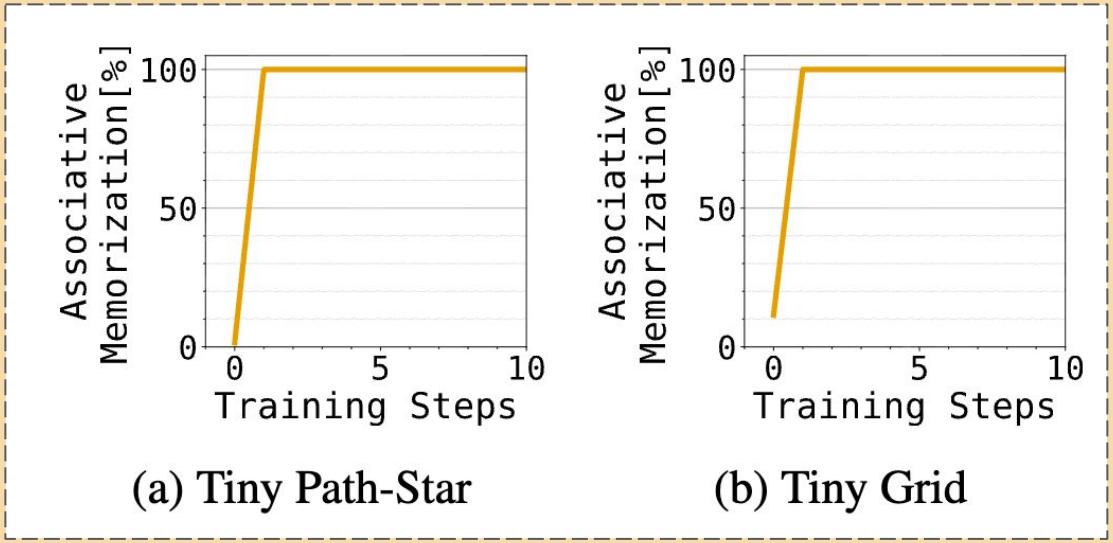
c→d, d→c

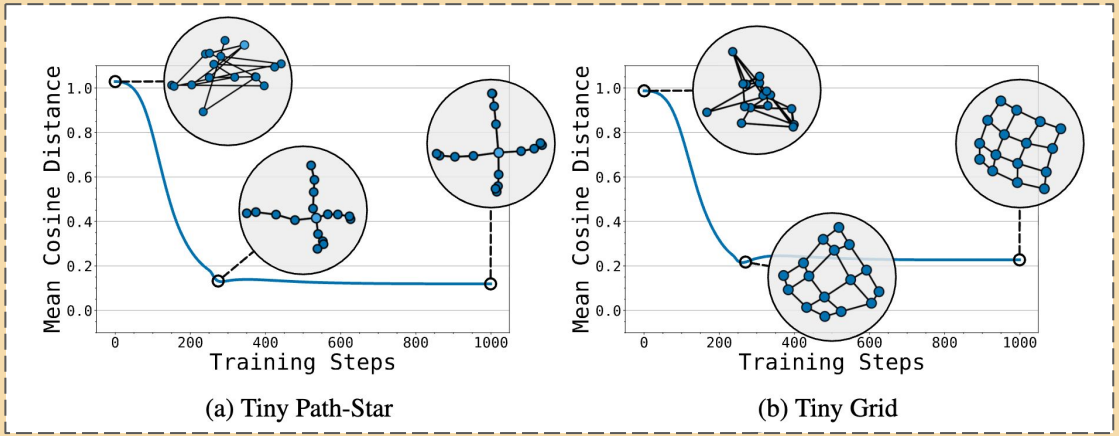
d→e , e→d





Highly organized
embeddings Φ_{geom}


Associative Memorization [%]
100
50
0
0
5
10
Training Steps
(a) Tiny Path-Star
Associative Memorization [%]
100
50
0
0
5
10
Training Steps
(b) Tiny Grid


Mean Cosine Distance
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Training Steps
(a) Tiny Path-Star
Mean Cosine Distance
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Training Steps
(b) Tiny Grid