Memorization vs. Reasoning: What the vec happens inside LLMs (when solving math questions)?

Marthe Ballon | Junior Colloquium UCL 27/11/2024 | Vincent Ginis

Design by Seorin Kim

Part 1: What are Large Language Models? (the basics)

What are Large Language Models?

Imagine your company hired a new colleague who is...

Smart

Very eloquent

Optimistic

Kind of a sycophant

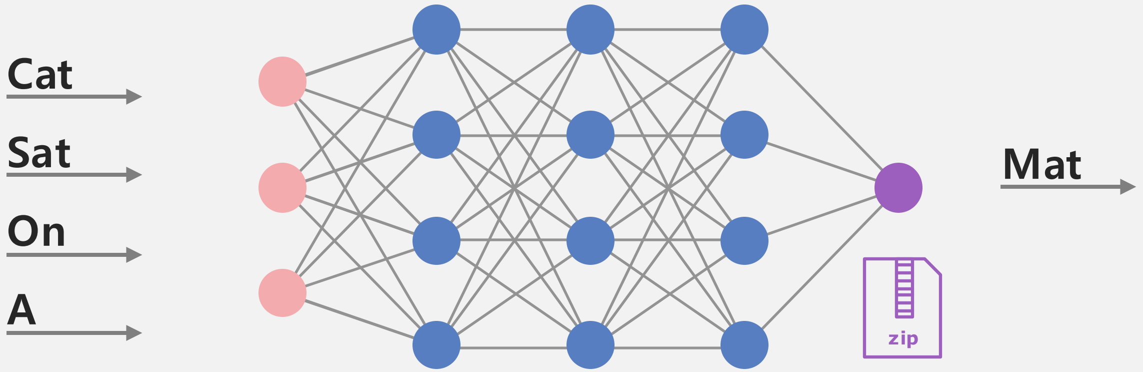
Sometimes a bit drunk

Versed in various scientific disciplines

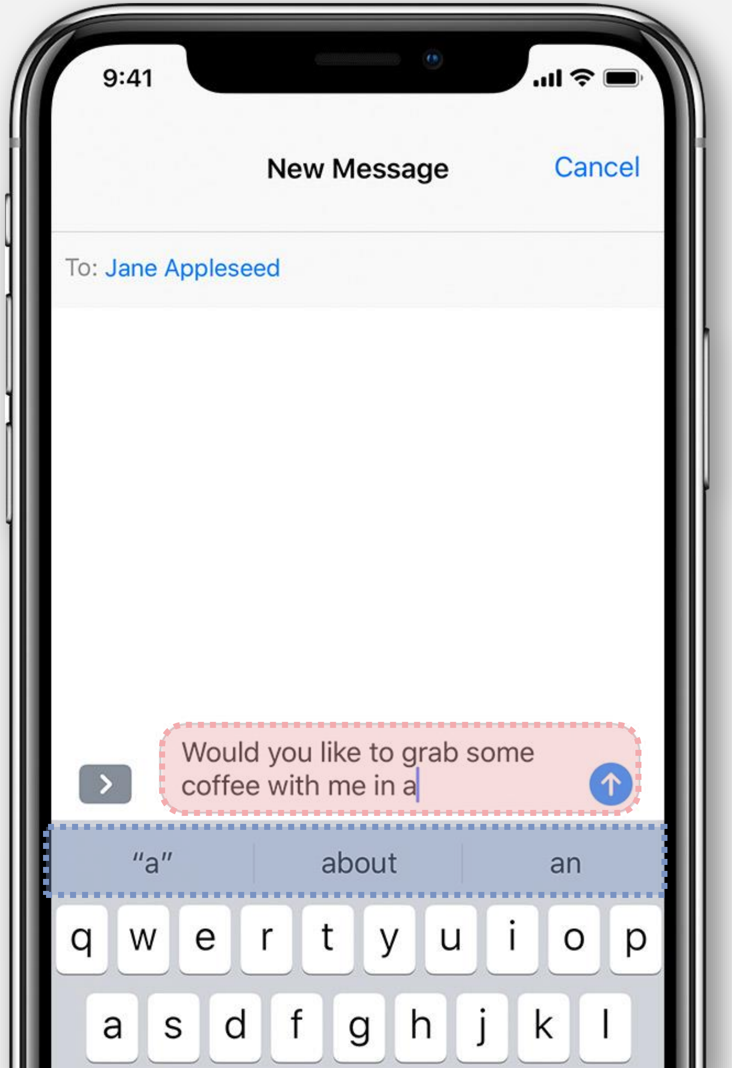
What are Large Language Models?

Generative Pre-trained Transformer (GPT)

The model predicts the next token based on a given context/prompt.

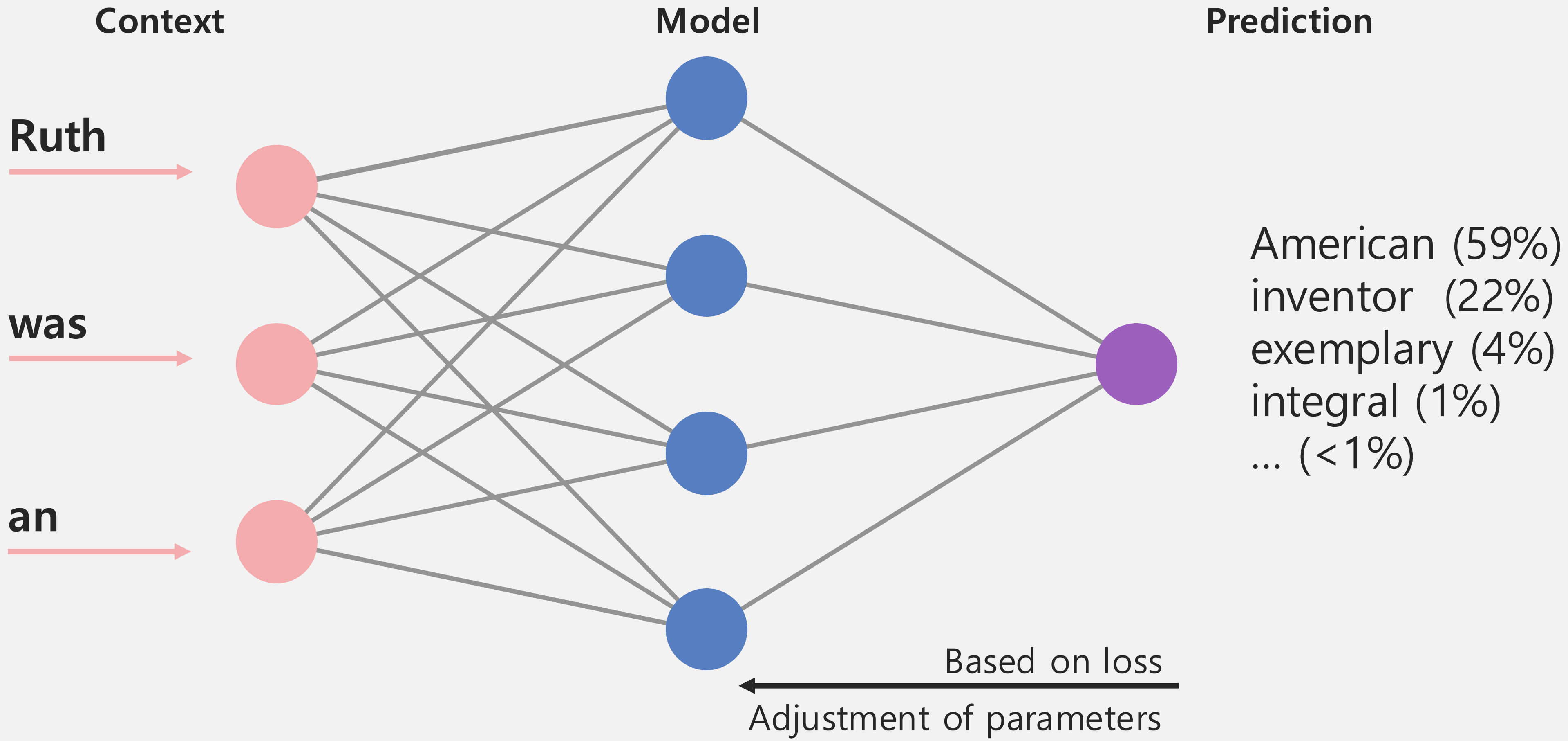


Context / Prompt   Model   Next token prediction



What are Large Language Models?

Pre-training is compressing information through prediction



What are Large Language Models?

GPTs are pre-trained on internet data

Ruth Handler

Ruth Marianna Handler (née Mosko; November 4, 1916 – April 27, 2002) was an American businesswoman and inventor. She is best known for inventing the Barbie doll in 1959,[2] and being co-founder of toy manufacturer Mattel with her husband Elliot, as well as serving as the company's first president from 1945 to 1975.[3]

The Handlers were forced to resign from Mattel in 1975 after the Securities and Exchange Commission investigated the company for falsifying financial documents.[3][4]

- Wikipedia

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
Sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

What are Large Language Models?

Fine-tuning base models into assistants

Going from GPT to ChatGPT in two steps: Step 1

Instruction tuning by training on curated data with a question-answer template.

<USER>

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

<ASSISTANT>

"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions…

Less focus on knowledge compression and more on how to extract knowledge via prompts.

What are Large Language Models?

Fine-tuning base models into assistants

Going from GPT to ChatGPT in two steps: Step 2

Reinforcement Learning through Human Feedback (RLHF).

Generate Multiple Answers

for each question and rank them
(human feedback)

Train the Model

to return the preferred answers
(via reinforcement learning)

What is new with OpenAI o1?

ChatGPT o1 talks to itself until it’s satisfied with its answer

oyfjdnisdr rtqwainr acxz mynzbhhx -> Think step by step

Use the example above to decode:

oyekaijzdf aaptcg suaokybhai ouow aqht mynznvaatzacdfoulxxz

Thought for 19 seconds

Cracking the code

OK, let me see. The user provides a cipher method, asking to decode another message by averaging letter pairs' numerical positions. This approach needs clarity for precise decoding.

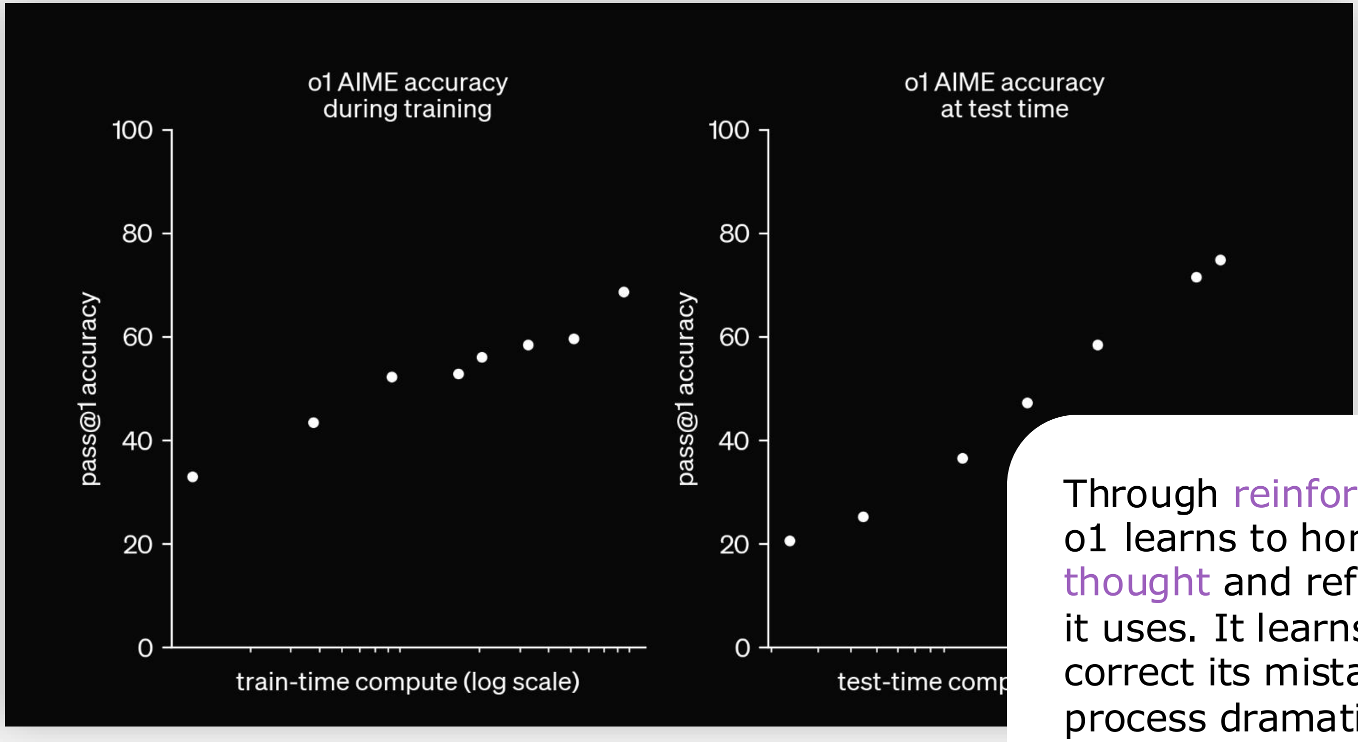
Analyzing encryption

OK, let me see. I'm examining the encrypted message "oyekaijzdf aaptcg suaokybhai ouow aqht mynznvaatzacdfoulxxz" to decode it effectively, leaning towards a letter pair-based method.

Decoding the puzzle

What is new with OpenAI o1?

Increasing compute at generation time trains models to "think"



Through reinforcement learning, o1 learns to hone its chain of thought and refine the strategies it uses. It learns to recognize and correct its mistakes. […] This process dramatically improves the model's ability to reason.

Four misconceptions about LLMs

1

Reproducing a blurry copy of the internet

2

Only capable of dull, repetitive tasks

3

Bias in – bias out

4

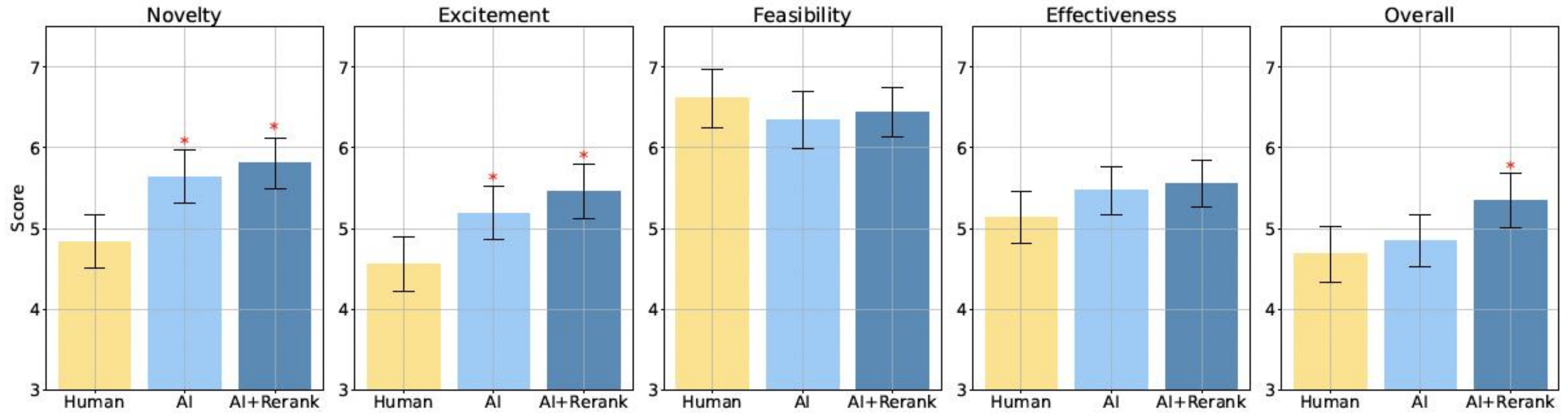
It only works with language

Why work with LLMs?

LLMs and Creative Work

Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers



Why work with LLMs?

LLMs and Creative Work

Article

Advancing mathematics by guiding human intuition with AI

https://doi.org/10.1038/s41586-021-04086-x

Received: 10 July 2021

Accepted: 30 September 2021

Published online: 1 December 2021

Open access

Check for updates

Alex Davies[1✉], Petar Veličković[1], Lars Buesing[1], Sam Blackwell[1], Daniel Zheng[1], Nenad Tomašev[1], Richard Tanburn[1], Peter Battaglia[1], Charles Blundell[1], András Juhász[2], Marc Lackenby[2], Geordie Williamson[3], Demis Hassabis[1] & Pushmeet Kohli[1✉]

The practice of mathematics involves discovering patterns and using these to formulate and prove conjectures, resulting in theorems. Since the 1960s, mathematicians have used computers to assist in the discovery of patterns and formulation of conjectures[1], most famously in the Birch and Swinnerton-Dyer conjecture[2], a Millennium Prize Problem[3]. Here we provide examples of new fundamental results in pure mathematics that have been discovered with the assistance of machine learning—demonstrating a method by which machine learning can aid mathematicians in discovering new conjectures and theorems. We propose a process of using machine learning to discover potential patterns and relations between mathematical objects, understanding them with attribution techniques and using these observations to guide intuition and propose conjectures. We outline this machine-learning-guided framework and demonstrate its successful application to current research questions in distinct areas of pure mathematics, in each case

Why work with LLMs?

LLMs and Creative Work

Conjecture. There exist constants c_1 and c_2 such that, for every hyperknot K, $|2\sigma(K) - \mathrm{slope}(K)| < c_1 \mathrm{vol}(K) + c_2$.

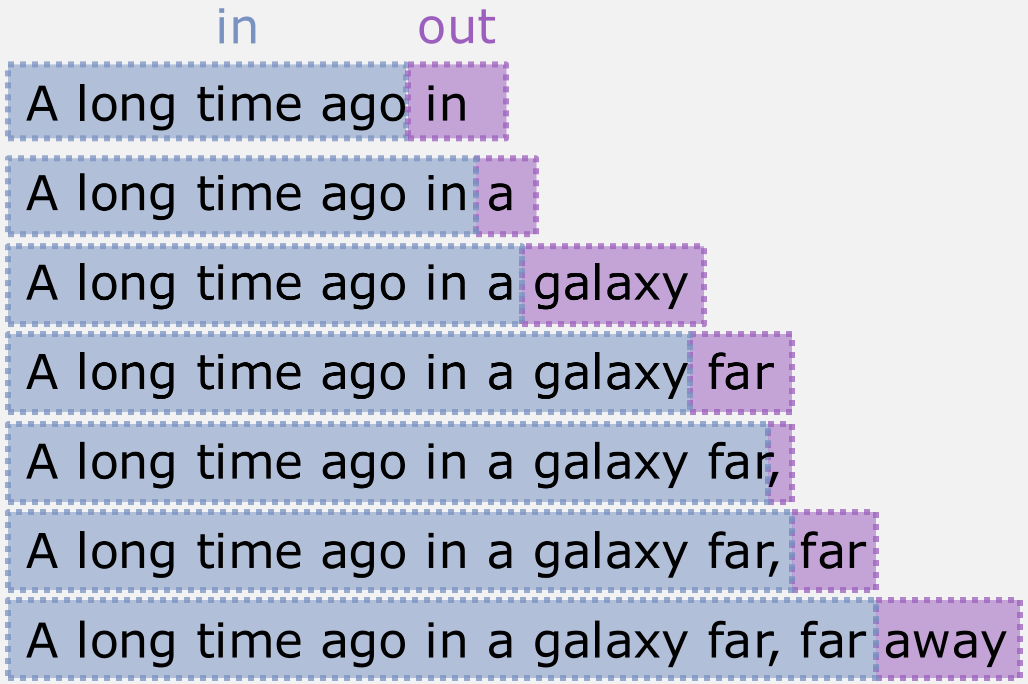
z: **Knot**	*X*(*z*): **Geometric invariants**				*Y*(*z*): **Algebraic invariants**		
	Volume	Chern–Simons	Meridional translation	...	Signature	Jones polynomial	...
	2.0299	0	i	...	0	$t^{-2} - t^{-1} + 1 - t + t^2$	...
	2.8281	–0.1532	$0.7381 + 0.8831i$	...	–2	$t - t^2 + 2t^3 - t^4 + t^5 - t^6$	...
	3.1640	0.1560	$-0.7237 + 1.0160i$	...	0	$t^{-2} - t^{-1} + 2 - 2t + t^2 - t^3 + t^4$	...

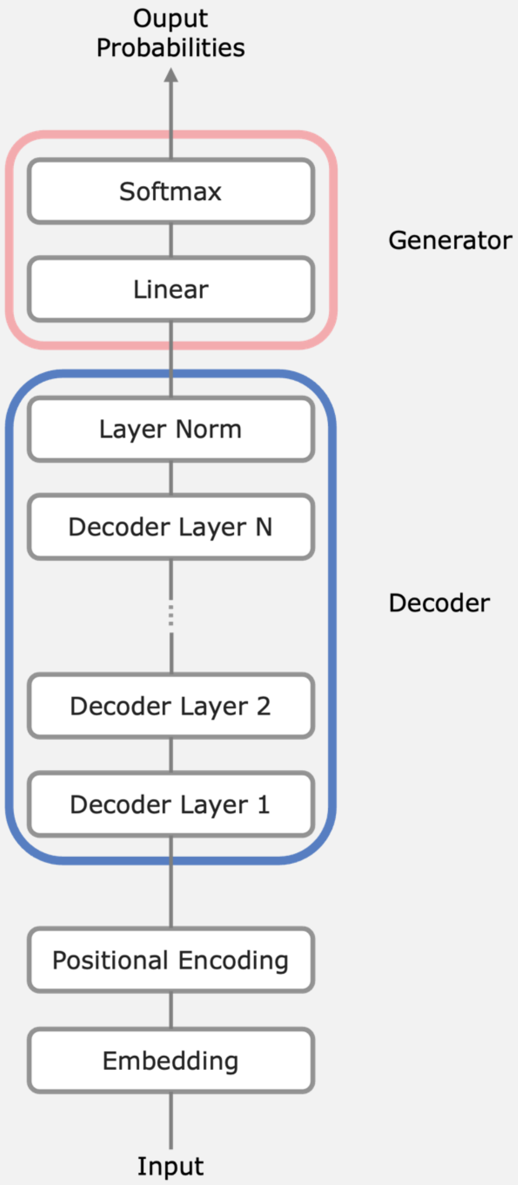
A big scientific debate

Memorization vs. Reasoning

The internal machinery of LLMs

LLMs convert text tokens to vector embeddings… and back





Part 2: How can we leverage vector embeddings?

How can we leverage embeddings?

#1: The Linear Representation Hypothesis and the Geometry of LLMs (K. Park 2024)

> "High-level concepts are represented linearly in the representation space of a model"

English/French

Female/Male

Upper case/Lower case

Past Tense/Present Tense

How can we leverage embeddings?

LLMs keep track of an embedding and unembedding space

Context

'He is the' X ⟶ Y

Next Word

'king',
'man',
'phd',... $\in V$

Embedding

$x \longmapsto \lambda(x) \in \Lambda \cong \mathbb{R}^d$ ⟶

Softmax

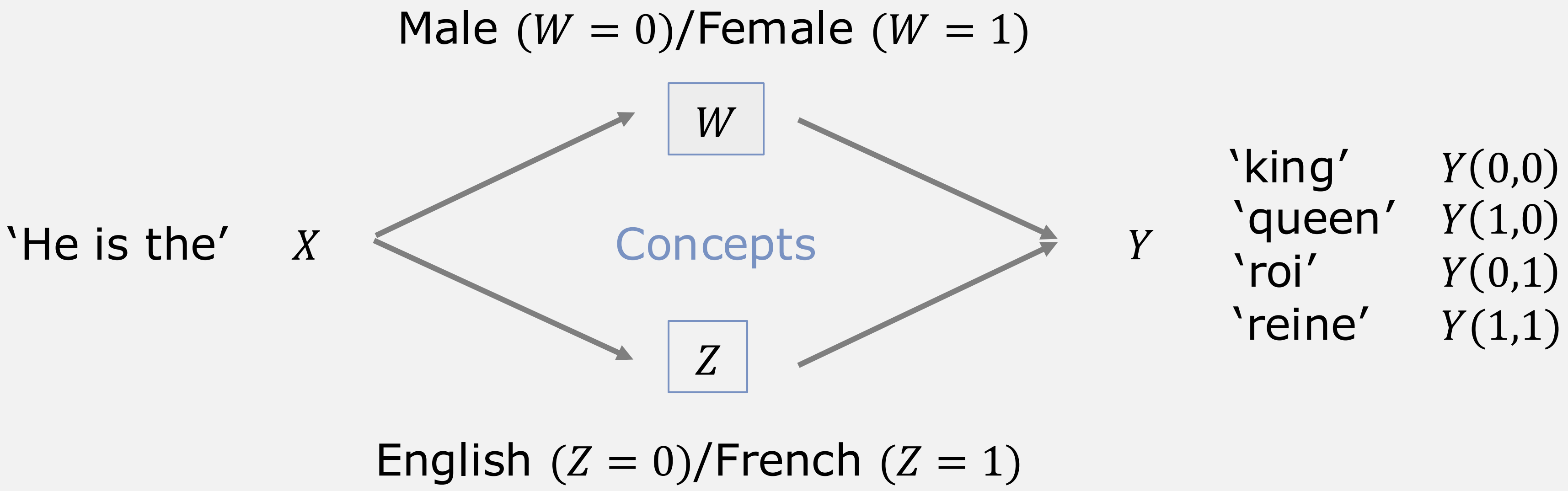
$p(x|y) \propto \exp \lambda(x)^T \gamma(y)$

Unembedding

⟵ $y \longmapsto \gamma(y) \in \Gamma \cong \mathbb{R}^d$

How can we leverage embeddings?

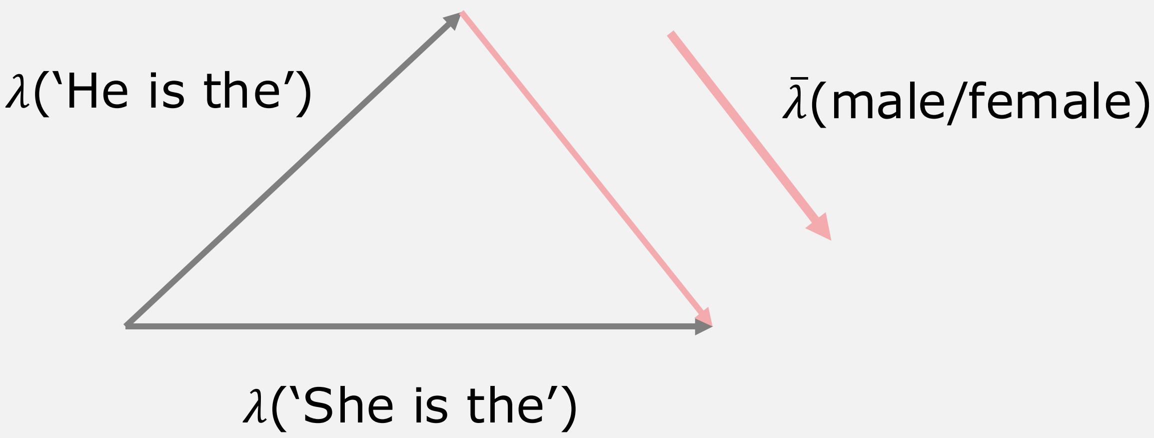
A concept is a factor of variation that can be changed in isolation



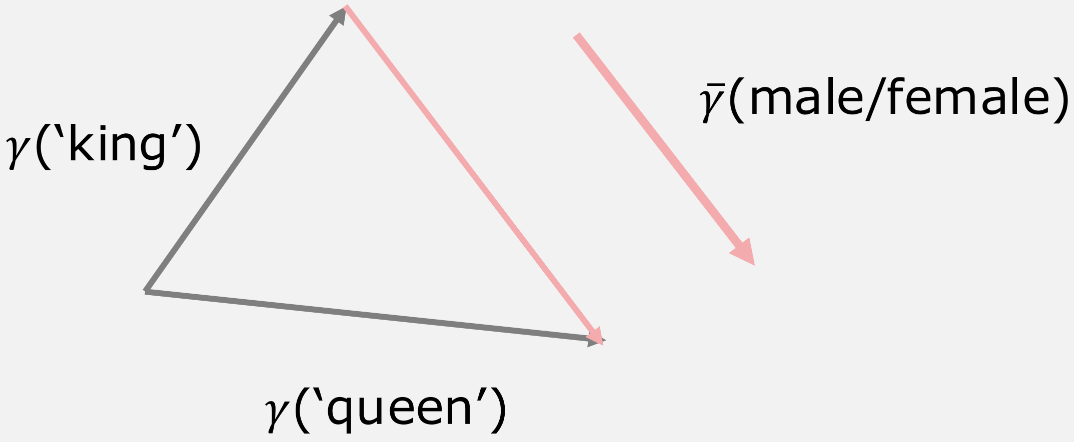
How can we leverage embeddings?

The linear representation of a concept W is the direction determined by pairs $(Y(0), Y(1))$

Embedding Space

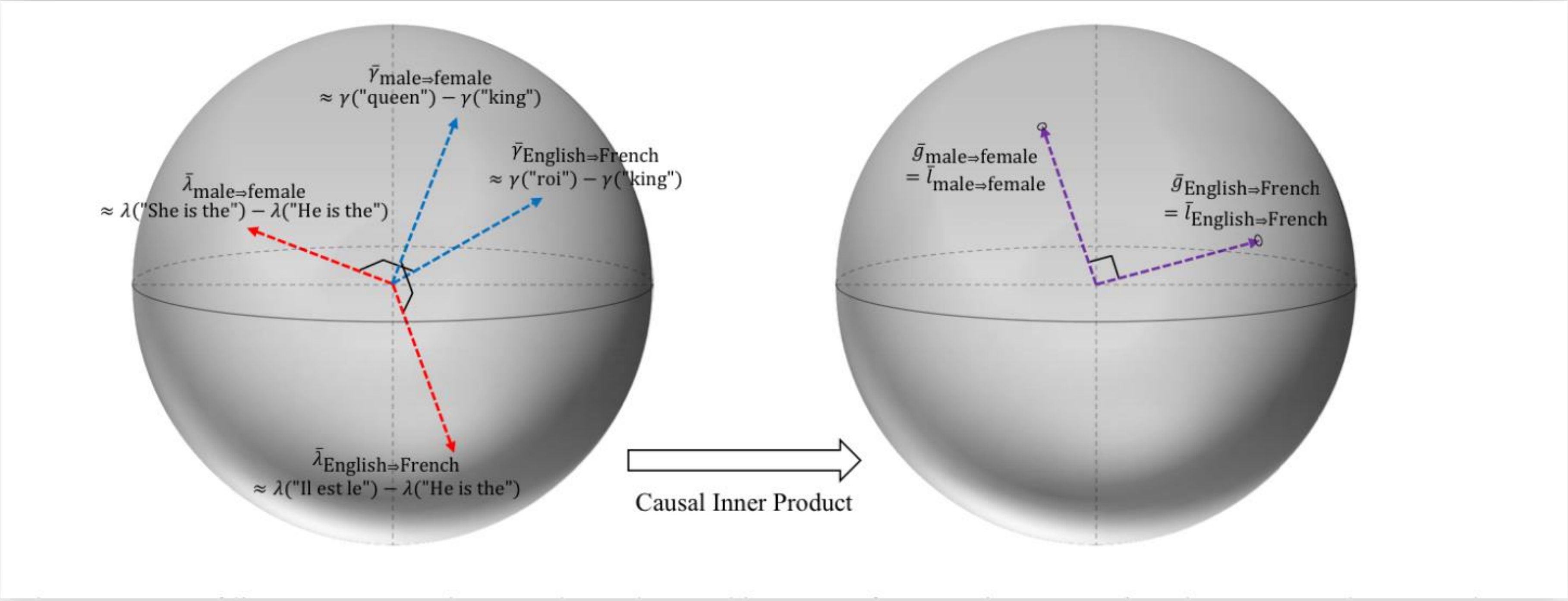


Unembedding Space



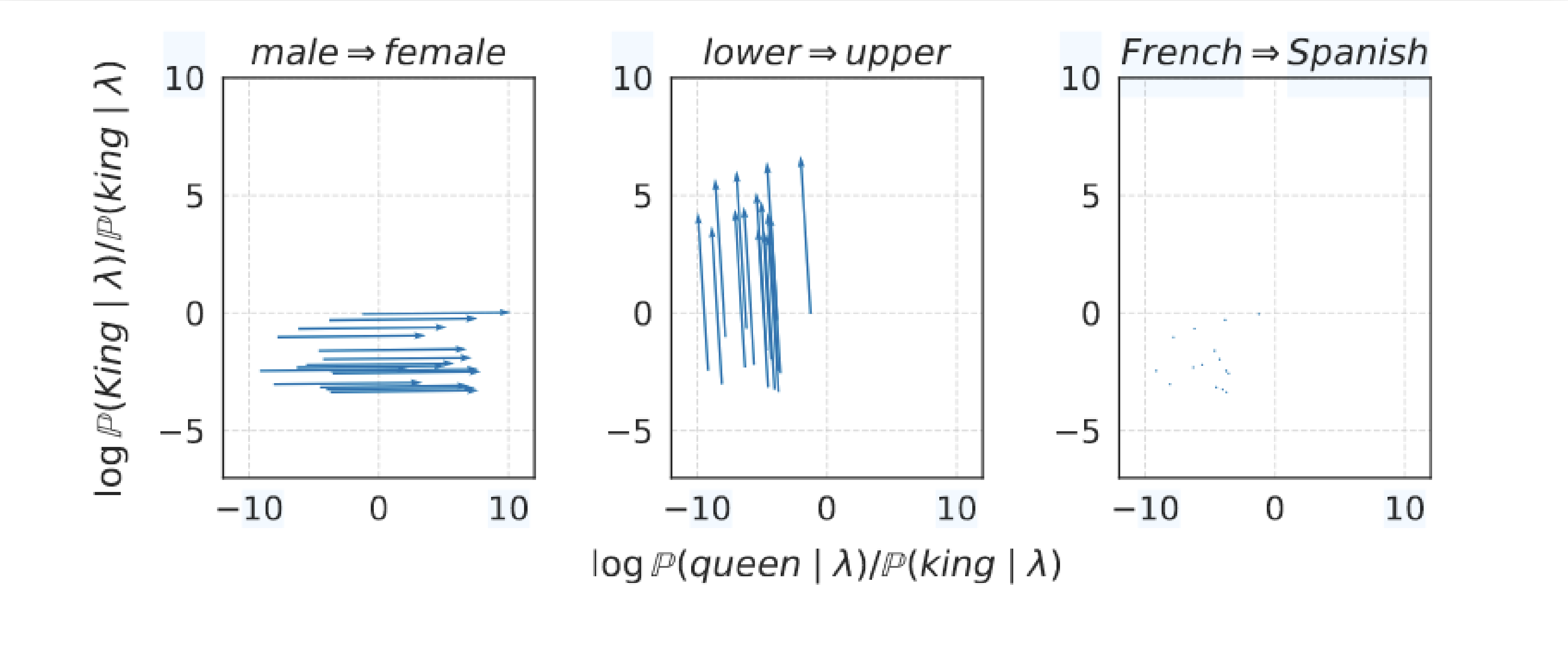
How can we leverage embeddings?

There exists a causal inner product that maps the embedding and unembedding spaces to each other



How can we leverage embeddings?

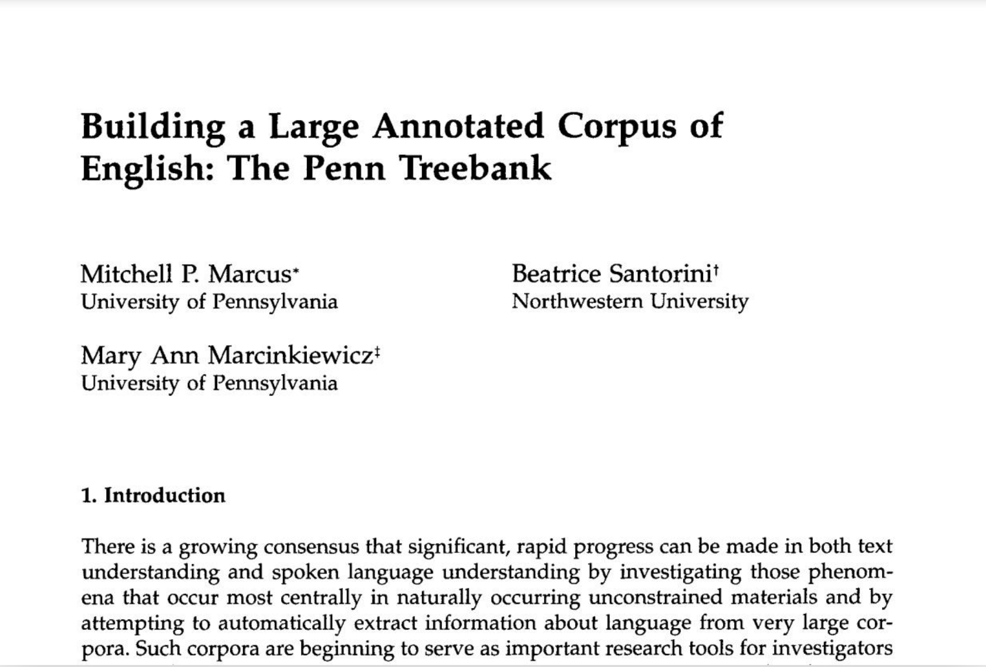
Use linear algebra operations for measurement and manipulation



How can we leverage embeddings?

#2 Isotropy in the Contextual Embedding Space: Clusters and Manifolds (X. Cai 2021)

“ GPT-2 embeds tokens in low-dimensional manifolds in the embedding space ”



Building a Large Annotated Corpus of English: The Penn Treebank

Mitchell P. Marcus*
University of Pennsylvania

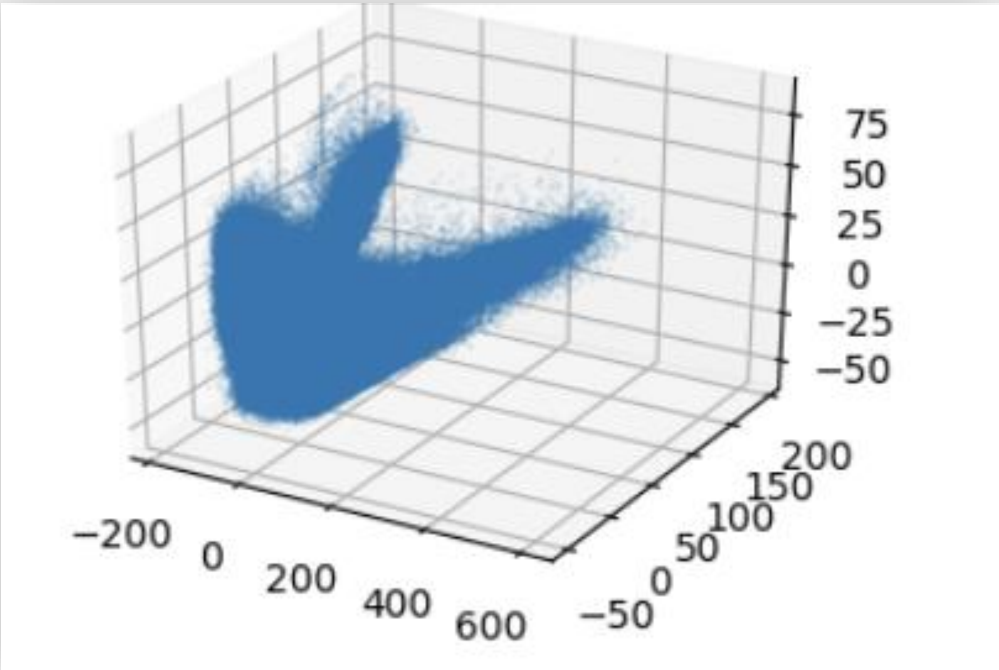
Beatrice Santorini†
Northwestern University

Mary Ann Marcinkiewicz‡
University of Pennsylvania

1. Introduction

There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora. Such corpora are beginning to serve as important research tools for investigators

?



How can we leverage embeddings?

Dimensionality reduction via Principal Component Analysis

- $E \in \mathbb{R}^{n \times d}$ matrix of vector embeddings
- Dimension reduction $d \to k$ = finding best fitting k-dim subspace of $\mathbb{R}^d$

Algorithm 1 Best Fit k-subspace

Input: A $n \times d$ matrix E
Output: Vectors $v_1, \ldots, v_r$

Determine $v_1 = \arg\max_{||v||=1} ||Av||$
$k = 2$
repeat
Determine $\alpha = \langle v_1, \ldots, v_{k-1} \rangle$
Determine $v_k = \arg\max_{v \perp \alpha,\, ||v||=1} ||Av||$
$k = k + 1$
until $v_k = 0$
return $v_1, \ldots, v_{k-1}$

How can we leverage embeddings?

Dimensionality reduction via Principal Component Analysis

- Singular Value Decomposition: $\exists\, U \in \mathbb{R}^{n\times n}, S \in \mathbb{R}^{n\times d}, V \in \mathbb{R}^{d\times d}$ s.t. $E = USV^T$
- $VV^T = I_d \implies F^T F = V^T E^T E V$ diagonal, $f_{ii} = \sigma_i(E)^2$ ordered squared singular values
- $\implies$ $\mathrm{Cov}(\mathrm{F})$ diagonal $\implies$ PCs correspond to directions with most variance
- Explained variance ratio $r_k \coloneqq \frac{\sum_{i=0}^{k} \sigma_i}{\sum_{i=0}^{d} \sigma_i}$, σ_i i-th largest eigenvalue of $\mathrm{Cov}(E)$
- ϵ-effective dimension $d(\epsilon) = \arg\min r_k \geq \epsilon$

How can we leverage embeddings?

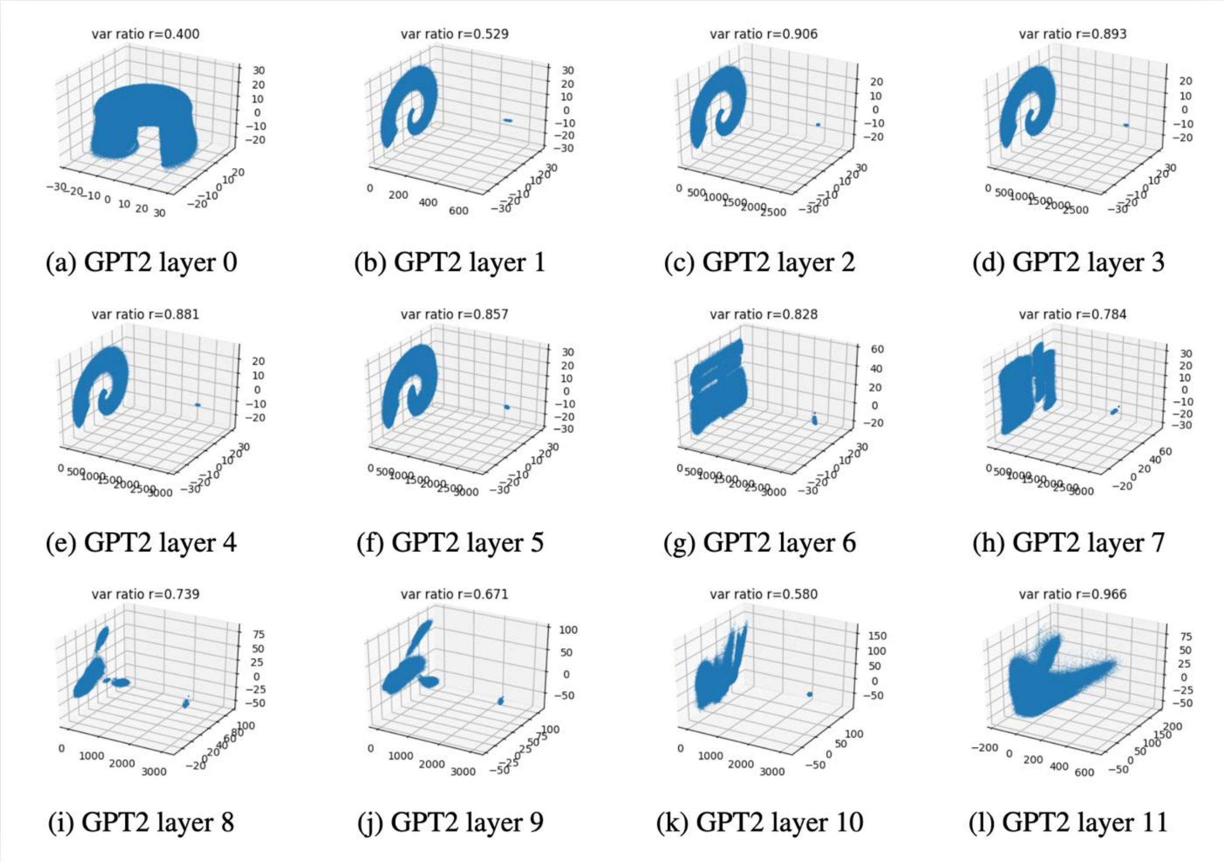
The embedding vectors of GPT-2 fall in a very narrow cone

The effective dimension $d(0.8)$

Layer	0	1	2	3	4	5	6	7	8	9	10	11
BERT	262	273	271	273	276	283	288	282	282	282	283	270
D-BERT	244	226	232	227	217	175						
GPT	265	141	65	76	173	210	205	217	221	253	269	307
GPT2	114	73	1	1	1	1	1	8	26	66	116	1
ELMo	455	367										

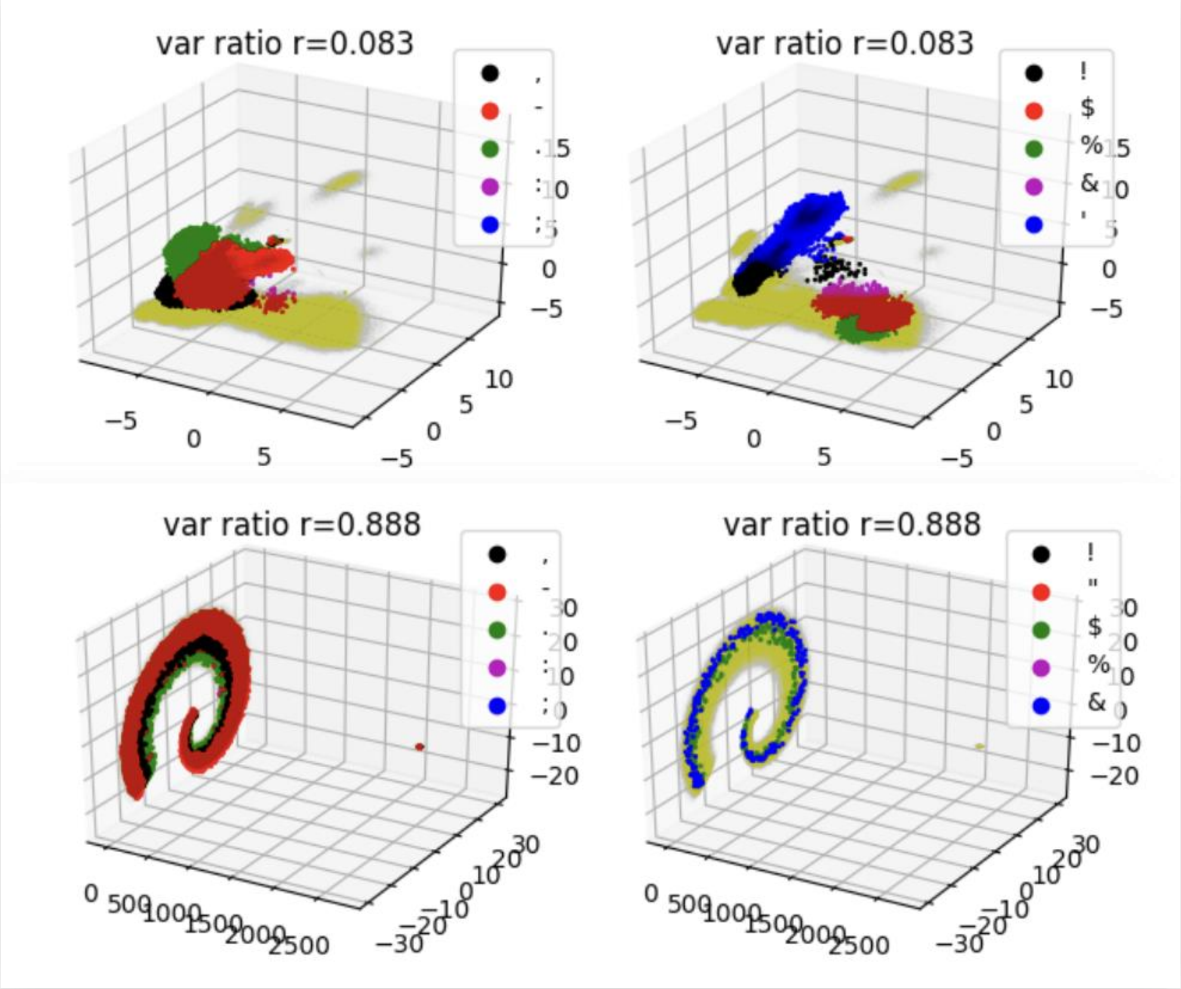
How can we leverage embeddings?

The 3-dimensional embedding space of GPT-2 consists of isolated clusters



How can we leverage embeddings?

The embedding space of GPT-2 has a different geometry than BERT



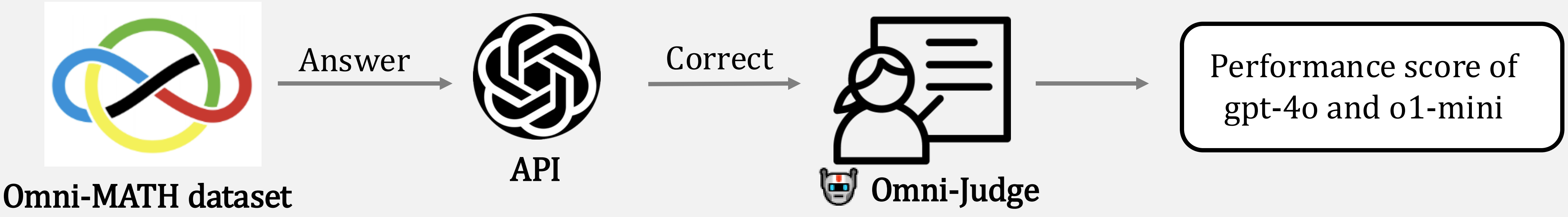
BERT

GPT-2

Part 3: How does GPT-4 (and o1) solve math questions?

How does GPT-4 solve math questions?

Experiment pipeline (part 1)



How does GPT solve math questions?

A new challenging math benchmark

Domain: Applied Mathematics

Difficulty: 3.0

Source: HMMT

Problem: You have infinitely many boxes, and you randomly put 3 balls into them. The boxes are labelled $1, 2, \ldots$ Each ball has probability $1/2^n$ of being put into box n. The balls are placed independently of each other. What is the probability that some box will contain at least 2 balls?

Solution: Notice that the answer is the sum of the probabilities that boxes $1, 2, \ldots$ respectively, contain at least 2 balls, since those events are mutually exclusive. For box n, the probability of having at least 2 balls is

$$3\left[\left(\frac{1}{2^n}\right)^2\left(1-\frac{1}{2^n}\right)\right]+\left(\frac{1}{2^n}\right)^3=\frac{3}{2^{2n}}-\frac{2}{2^{3n}}=\frac{3}{4^n}-\frac{2}{8^n}.$$

Summing to infinity using the geometric series formula, we get the answer $\frac{\frac{3}{4}}{1-\frac{1}{4}}-\frac{\frac{2}{8}}{\left(1-\frac{1}{8}\right)}$ which is equal to $\frac{5}{7}$.

Answer: $\frac{5}{7}$

How does GPT-4 solve math questions?

A new challenging math benchmark

Domain: Algebra

Difficulty: 5.0

Source: HMMT

Problem: A sequence of positive integers is defined by $a_0 = 1$ and $a_{n+1} = a_n^2 + 1$ for each $n \geq 0$. Find $\gcd(a_{999}, a_{2004})$.

Solution: If d is the relevant greatest common divisor, then $a_{1000} = a_{999}^2 + 1 \equiv 1 = a_0 \ mod\ d$ which implies (by induction) that the sequence is periodic modulo d, with period 1000. In particular, $a_4 \equiv a_{2004} \equiv 0$. So d must divide a_4. Conversely, we can see that $a_5 = a_4^2 + 1 \equiv 1 = a_0 \ mod\ a_4$ so (again by induction) the sequence is periodic modulo a_4 with period 5, and hence a_{999}, a_{2004} are indeed both divisible by a_4. So the answer is a_4, which we can compute directly; it is 677.

Answer: 677

How does GPT-4 solve math questions?

A new challenging math benchmark

Domain: Discrete Mathematics

Difficulty: 7.0

Source: USA team selection test

Problem: A tournament is a directed graph for which every pair of vertices has a single directed edge from one vertex to the other. Let us define a proper directed-edge-coloring to be an assignment of a color to every (directed) edge, so that for every pair of directed edges $\overrightarrow{uv}$ and $\overrightarrow{vw}$ those two edges are in different colors. […] For each n, determine the minimum directed-edge-chromatic-number over all tournaments on n vertices.

Solution: […] For each \\(n\\), we aim to determine the minimum directed-edge-chromatic-number over all tournaments on \\(n\\) vertices. A construction is as follows: Label the vertices \\(\\{1, 2, 3, \\ldots, n\\}\\) and write each label in binary. Then for any two vertices \\(u, v\\), let \\(k\\) be the first position from the right which is different in their binary representations. […]

Answer: $\lceil \log_2 n \rceil$

How does GPT-4 solve math questions?

A new challenging math benchmark

Domain: Geometry

Difficulty: 9.5

Source: IMO shortlist

Problem: A circle ω with radius 1 is given. A collection T of triangles is called *good*, if the following conditions hold:

1. Each triangle from T is inscribed in ω;
2. No two triangles from T have a common interior point.

Determine all positive real numbers t such that, for each positive integer n, there exists a *good* collection of n triangles, each of perimeter greater than t.

Solution: Consider a circle ω with radius 1. We will determine the set of all positive real numbers t such that for each positive integer n, there exists a *good* collection T of n triangles inscribed in ω where each triangle has a perimeter greater than t. [...] Since each triangle in T is inscribed in a circle ω with radius 1, [...]

Answer: $0 < t \leq 4$

How does GPT-4 solve math questions?

Automated calls to the completions endpoint of the openai API

Domain: Discrete Mathematics

Difficulty: 7.0

Problem: A tournament is a directed graph for which every pair of vertices has a single directed edge from one vertex to the other. Let us define a proper directed-edge-coloring to be an assignment of a color to every (directed) edge, so that for every pair of directed edges $\overrightarrow{uv}$ and $\overrightarrow{vw}$ those two edges are in different colors. [...] For each n, determine the minimum directed-edge-chromatic-number over all tournaments on n vertices.

Model generation: To solve the problem of determining the minimum directed-edge-chromatic-number over all tournaments on n vertices, we start by examining the constraints for proper directed-edge-coloring in a tournament. [...] Key Observations: **Out-degree and Patterns**: In any tournament, each vertex has an out-degree of $n-1$. Given that for each vertex v, the edges $\overrightarrow{uv}$ (u as any vertex) appear, and there are $n-2$ patterns $\overrightarrow{uv} \rightarrow \overrightarrow{vw}$, we need different colors for these patterns. **Chromatic Condition**: The directed-edge-coloring requires different colors for edges forming transitive triples. For any directed edge-coloring in a tournament [...]

$$\boxed{n-1}$$

How does GPT-4 solve math questions?

Another LLM for correcting the answers

Domain: Discrete Mathematics

Difficulty: 7.0

Problem: A tournament is a directed graph for which every pair of vertices has a single directed edge from one vertex to the other. [...]

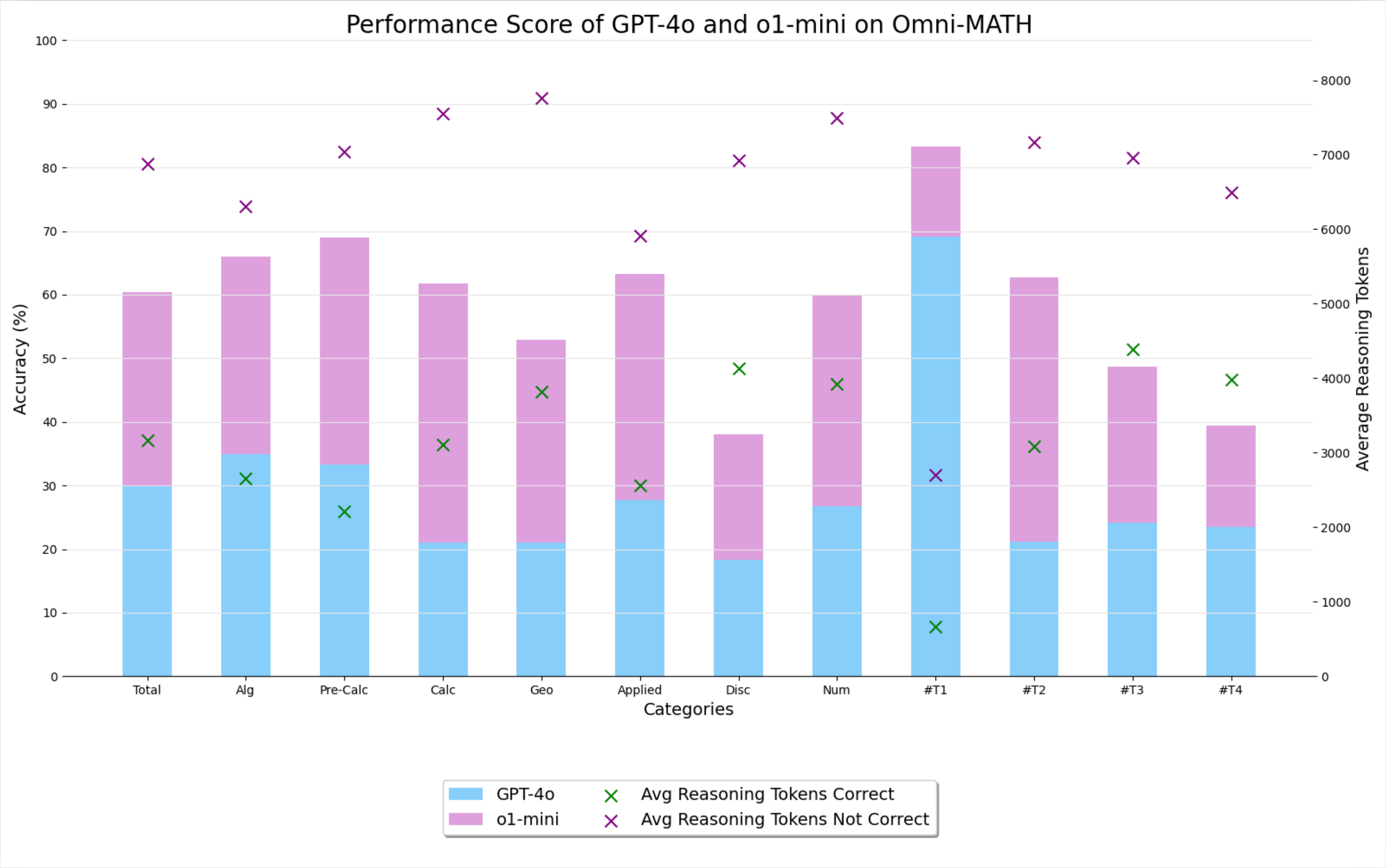
Model generation: To solve the problem of determining the minimum directed-edge-chromatic-number over all tournaments on n vertices, we start by [...]

$$\boxed{n-1}$$

Omni-judge: ## *Student Final Answer*: n – 1 ## *Equivalence Judgement*: FALSE ## *Justification*: The student's answer of $n-1$ does not share the same meaning as the reference answer, which is $\lceil \log_2 n \rceil$. [...] The student incorrectly assumes a direct relationship between the number of edges and the number of colors needed for proper directed-edge-coloring in a tournament. In reality, the problem requires considering the combinatorial structure of the tournaments and how it affects the coloring requirements. The reference answer provides an accurate expression that reflects this complexity. === report over ===

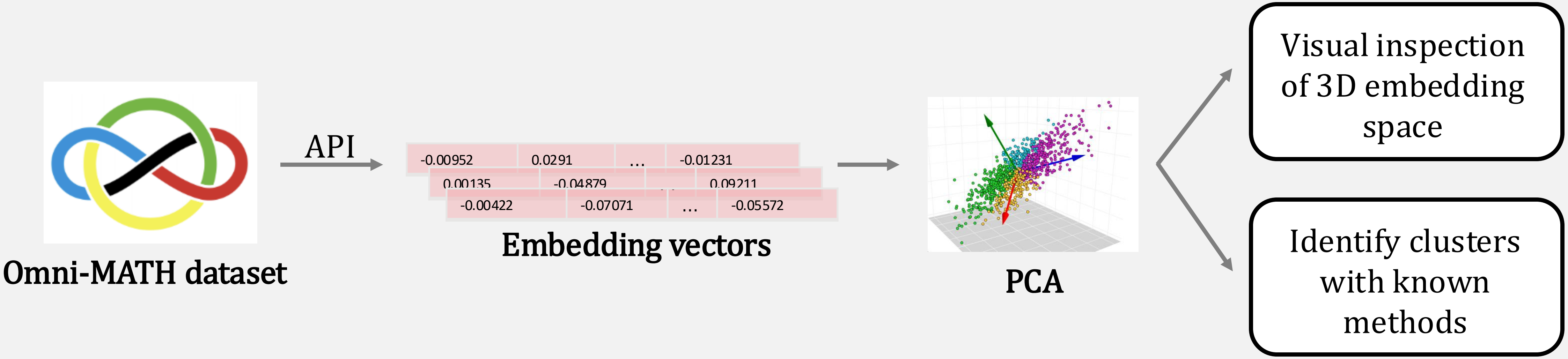
How does GPT-4 solve math questions?

Computing domain and difficulty specific accuracy



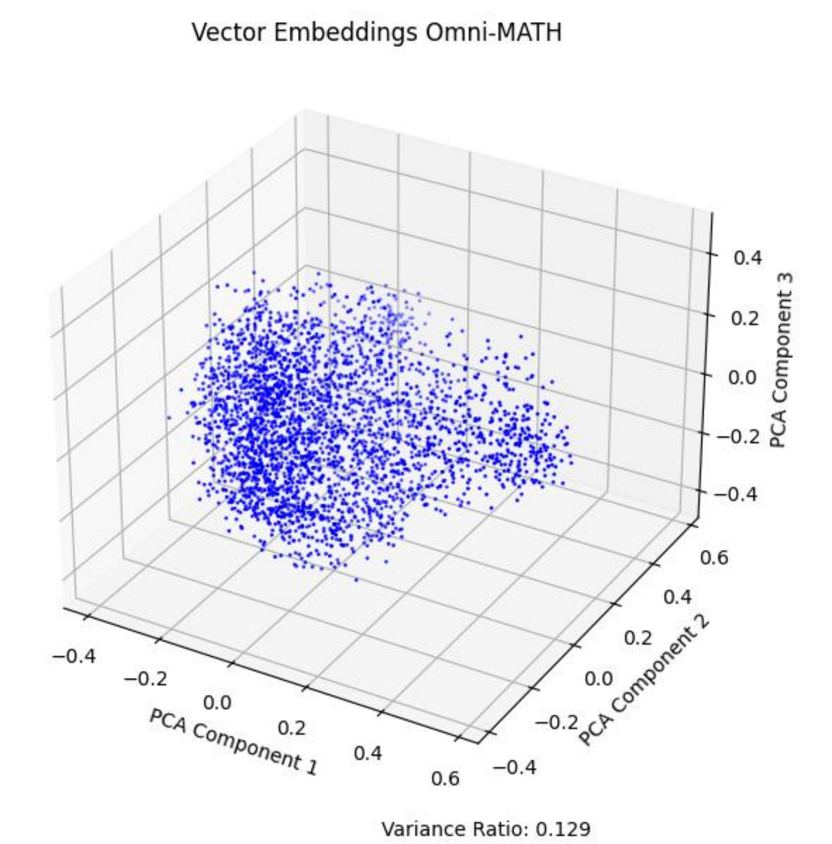
How does GPT-4 solve math questions?

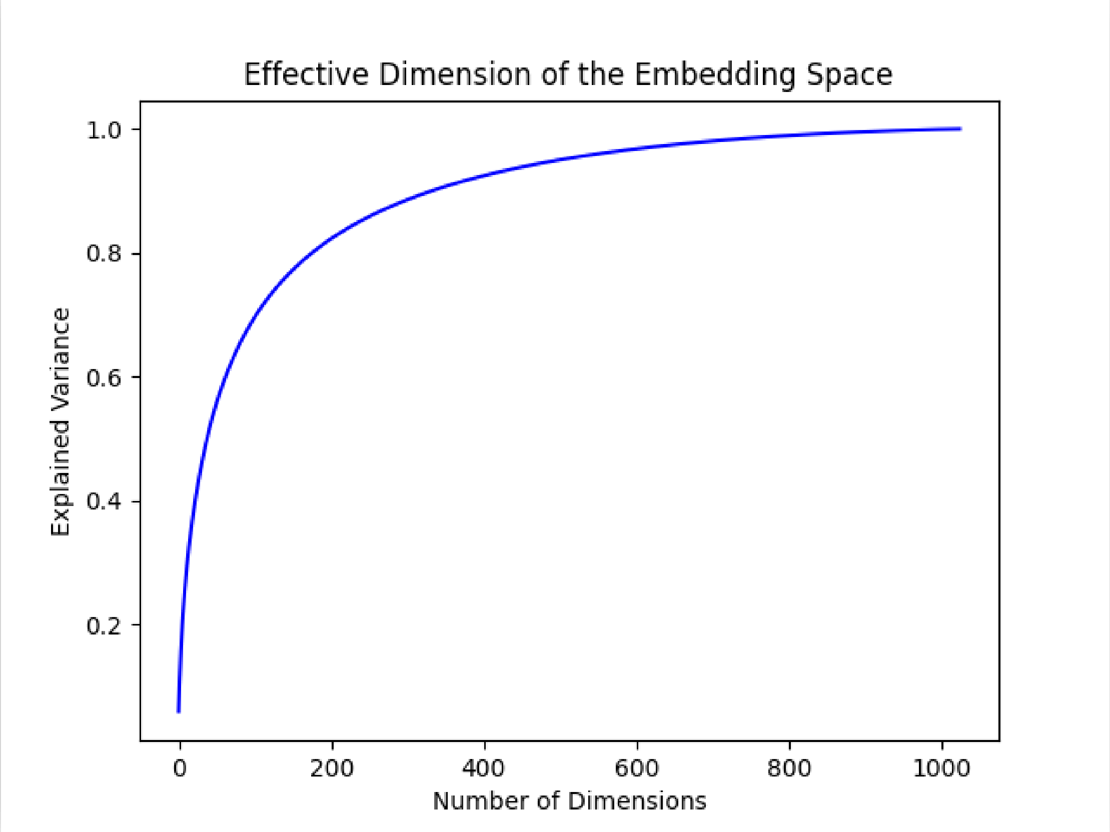
Experiment pipeline (part 2)



How does GPT-4 solve math questions?

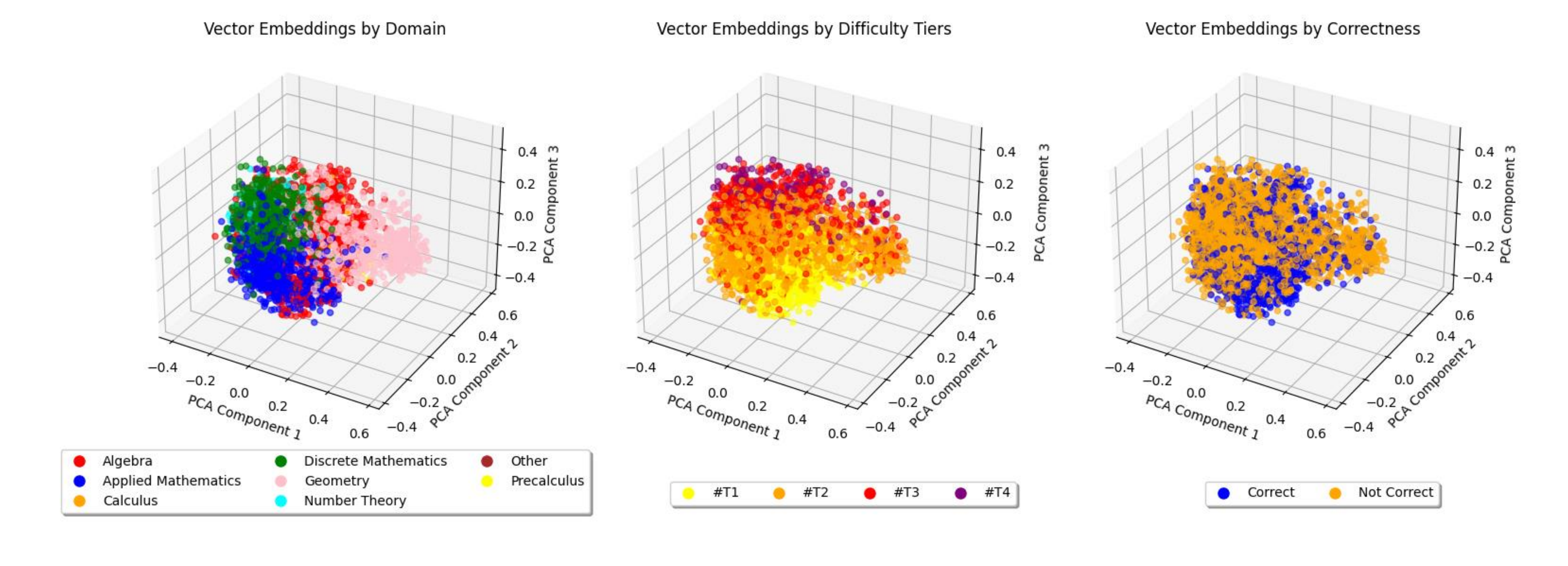
Dimensionality reduction via Principal Component Analysis





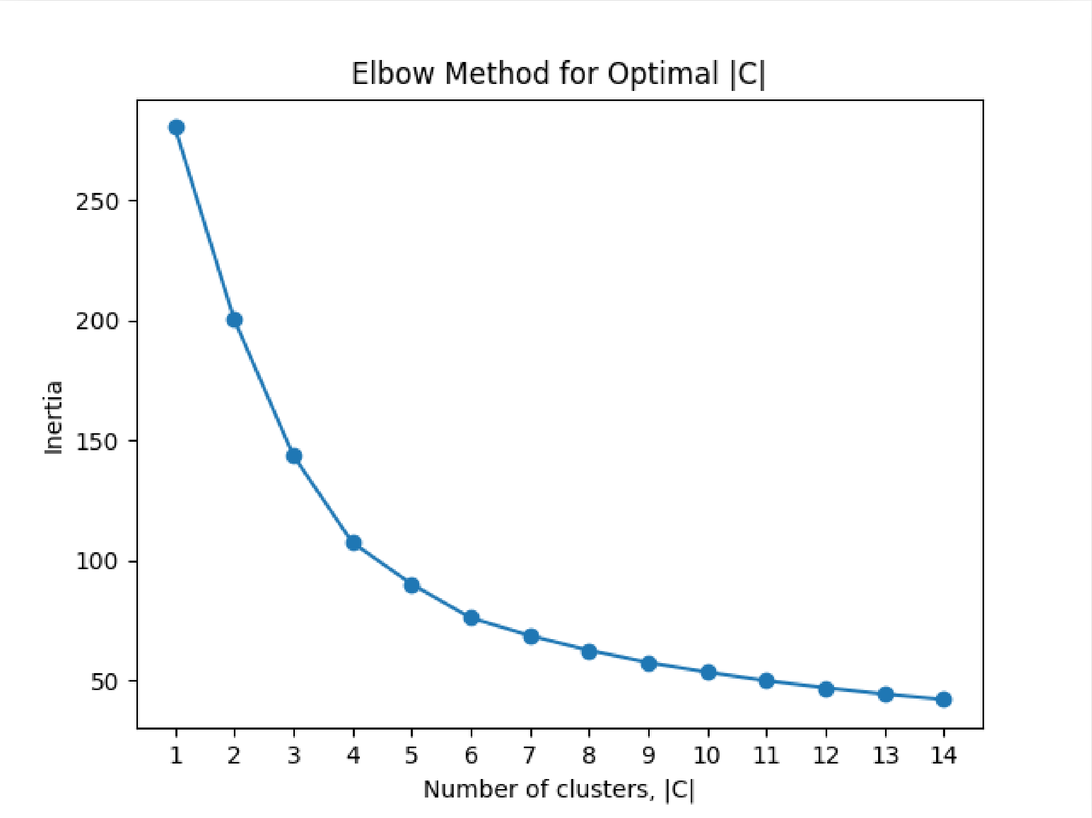
How does GPT-4 solve math questions?

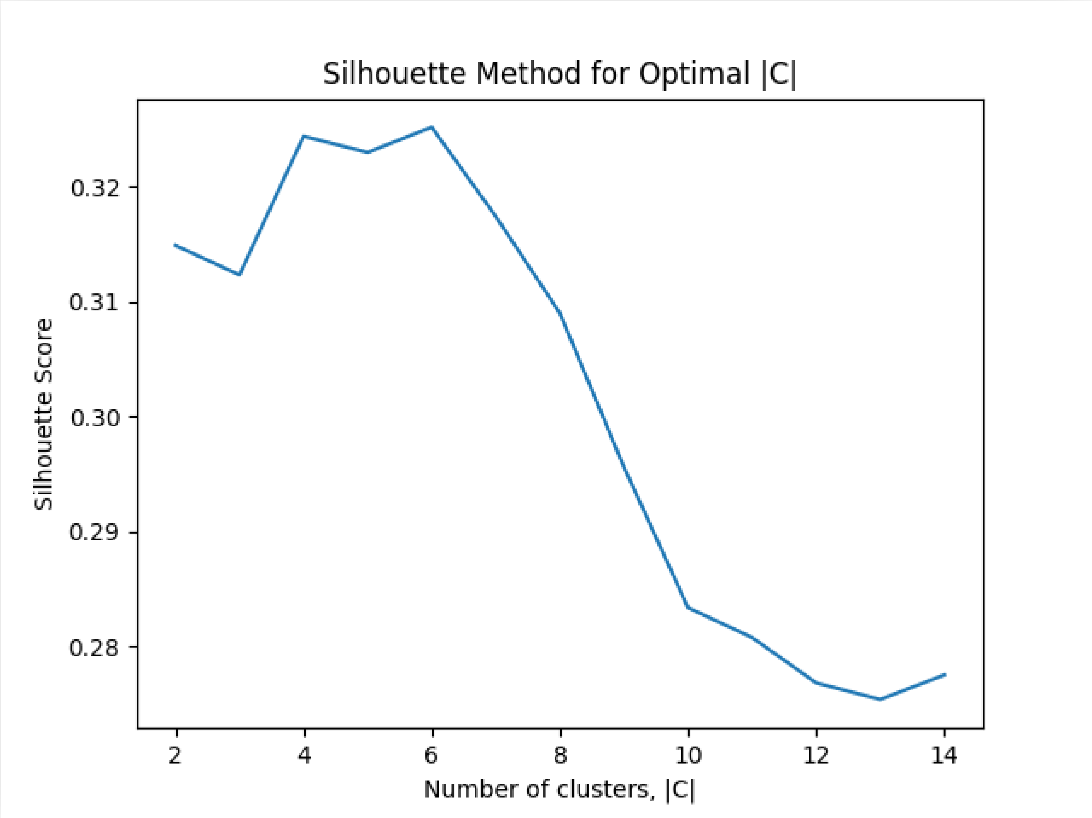
Where do which questions lie in the embedding space?



How does GPT-4 solve math questions?

Where do which questions lie in the embedding space? (part 2)

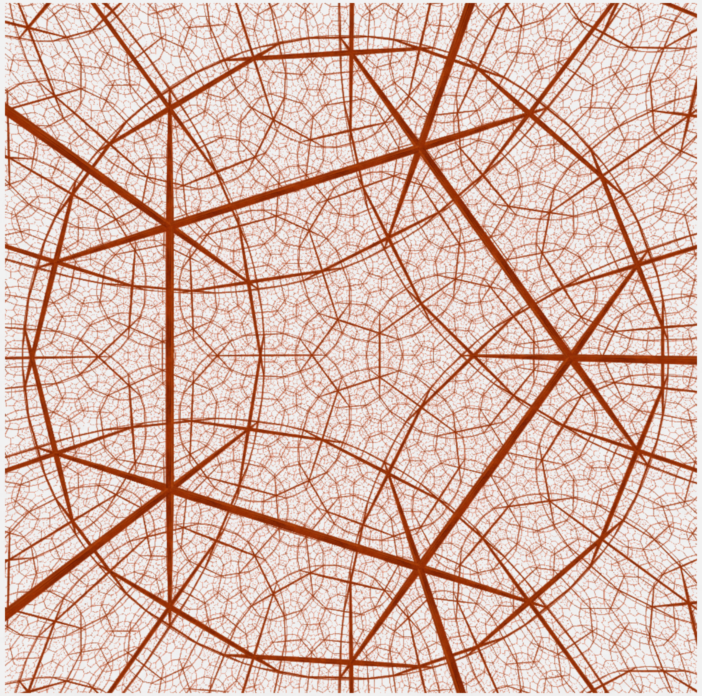




How does GPT-4 solve math questions?

There are endless possibilities for studying the geometry of embedding spaces…

- Euclidean ⟶ ?
- Non-linear dimensionality reduction
- Non-convex or hierarchical clustering
- Topological Data Analysis, Manifold Hypothesis
- …



Thank you!

More questions??

marthe.ballon@vub.be

SCAN ME

Memorization vs. Reasoning: What the vec happens inside LLMs (when solving math questions)?

Marthe Ballon | Junior Colloquium UCL 27/11/2024 | Vincent Ginis

Design by Seorin Kim