

2025/05/22 Stanford NLP Seminar

Closing the Modality Gap: Benchmarking and Improving Visual Understanding in Multimodal LLMs

Deqing Fu

University of Southern California

Hunt for Meaningful Failures

- ❖ (Multimodal) LLMs work quite well, but they can still fail in some situations.
- ❖ How can we **understand** these failures and potentially **mitigate** some of them?
- ❖ How to build meaningful **evaluations** beyond ranking models?

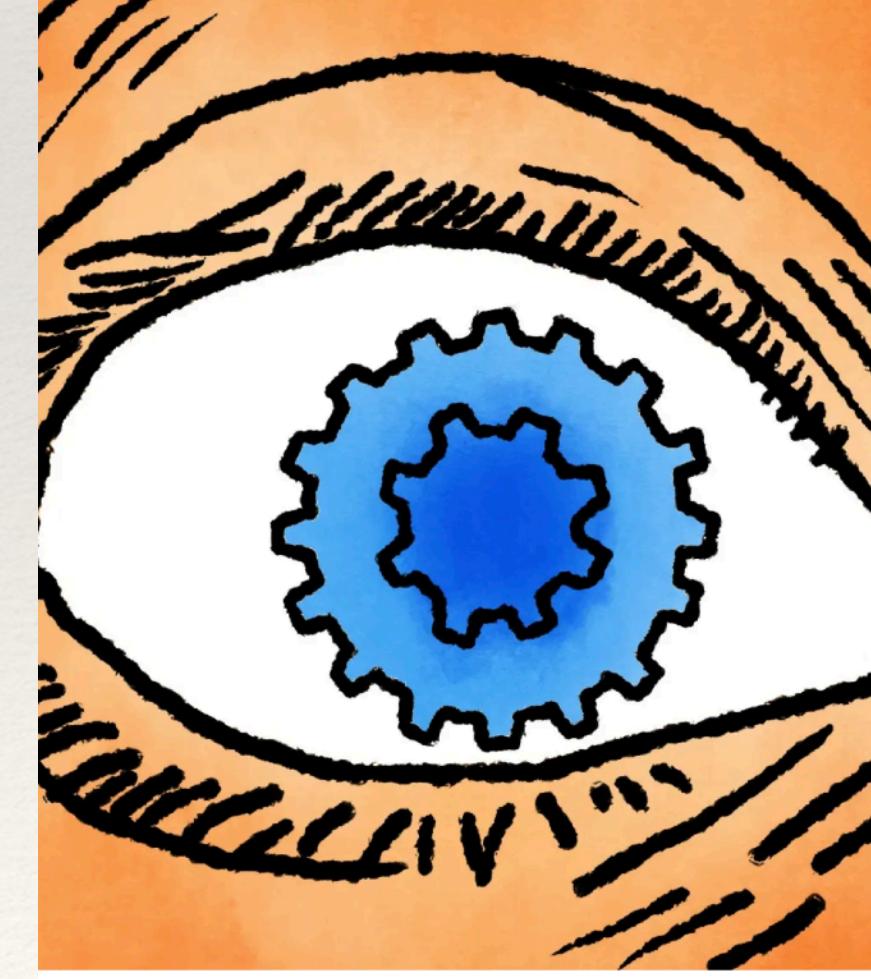
≡ **WIRED** SECURITY POLITICS THE BIG STORY MORE ▾ SIGN IN | SUBSCRIBE 

KYLE ORLAND, Ars Technica BUSINESS OCT 15, 2024 7:55 PM

Apple Engineers Show How Flimsy AI ‘Reasoning’ Can Be

The new frontier in large language models is the ability to “reason” their way through problems. New research from Apple says it’s not quite what it’s cracked up to be.

TC TechCrunch Latest Startups Venture Apple Security AI Apps Google I/O 2025 | Events Podcasts Newsletters  



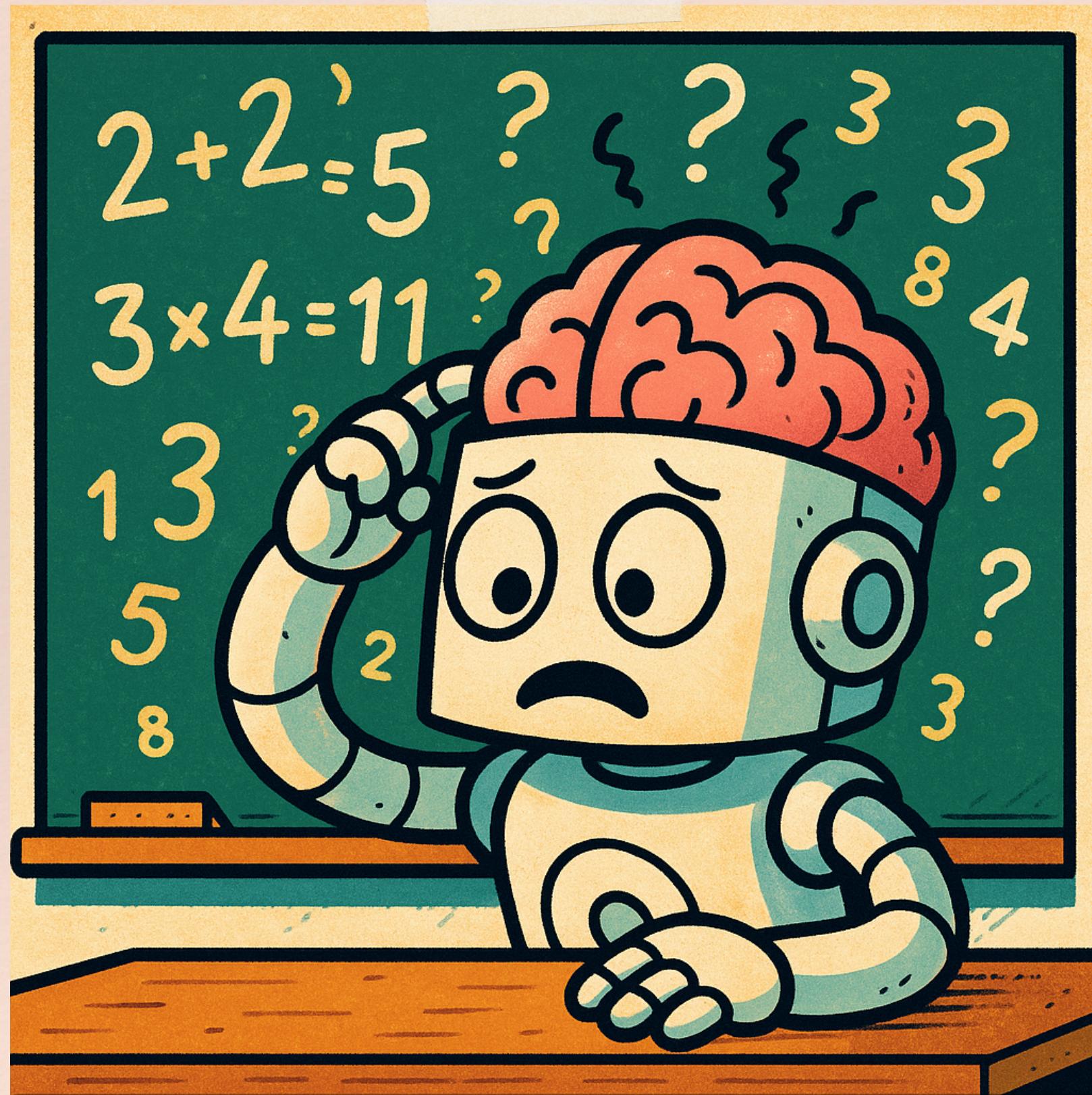
AI

“Visual” AI models might not see anything at all

Devin Coldewey — 10:41 AM PDT · July 11, 2024

Failures Covered In This Talk

- ❖ LLMs Are Bad at Arithmetics



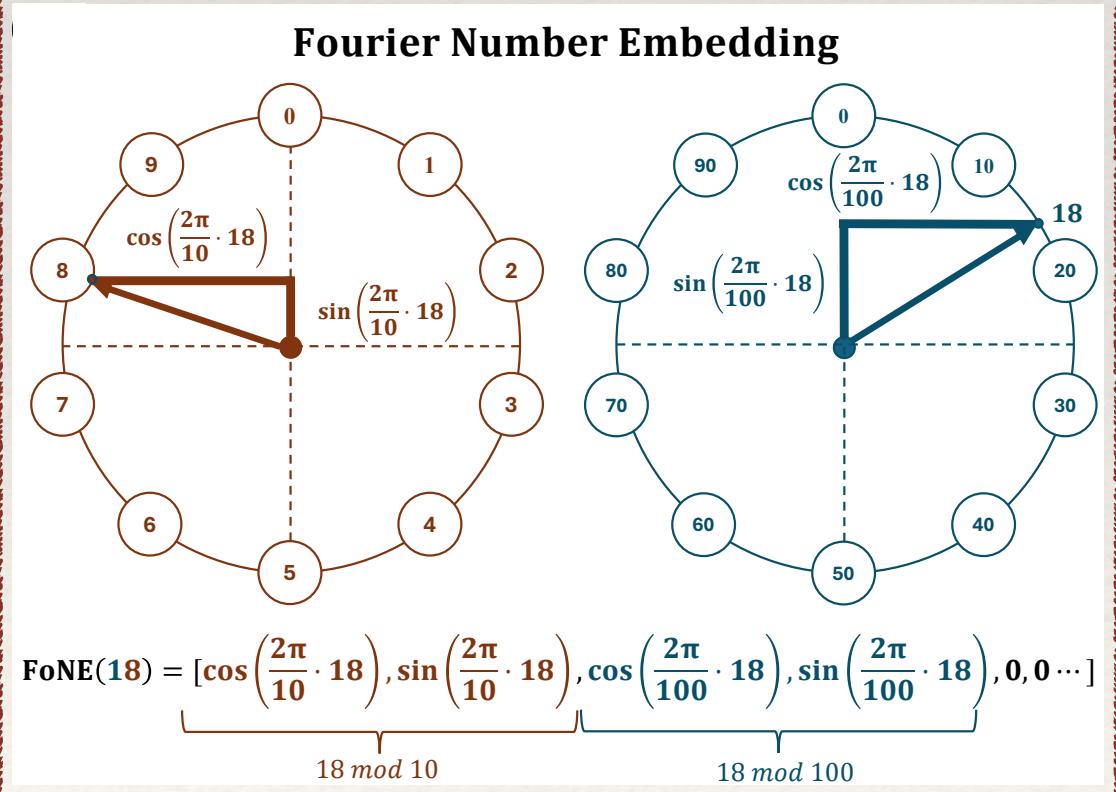
- ❖ Multimodal LLMs Hallucinate



Overview

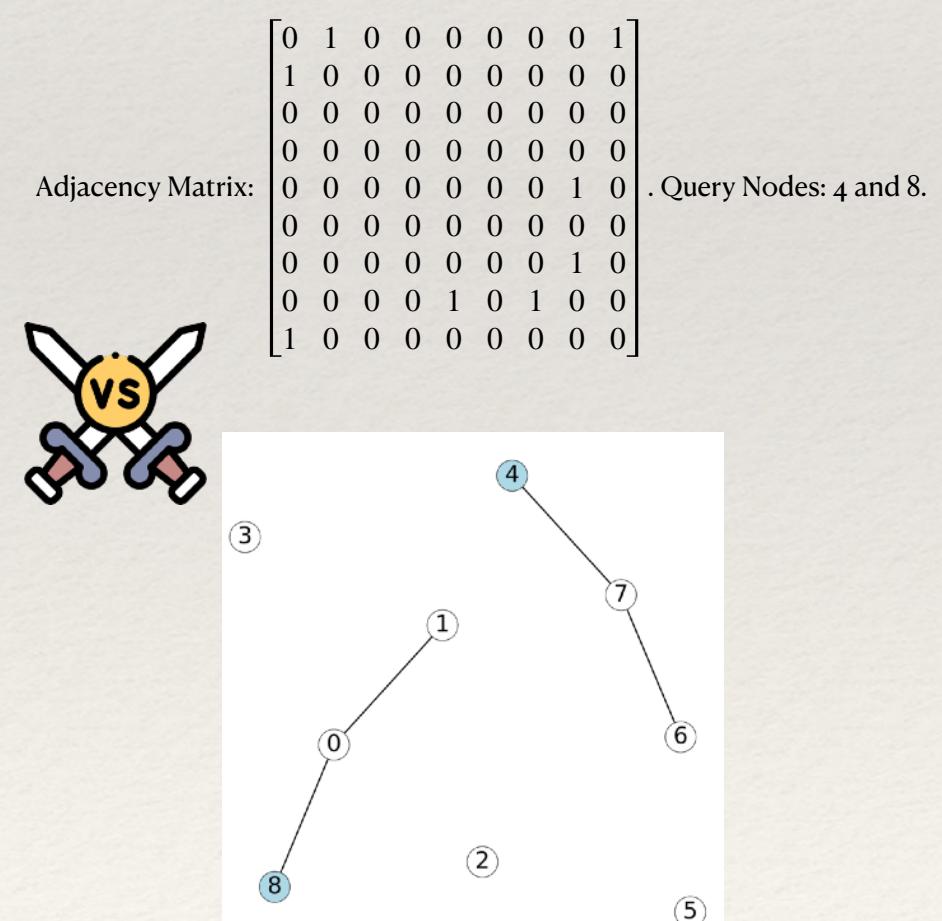
LLMs are Bad At Arithmetics.

Can we find a family of embeddings to bypass tokenization?



Evaluating Multimodal LLMs Beyond Ranking Models.

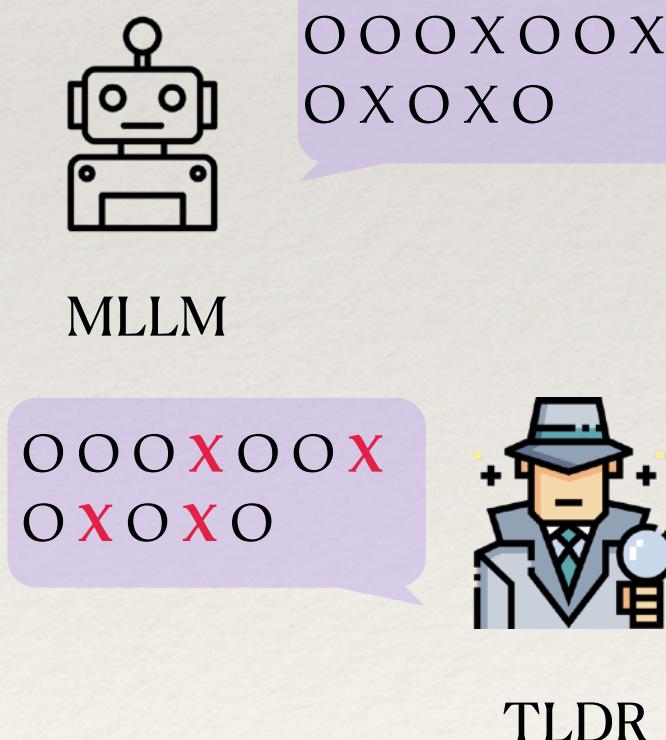
Humans have pictorial superiority. How about MLLMs?



Text LLMs are Teachers for MLLMs.

Fine-Grained Rewards.

Train Text Backbone
to be Token-Level
Reward Models.



Text Steering Vectors.

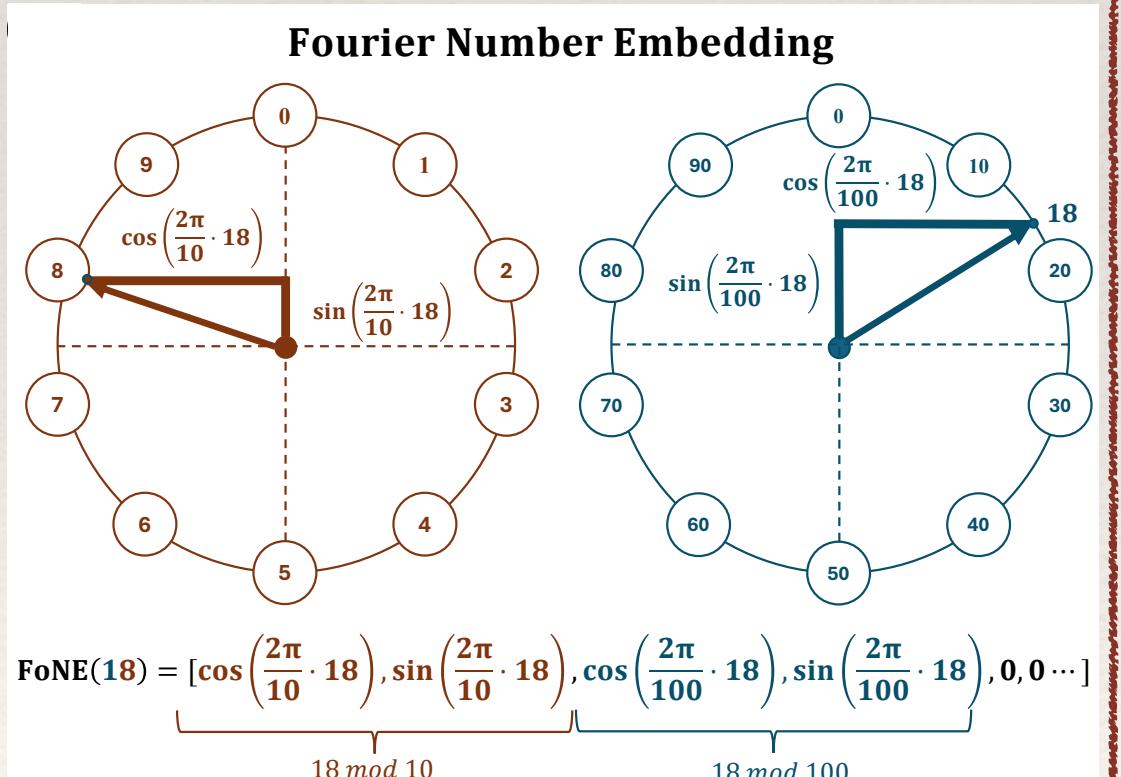
Leverage Text-Based
Interpretability Tools
for MLLMs.



Overview

LLMs are Bad At Arithmetics.

Can we find a family of embeddings to bypass
tokenization?



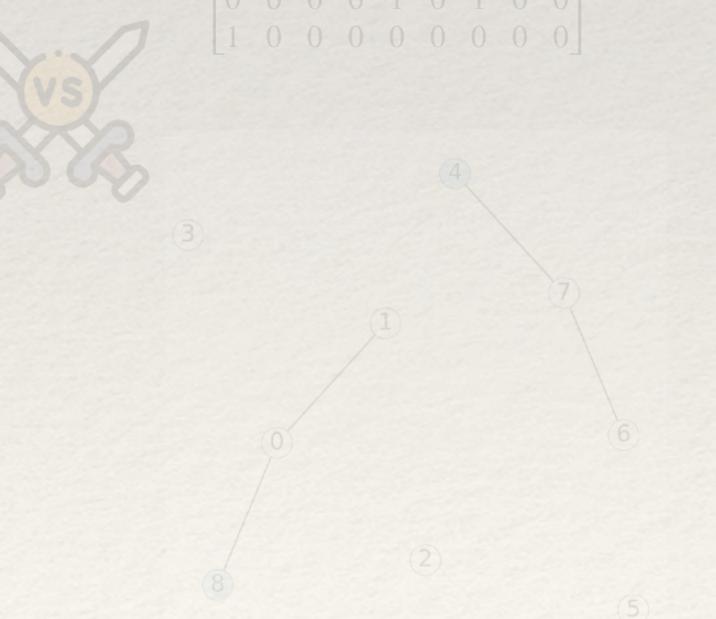
Evaluating Multimodal LLMs Beyond Ranking Models.

Humans have pictorial
superiority. How about
MLLMs?

Adjacency Matrix:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Query Nodes: 4 and 8.



Text LLMs are Teachers for MLLMs.

Fine-Grained Rewards.

Train Text Backbone
to be Token-Level
Reward Models.



Text Steering Vectors.

Leverage Text-Based
Interpretability Tools
for MLLMs.



FoNE: Precise Single-Token Number Embeddings via Fourier Features [In Submission]



Tianyi Zhou



Deqing Fu



Mahdi Soltanolkotabi



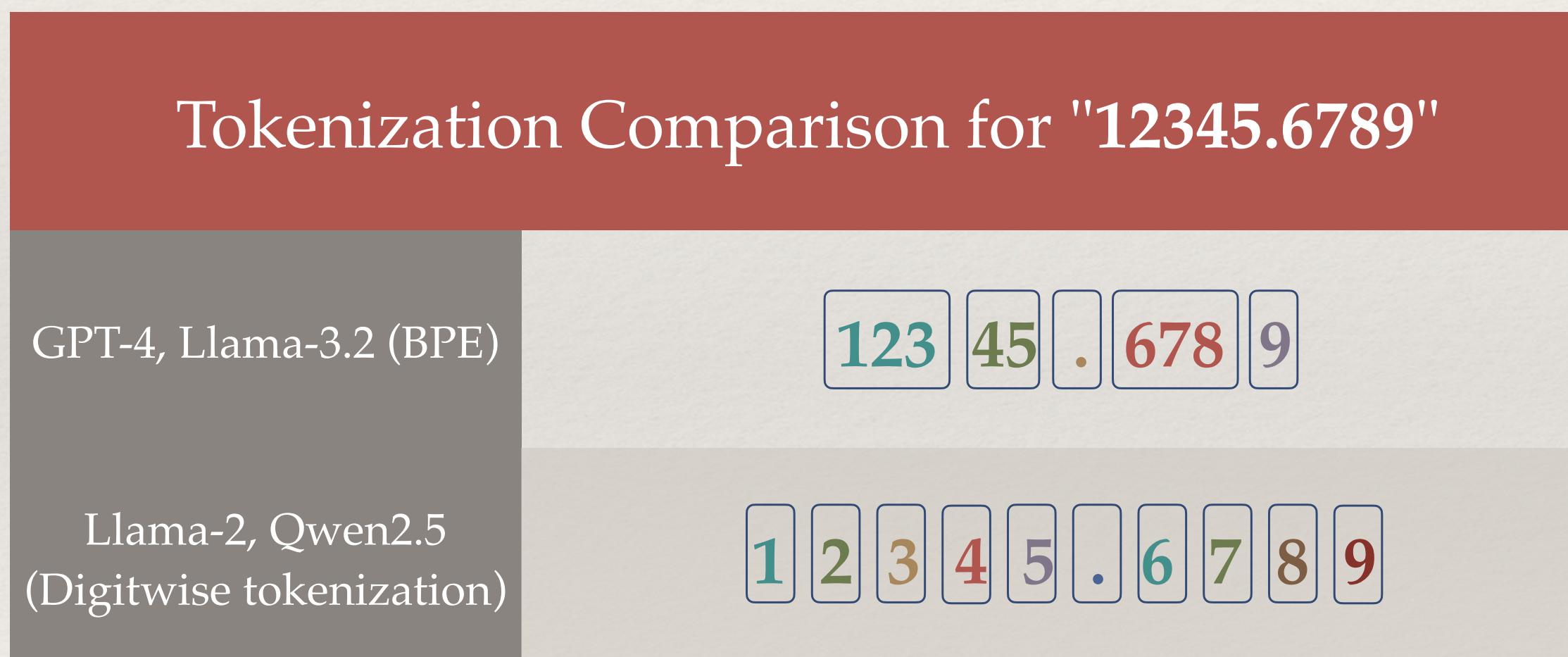
Vatsal Sharan



Robin Jia

A Famous Tweet

- ❖ **Tokenization** seems like a reason why LLMs are natively bad at simple arithmetics.



- ❖ Can we design a number embedding to bypass tokenization?

 Andrej Karpathy ✅
@karpathy

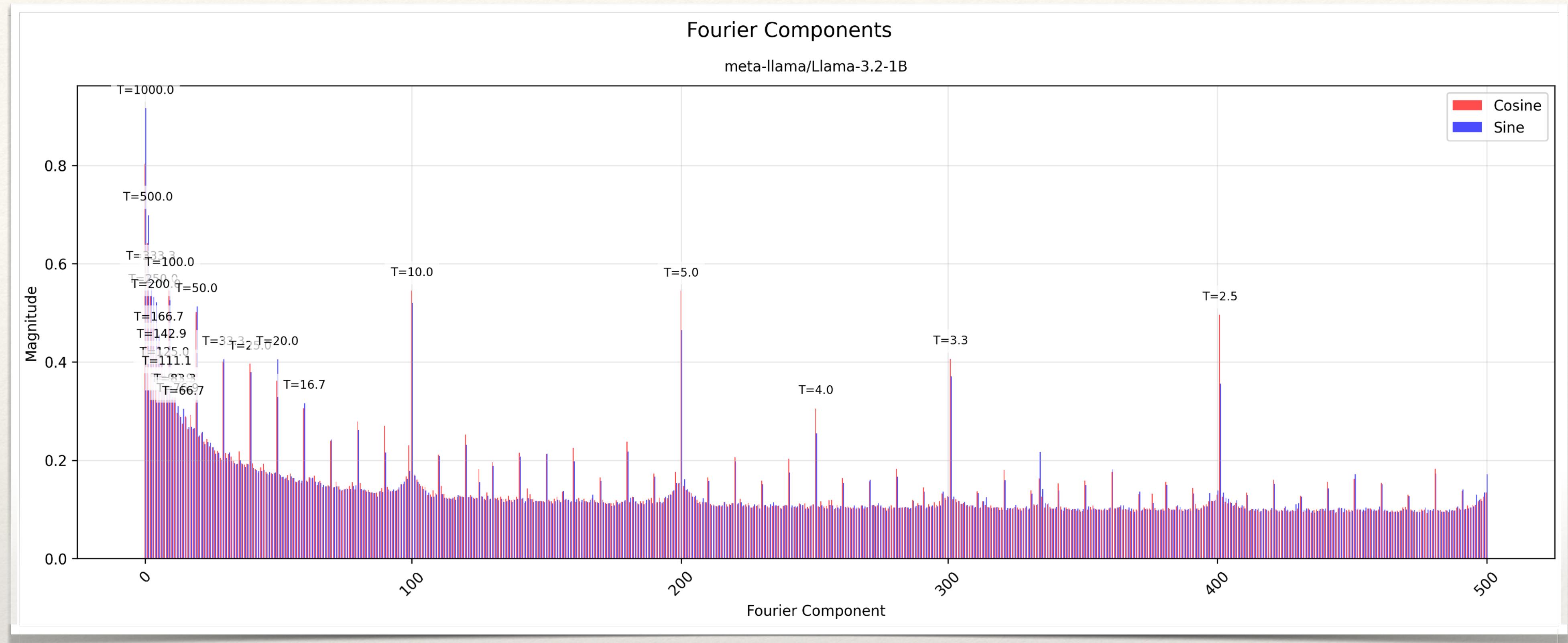
We will see that a lot of weird behaviors and problems of LLMs actually trace back to tokenization. We'll go through a number of these issues, discuss why tokenization is at fault, and why someone out there ideally finds a way to delete this stage entirely.

Tokenization is at the heart of much weirdness of LLMs. Do not brush it off.

- Why can't LLM spell words? **Tokenization**.
- Why can't LLM do super simple string processing tasks like reversing a string? **Tokenization**.
- Why is LLM worse at non-English languages (e.g. Japanese)? **Tokenization**.
- Why is LLM bad at simple arithmetic? **Tokenization**. Why is LLM bad at simple arithmetic? **Tokenization**.
- Why did GPT-2 have more than necessary trouble coding in Python? **Tokenization**.
- Why did my LLM abruptly halt when it sees the string "<|endoftext|>"? **Tokenization**.
- What is this weird warning I get about a "trailing whitespace"? **Tokenization**.
- Why the LLM break if I ask it about "SolidGoldMagikarp"? **Tokenization**.
- Why should I prefer to use YAML over JSON with LLMs? **Tokenization**.
- Why is LLM not actually end-to-end language modeling? **Tokenization**.
- What is the real root of suffering? **Tokenization**.

9:40 AM · Feb 20, 2024 · 744.5K Views

59 402 2.7K 1K

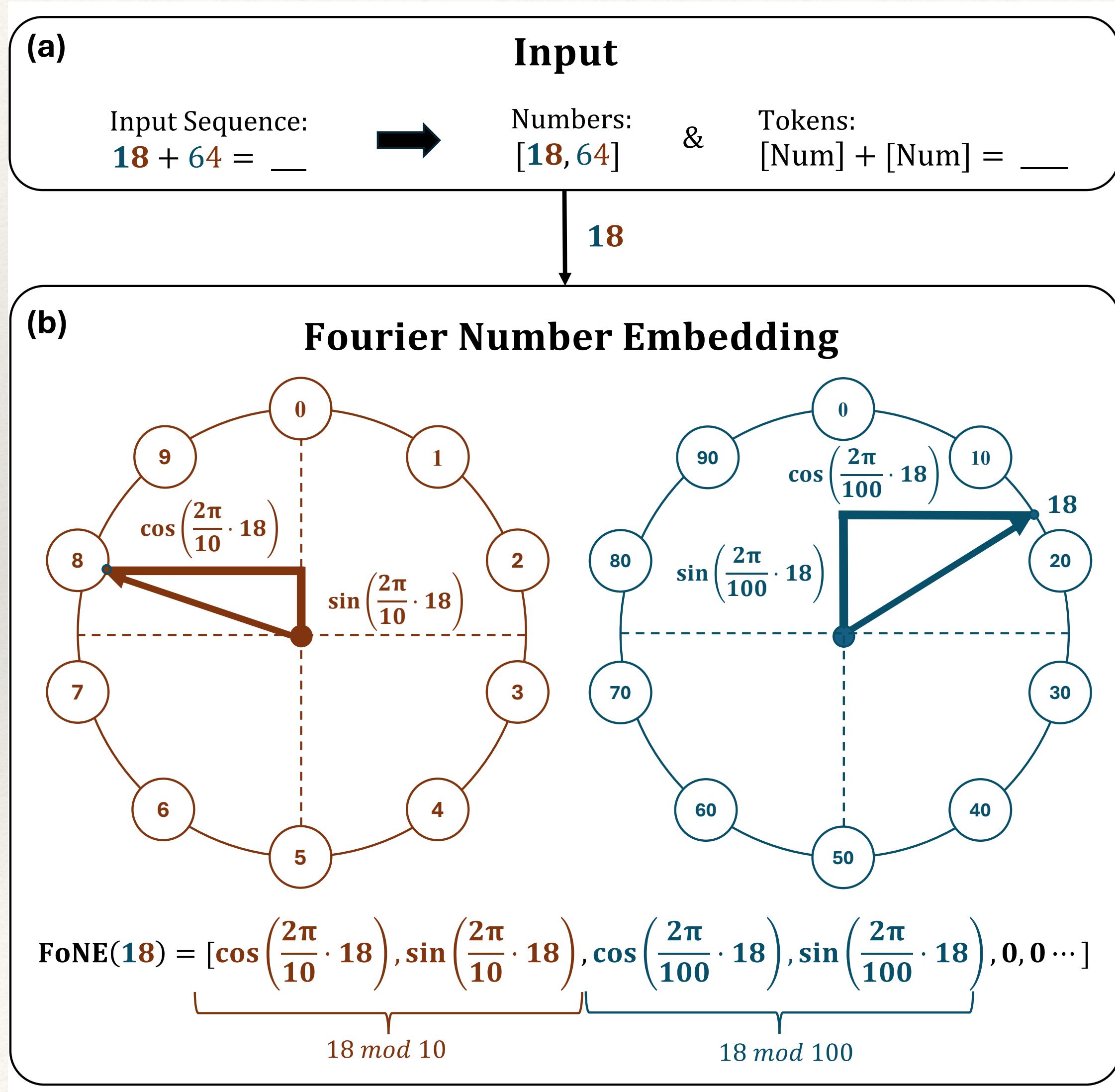


Zhou et. al. *Pre-trained Large Language Models Use Fourier Features to Compute Addition*. NeurIPS 2024

Pre-trained Number Tokens Are Embedded in a Fourier Space.

Can we Hard-Code Numbers?

FoNE: Fourier Number Embedding



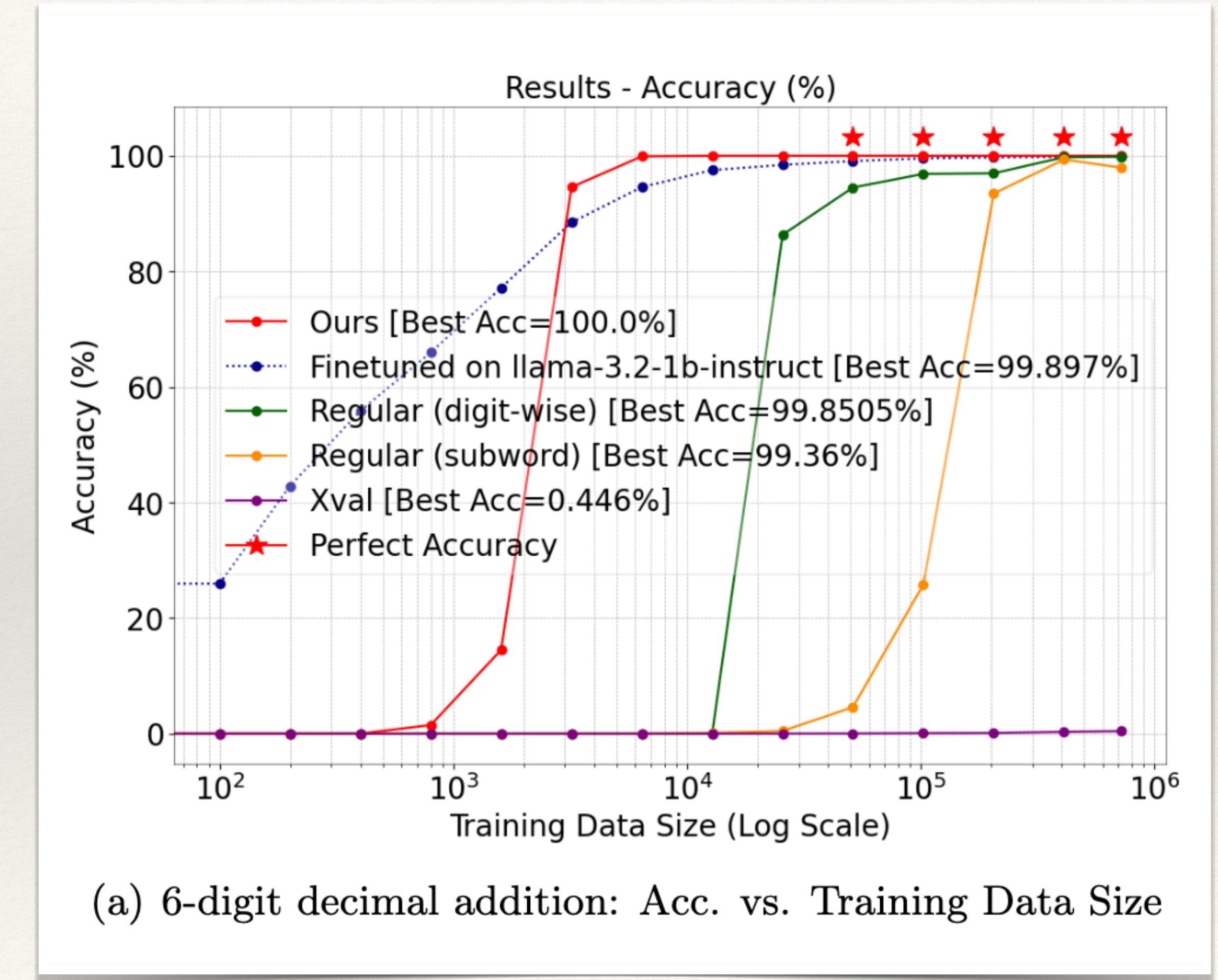
- ❖ We hard-code embeddings for any number as the following,

$$\text{FoNE}(x) = \begin{bmatrix} \cos\left(\frac{2\pi}{10} \cdot x\right) \\ \sin\left(\frac{2\pi}{10} \cdot x\right) \\ \cos\left(\frac{2\pi}{100} \cdot x\right) \\ \sin\left(\frac{2\pi}{100} \cdot x\right) \\ \vdots \end{bmatrix} = \begin{bmatrix} \cos\left(\frac{2\pi}{T_j} \cdot x\right) \\ \sin\left(\frac{2\pi}{T_j} \cdot x\right) \end{bmatrix}_j$$

- ❖ You may choose the periods T_j to be any number. For floating numbers, we can choose $T_j = 0.1, 0.01, 0.001, \dots$

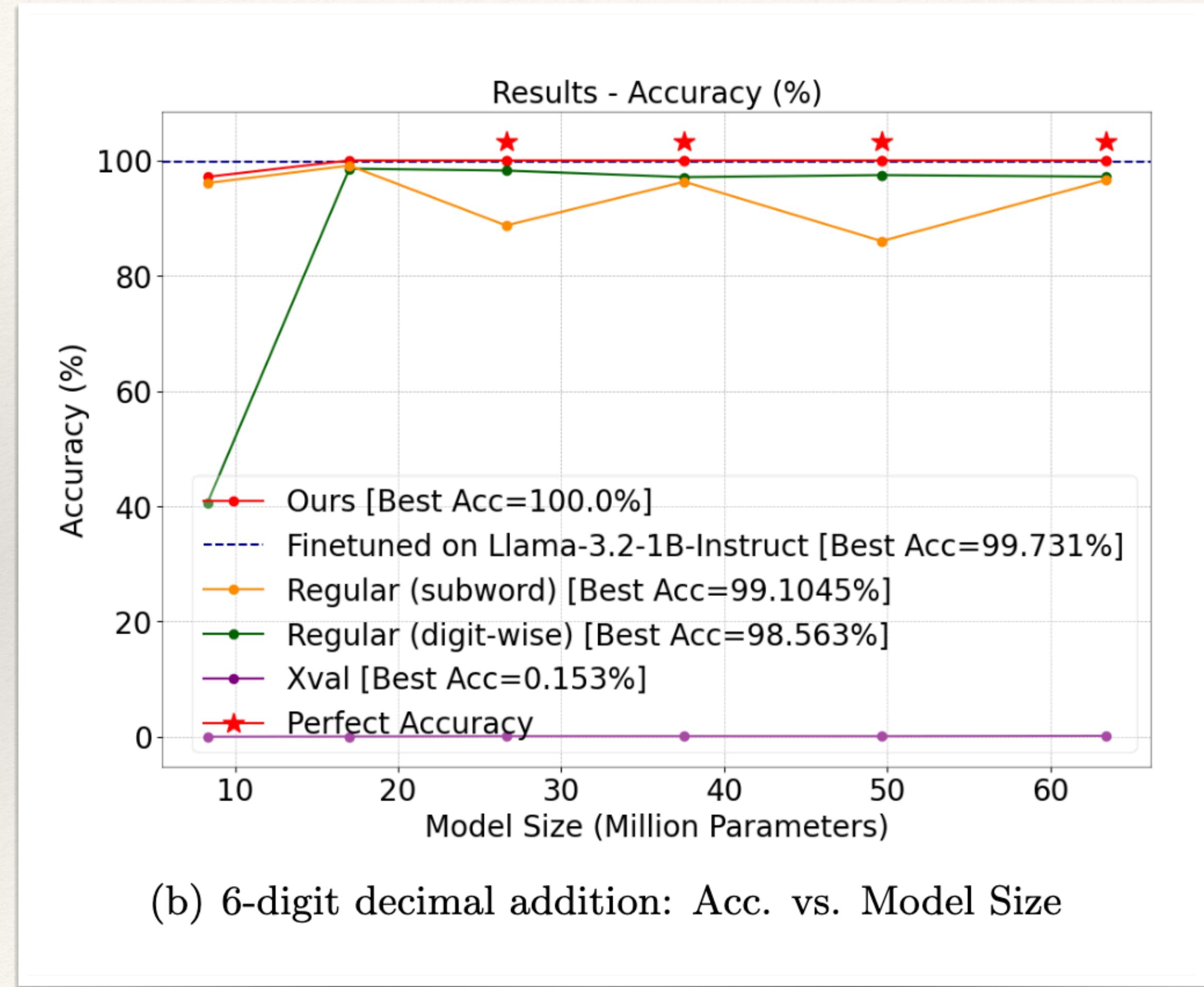
FoNE Enables Efficient Arithmetic Training

- ❖ We train a Transformer model with only 38M parameters with FoNE number embedding.
- ❖ Task: 6-digit decimal addition.
- ❖ FoNE can achieve 100% accuracy.
- ❖ To achieve >99% accuracy, FoNE requires much less data. FoNE is **data efficient**.



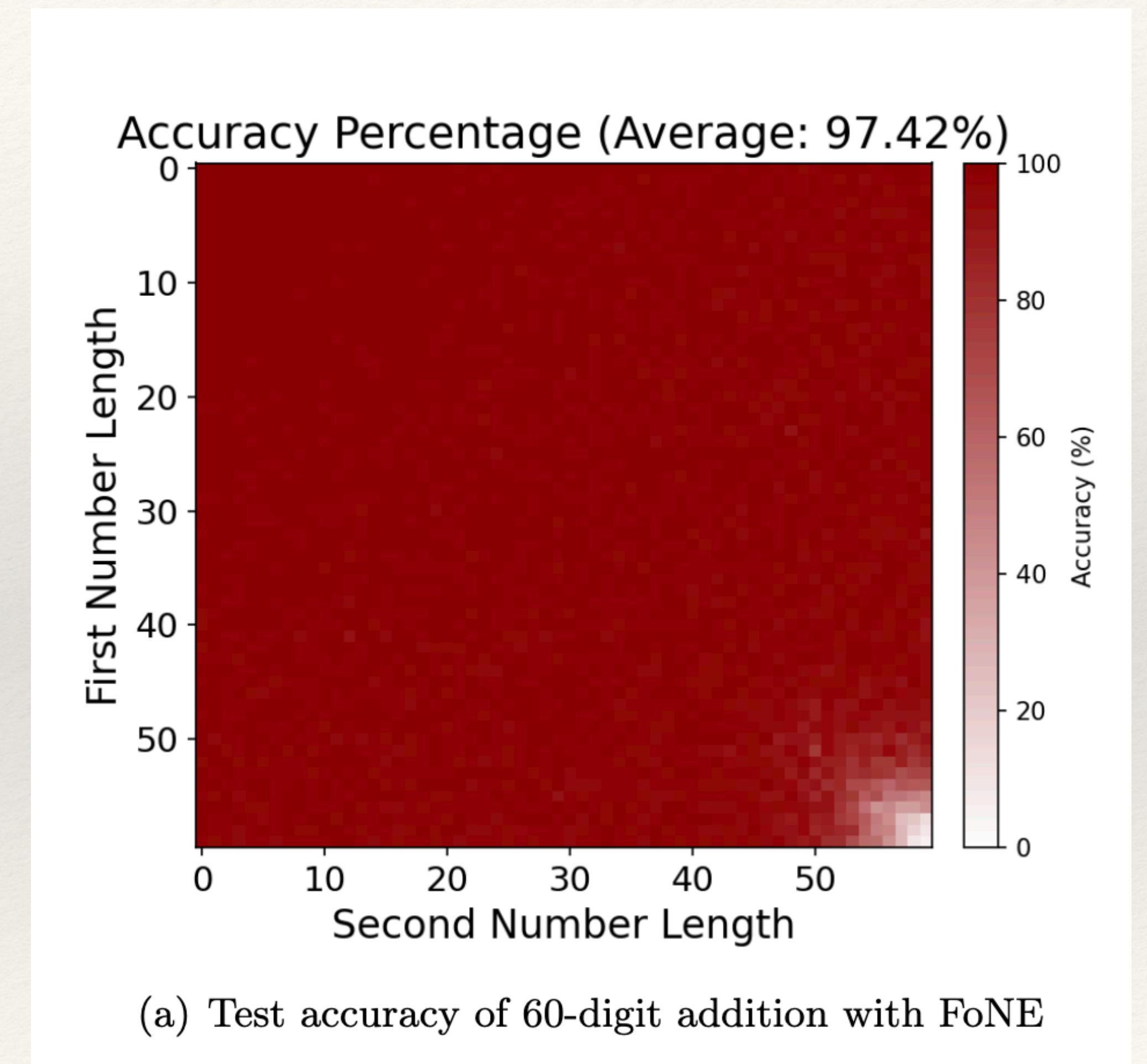
FoNE Enables Efficient Arithmetic Training

- ❖ We ablate # parameters of Transformers for training FoNE.
- ❖ FoNE is **parameter efficient**.
- ❖ We extend tasks to addition, subtraction, and multiplication. The conclusions are similar.



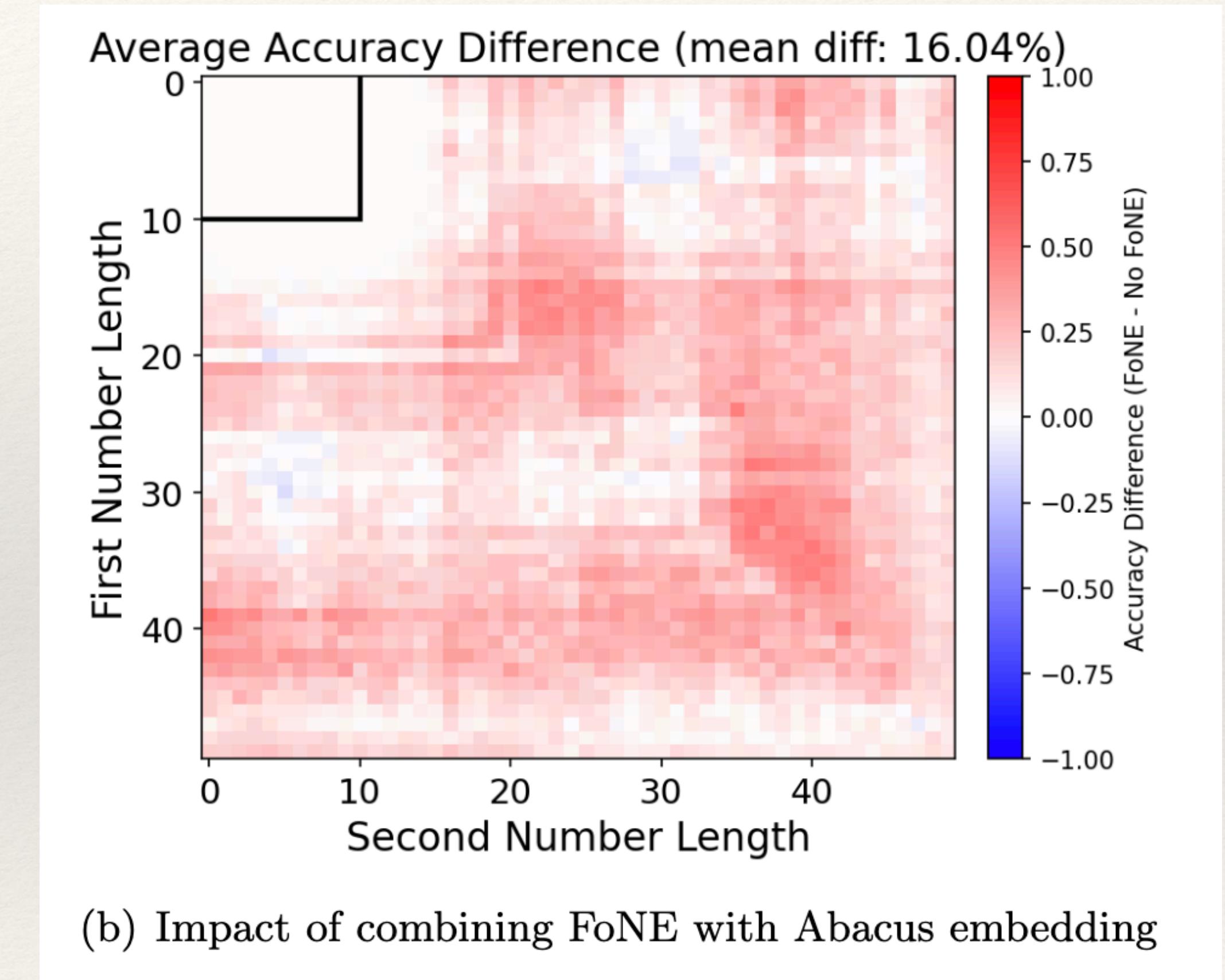
FoNE Can Handle Large Numbers

- ❖ Train and test on **60-digit addition**, FoNE could achieve **97%** accuracy on average.
- ❖ It demonstrates the effectiveness of FoNE in handling **long sequences**.



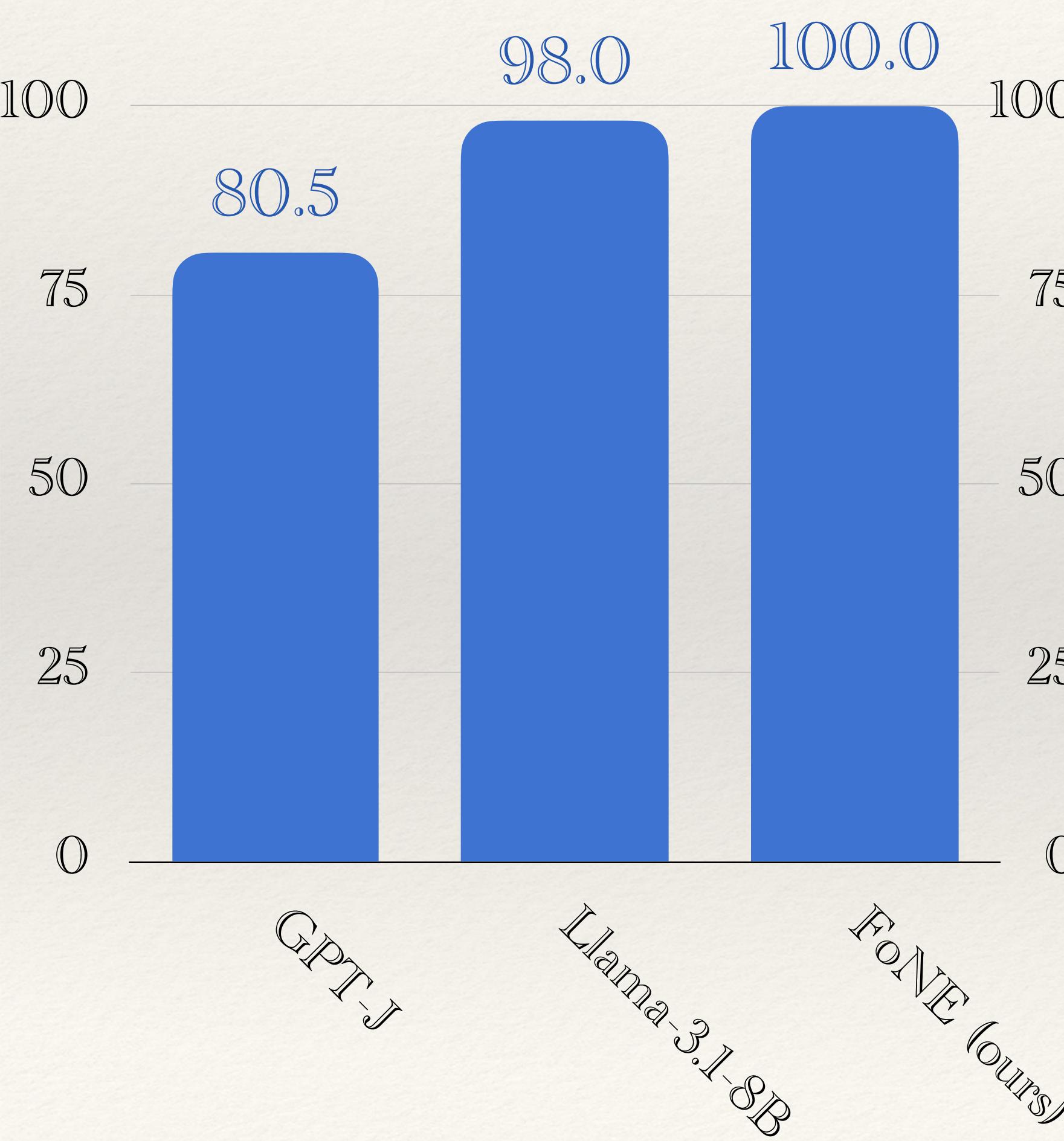
FoNE Enables Better Length Generalization

- ❖ We integrate FoNE with the Abacus embedding method [MBSJ+24], which operates on digit-wise tokenization.
- ❖ In this setup, the embeddings for each digit (0–9) are replaced with their corresponding Fourier Number Embeddings.
- ❖ Complementary benefits of combining FoNE with other positional embedding strategies.

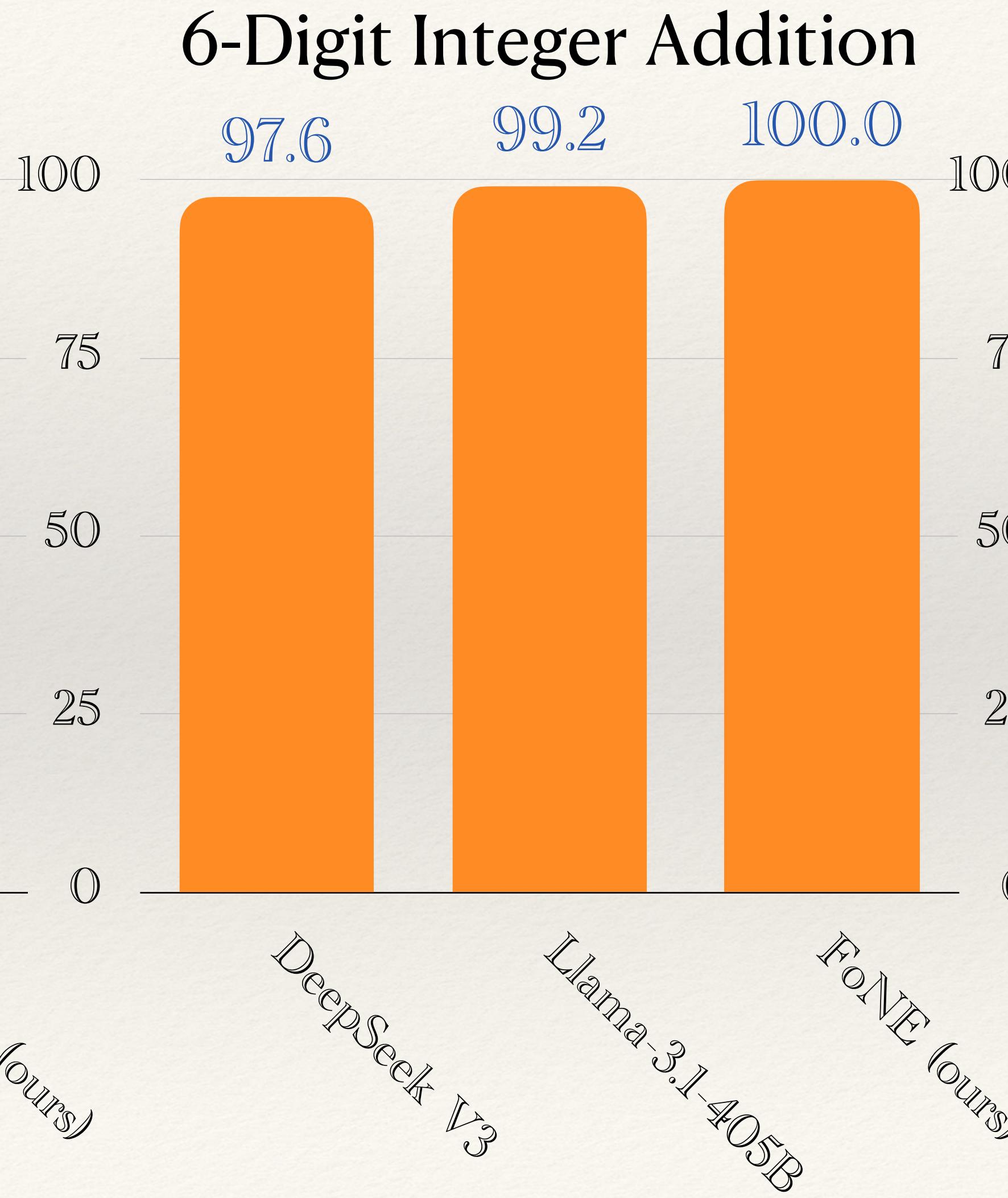


FoNE Could Fix Tokenization Issues for Numbers

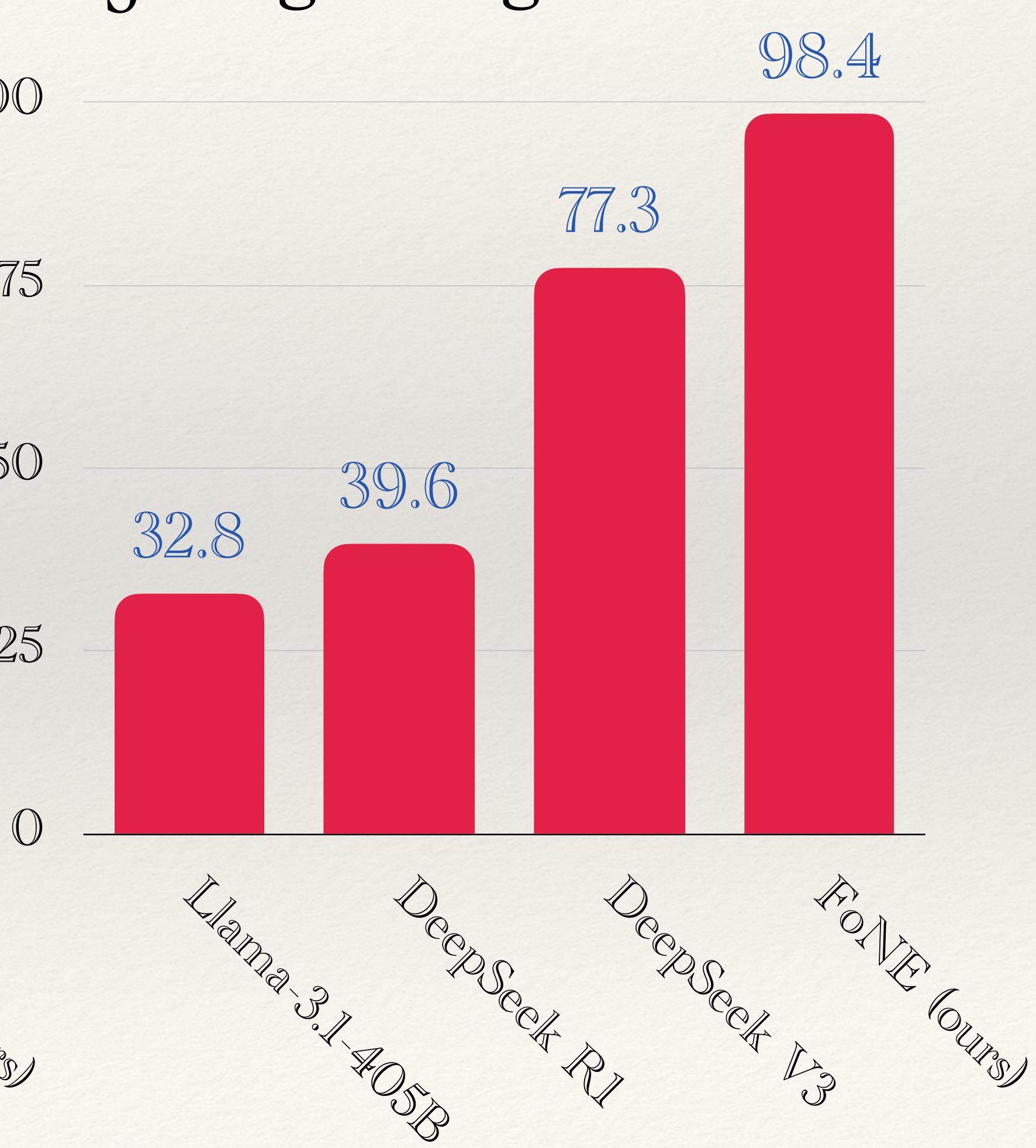
2-Digit Integer Addition



6-Digit Integer Addition



50-Digit Integer Addition

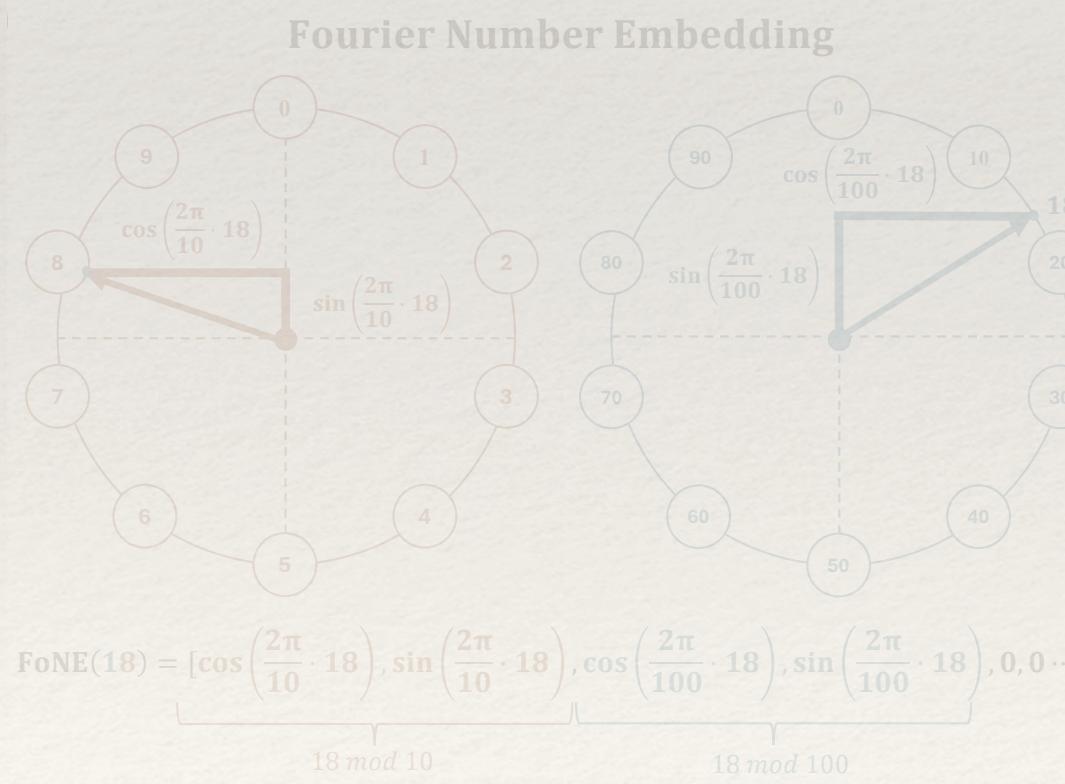


Conclusion

- ❖ Leveraging Fourier features, FoNE directly maps numbers into a compact and precise representation.
- ❖ FoNE bypasses tokenization — a commonly believed root cause for arithmetic failures.
- ❖ **Future Plans for FoNE:**
 - ❖ Training a foundation model with FoNE for numbers.
 - ❖ Unleashing FoNE's efficiency and precision for science and time-series tasks.

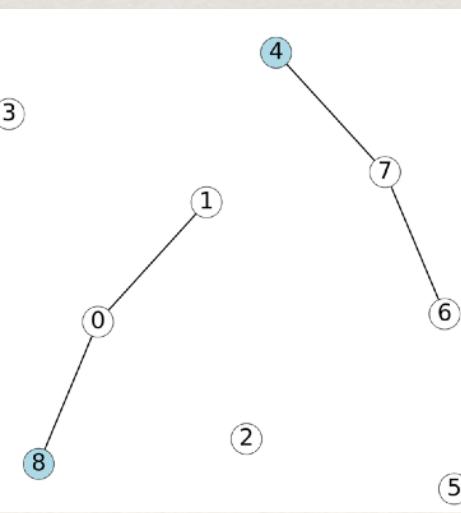
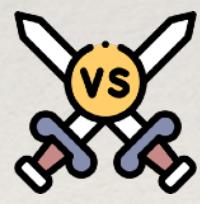
Overview

LLMs are Bad At
Arithmetics.
Can we find a family of
embeddings to bypass
tokenization?



Evaluating Multimodal
LLMs Beyond Ranking
Models.
Humans have pictorial
superiority. How about
MLLMs?

Adjacency Matrix:
$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
. Query Nodes: 4 and 8.



Text LLMs are Teachers for MLLMs.

Fine-Grained
Rewards.
Train Text Backbone
to be Token-Level
Reward Models.



MLLM

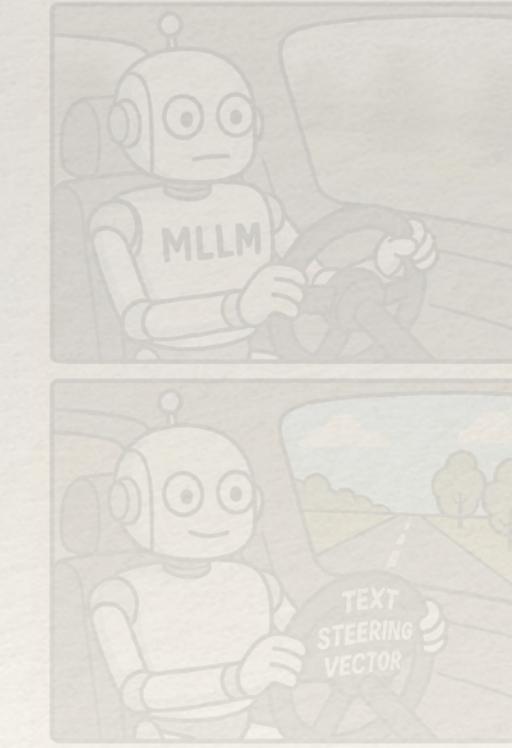
OOOXOOX
OXOXO



TLDL

OOOXOOX
OXOXO

Text Steering
Vectors.
Leverage Text-Based
Interpretability Tools
for MLLMs.



IsoBench: Benchmarking Multimodal Foundation Models on *Isomorphic Representations* [COLM 2024]



Deqing
Fu*



Ruohao
Guo*



Ghazal
Khalighinejad*



Ollie
Liu*



Bhuwan
Dhingra



Dani
Yogatama



Robin
Jia

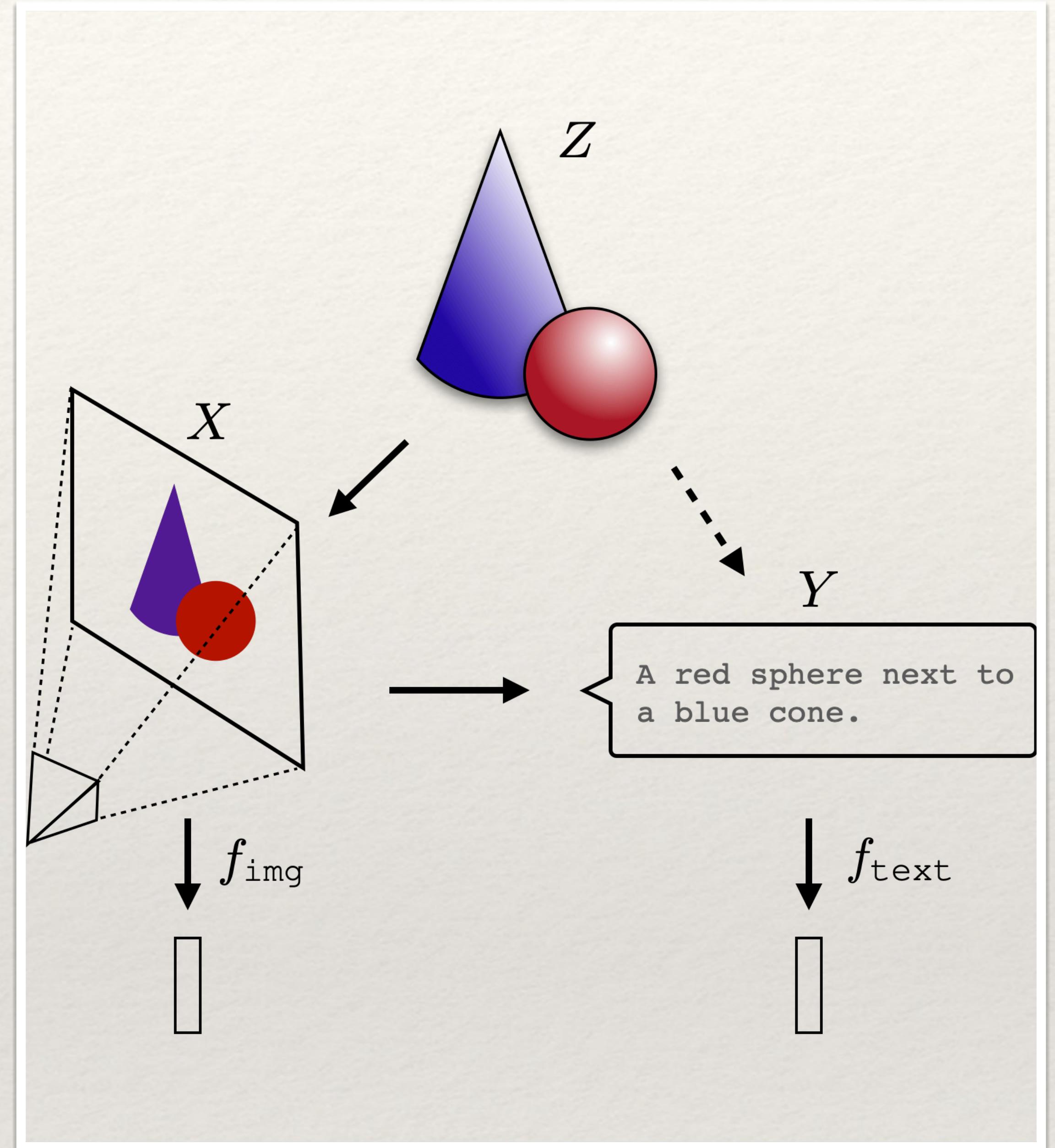


Willie
Neiswanger

In a hypothetical perfect world

The Platonic Representation Hypothesis

“Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.”

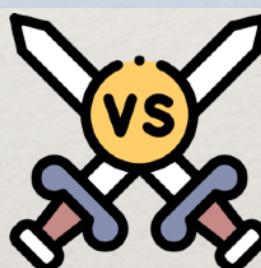


Prompts

Problem: Check Function Parity

<Definition of Function Parity>

Image Representation



Text Representation

Here is the expression $f(x)$:

$$f(x) = \frac{7x^6 - 3x^5 + x^2 - 21.76}{2x^6 + 4x^4 + 3x^2 + 11.34}$$

Responses

The plot of the function $f(x)$ shows symmetry about the y-axis. This symmetry indicates that $f(x) = f(-x)$ for all x in the domain of the function, which is the definition of an even function. Therefore, based on the plot, $f(x)$ is an even function.



Substituting $-x$ into the original function, we get

$$f(-x) = \frac{7(-x)^6 - 3(-x)^5 + (-x)^2 - 21.76}{2(-x)^6 + 4(-x)^4 + 3(-x)^2 + 11.34} = \frac{7x^6 + 3x^5 + x^2 - 21.76}{2x^6 + 4x^4 + 3x^2 + 11.34}$$

This expression is not equal to the original function $f(x)$, indicating that the function is not even. Additionally, the expression is not the negative of the original function, indicating that the function is not odd.

Therefore, $f(x)$ is neither even nor odd.



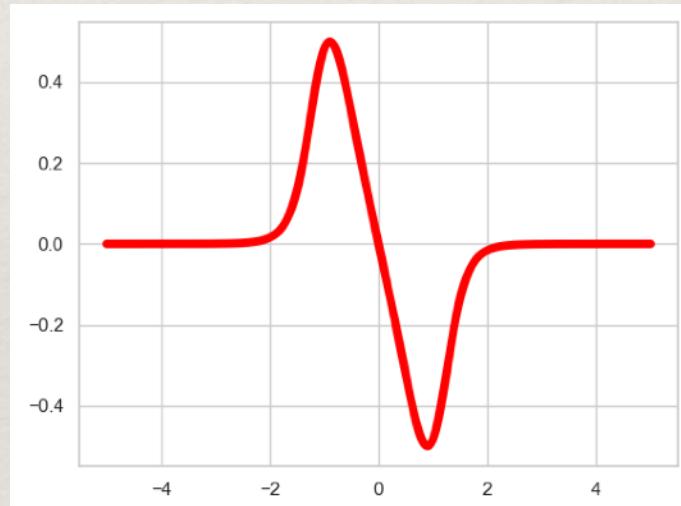
IsoBench Tasks

Task: Function Parity

Text Representation

$$f(x) = -\frac{18x}{2x^{10} + 16x^4 - 10x^2 + 29.34}$$

Image Representation



Mathematical Functions

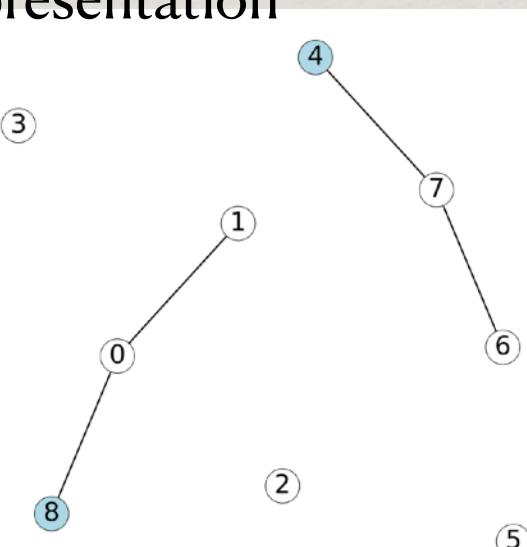
Graph Algorithms

Task: Connectivity

Text Representation

Adjacency Matrix:
$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
. Query Nodes: 4 and 8.

Image Representation



Answer: Not Connected

Task: Winner Identification

Text Representation

r . . . k .
p R . . . p p p
.
Q
P . P . p
. . P . K q . . .
. . . B . P . P .
. . . . R

Image Representation



Answer: Black Wins

Chess Games

Science Questions

Task: Physics QA

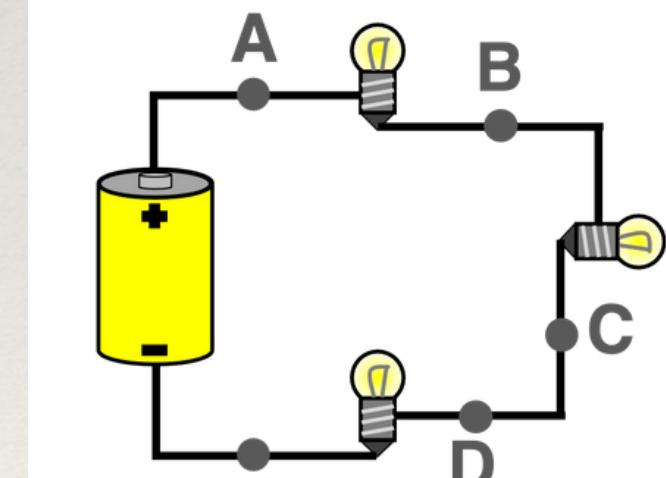
Text Representation

Question: Consider the given circuit. What is the current in amperes if 9.9 Coulombs of charge flow past point A in 1.1 seconds?

- (A) 6 (B) 7 (C) 8 (D) 9

Description: The circuit consists of a single battery connected to three light bulbs arranged in a parallel configuration. A, B, C, D, and E are points in the circuit. The battery provides the electrical energy that powers the light bulbs.

Image Representation

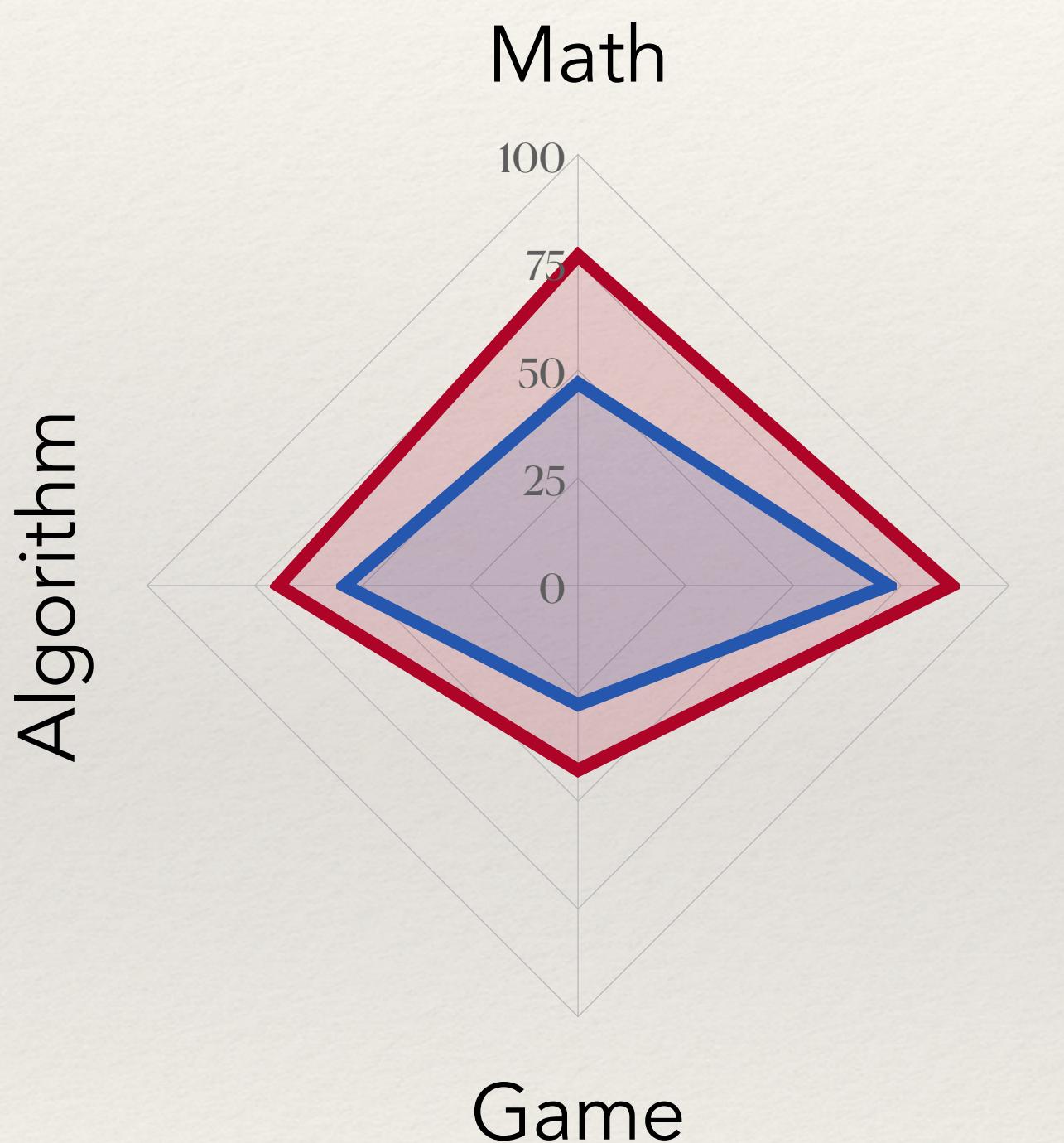


Answer:(D) 9 Amps

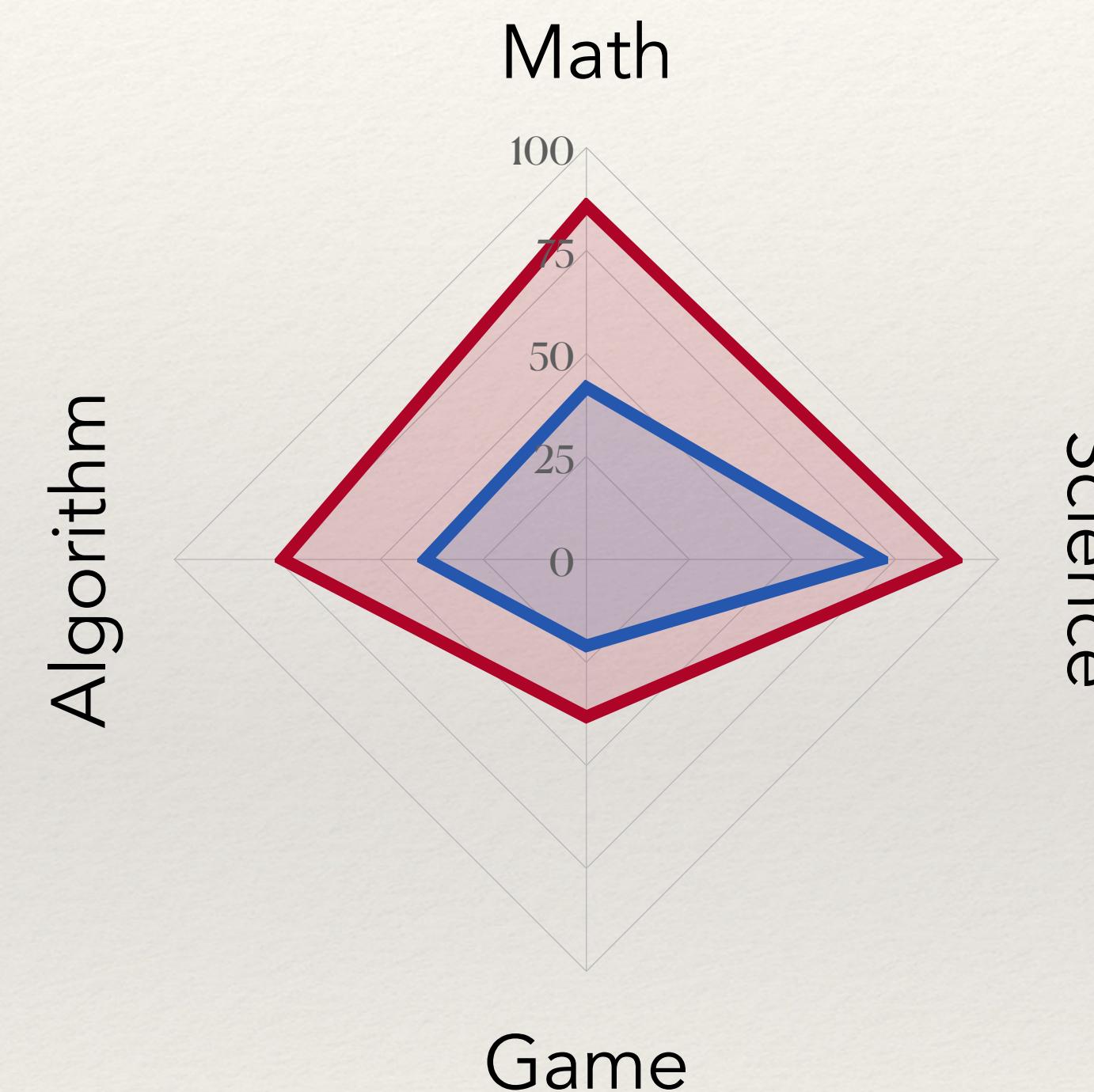
In total, IsoBench comprises 10 tasks, spanning these four sub-categories.

Significant Modality Gap

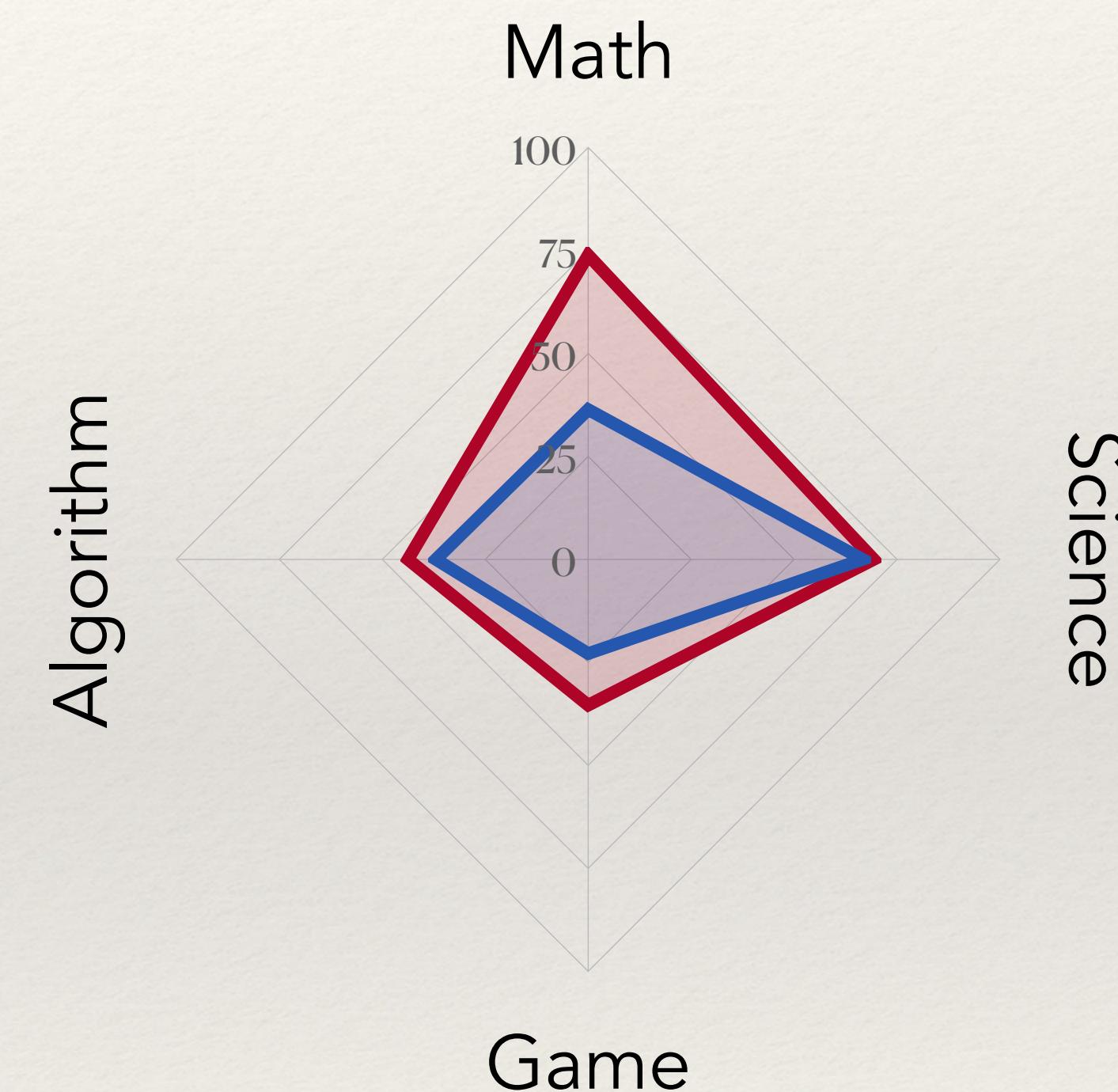
GPT-4-Turbo



Claude-3-Opus



Gemini-1.0-Pro



On average, GPT-4-Turbo has a **modality gap of 18.7**; Claude-3-Opus **14.9**; and Gemini-1.0-Pro **28.7**.

Image
 Text

Significant Modality Gap

Modality	Gemini 1.5				Claude				GPT			
	Pro		Flash		Opus		Sonnet		Haiku		4o	
	Perf.	Δ	Perf.	Δ	Perf.	Δ	Perf.	Δ	Perf.	Δ	Perf.	Δ
Text	77.5	-	69.9	-	77.7	-	77.4	-	72.5	-	71.5	-
Image	57.3	20.2↓	36.3	33.6↓	26.9	50.8↓	18.8	58.6↓	9.9	62.6↓	60.1	11.4↓
Audio	56.6	20.9↓	53.9	16.0↓	-	-	-	-	-	-	-	-
Video	36.3	41.2↓	15.1	54.8↓	-	-	-	-	-	-	53.1	18.4↓

Extract-Then-Answer (ETA) Prompting												
Image	73.5	4.0↓	68.1	1.8↓	62.6	15.1↓	48.1	29.3↓	43.2	29.3↓	66.7	4.8↓
Audio	69.9	7.6↓	63.6	6.3↓	-	-	-	-	-	-	-	-
Video	48.6	28.9↓	42.8	27.1↓	-	-	-	-	-	-	25.0	46.5↓

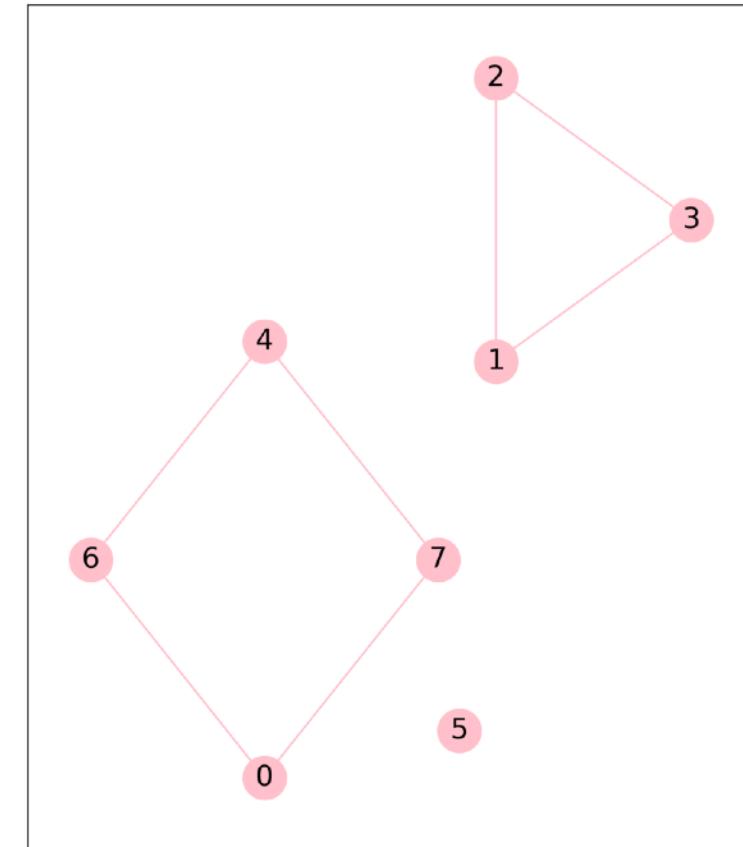
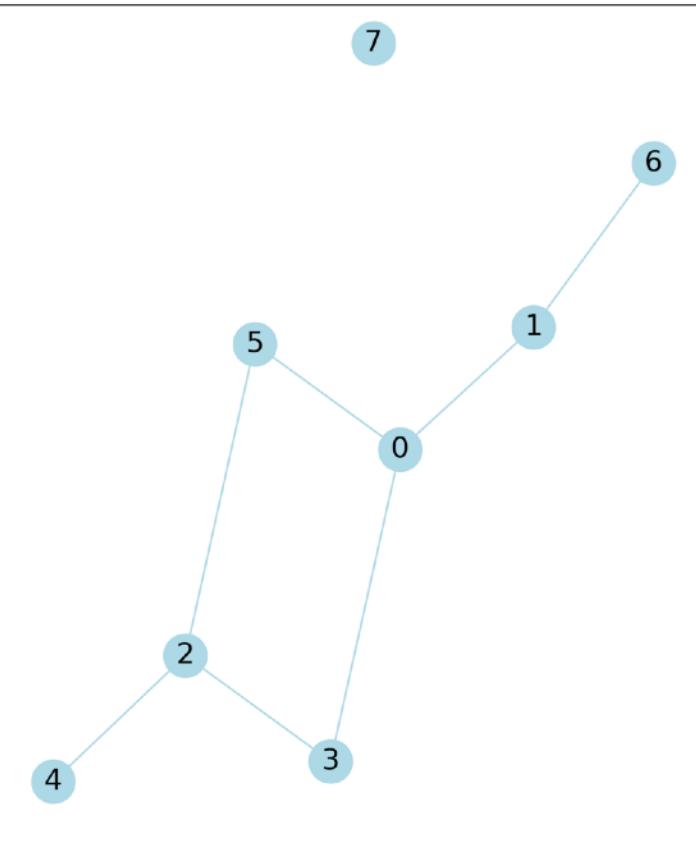
We can design *simple* and *deliberately contrived* mechanisms
to further study the performance of MLLMs
on tasks with *isomorphic representations*

IsoCombination

Instruction about the Task: Graph Isomorphism

Definition of Graph Isomorphism

Image Representation



Text Representation 1

Adjacency Matrices:

$$G = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Text Representation 2

Description of Graph G:

There are 8 nodes in the graph.
Node 0 is connected to Node 1 and Node 3
and Node 5.
[...]

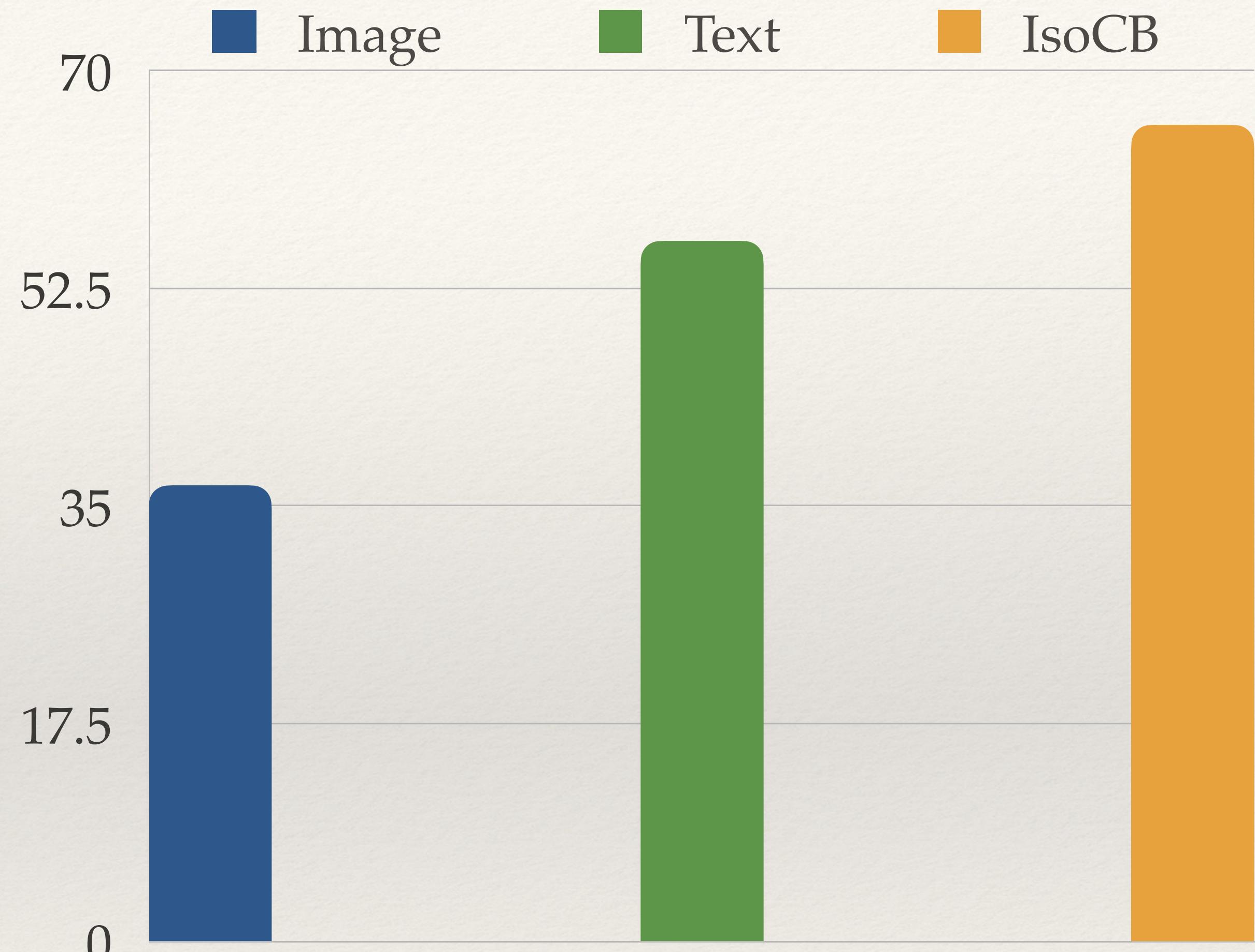
Description of Graph H:

There are 8 nodes in the graph.
Node 0 is connected to Node 6 and Node 7.
Node 1 is connected to Node 2 and Node 3.
[...]

Instruction on Output Format

IsoCombination

- Combining **Text** and **Image** inputs could improve the model significantly.
- Sometimes IsoCB is better than text-only inputs.
- However, it's “*cheating*” in the sense that the text inputs are given.



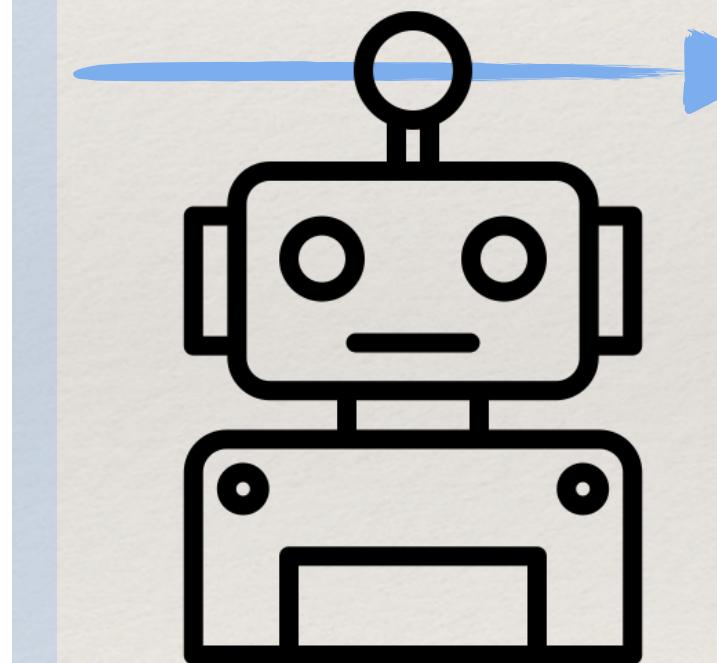
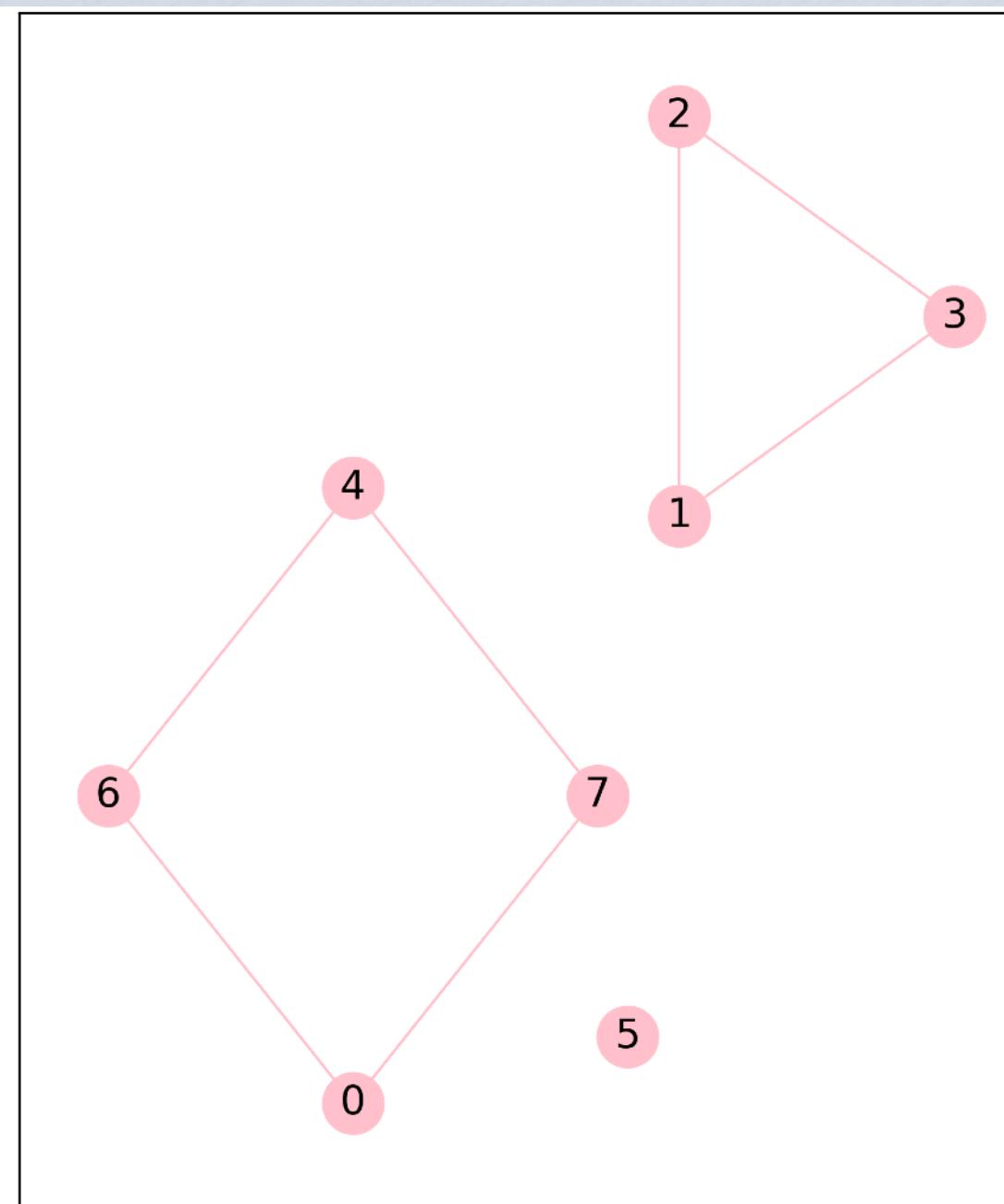
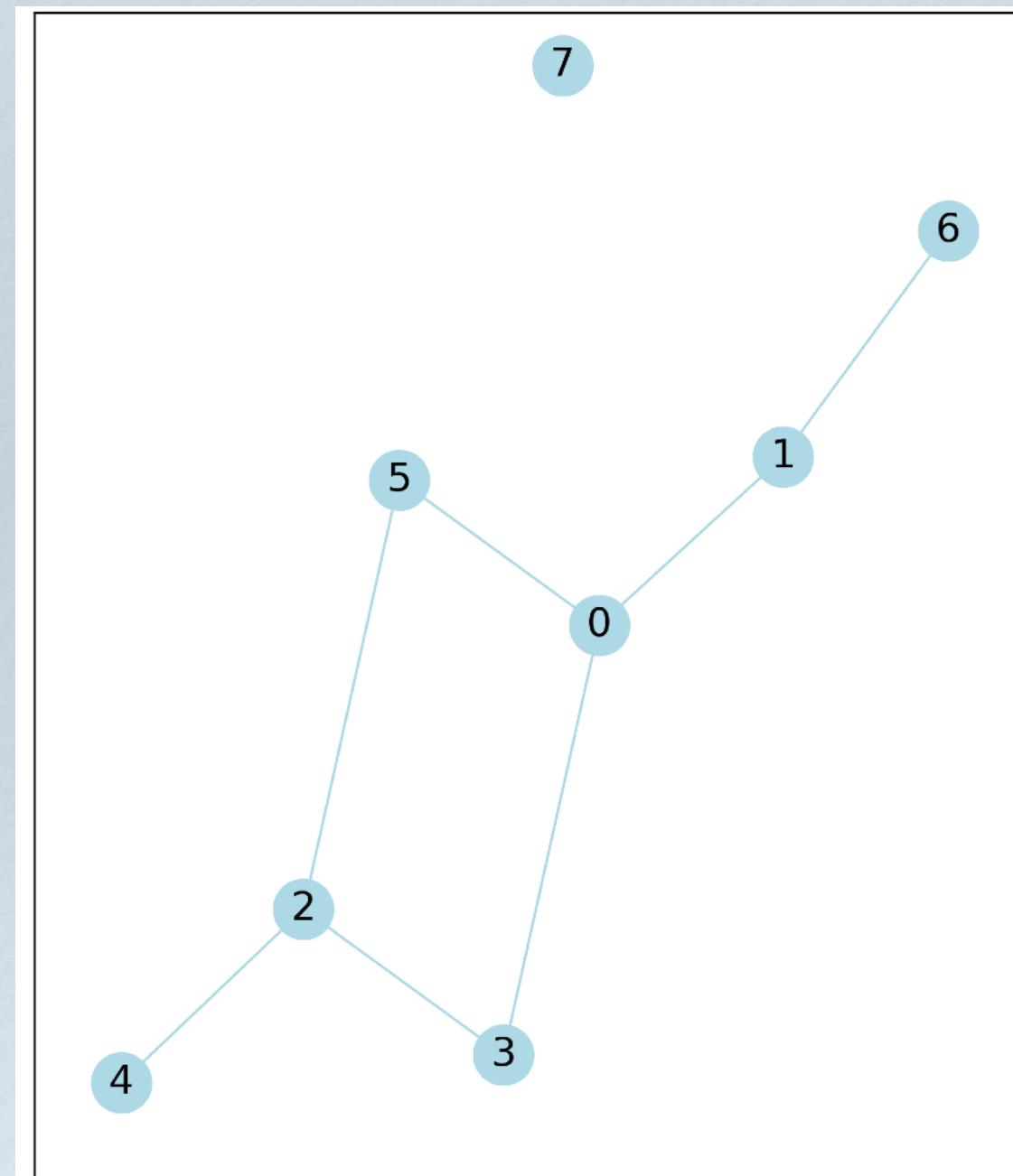
GPT-4-Turbo on MaxFlow Problems

IsoScratchPad

1

Instruction: Describe the Image

Image Representation



Model Scratch Pad

The image displays two graphs.

The Left Graph consists of eight nodes numbered 0 to 7, with the following connections:

- 0 is connected to 1, 3, and 5.
- 1 is connected to 6.
- 2 is connected to 3, 4 and 5.
- 7 is an isolated node.

The Right Graph consists of eight nodes numbered 0 to 7, with the following connections:

- 0 is connected to 6 and 7.
- 1 is connected to 2 and 3.
- 2 is connected to 3.
- 4 is connected to 6, 7.
- 5 is an isolated node.

IsoScratchPad

Instruction about the Task: Graph Isomorphism

Definition of Graph Isomorphism

Model Scratch Pad

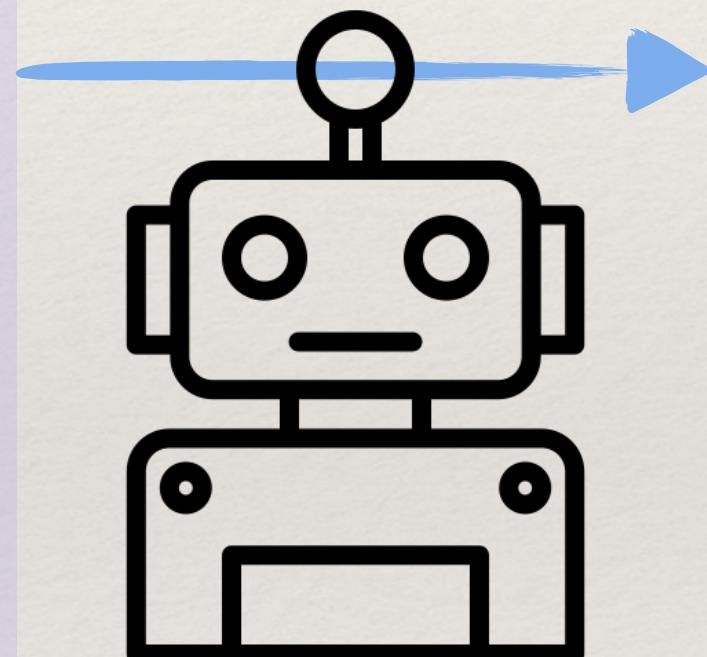
The image displays two graphs.

The Left Graph consists of eight nodes numbered 0 to 7, with the following connections:

- 0 is connected to 1, 3, and 5.
- 1 is connected to 6.
- 2 is connected to 3, 4 and 5.
- 7 is an isolated node.

The Right Graph consists of eight nodes numbered 0 to 7, with the following connections:

- 0 is connected to 6 and 7.
- 1 is connected to 2 and 3.
- 2 is connected to 3.
- 4 is connected to 6, 7.
- 5 is an isolated node.



Model Final Response

No, the two graphs G (Left Graph) and H (Right Graph) are not isomorphic.

Explanation:

In graph G, there are two nodes (0 and 2) that each have a degree of 3, meaning they are connected to three other nodes. However, in graph H, there is no node with a degree of 3. The nodes in graph H have the following degrees:

- Nodes 0, 1, 2, 3, 4, 6, 7 have a degree of 2
- Node 5 has a degree of 0

For two graphs to be isomorphic, they must have the same number of nodes with each degree. Since graph G has two nodes with a degree of 3 and graph H has no nodes with a degree of 3, the two graphs cannot be isomorphic.

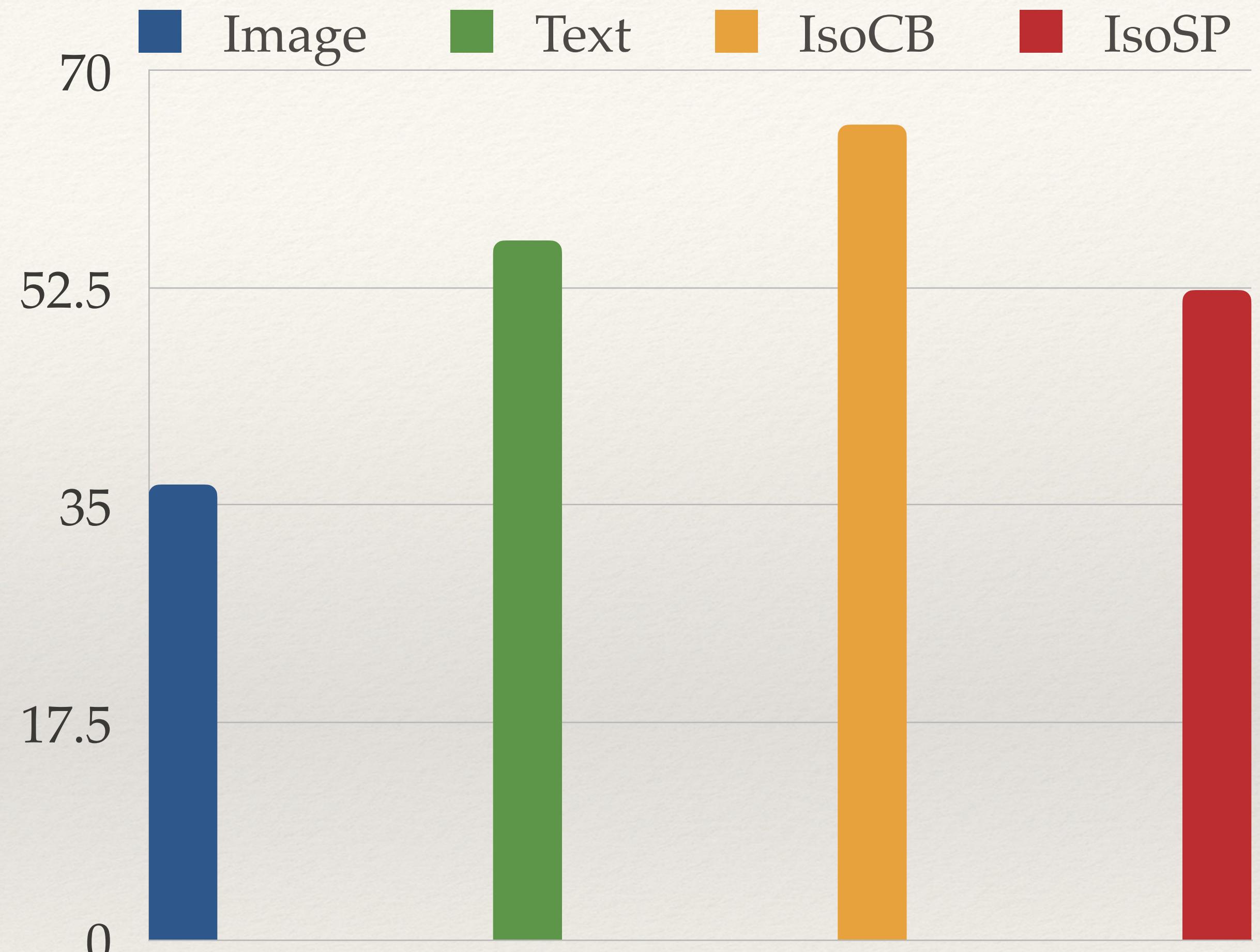
Furthermore, graph H has a node (Node 3) that is connected to two nodes (Node 1 and Node 2) which are also connected to each other, forming a triangle. In graph G, there is no such triangle structure.

These structural differences between the two graphs prove that they are **not isomorphic**, as there is no bijection that can preserve the edge relationships between the nodes while mapping one graph to the other.

Instruction on Output Format

IsoScratchPad

- Prompting MLLMs to generate texts as *scratchpad* could fill in the modality gap slightly.
- Due to model *hallucinations* in the modality transferring steps, IsoScratchPad is still worse than the text-only variant.



GPT-4-Turbo on MaxFlow Problems

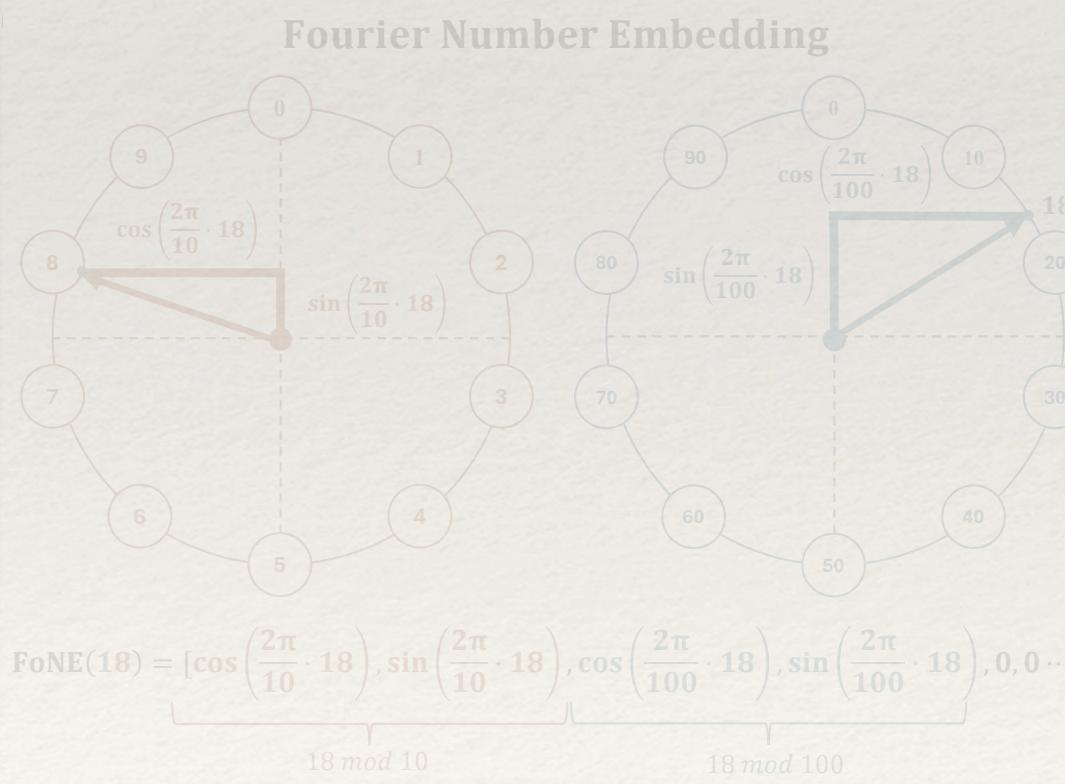
Conclusions

- ❖ Humans have pictorial superiority but MLLMs seem to have textual superiority.
- ❖ [Hu et al. *Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models*. NeurIPS 2024] Proposes tool usage to draw auxiliary lines to help MLLMs reason over images.
- ❖ [Li et al. *Zebra-CoT: A Dataset for Interleaved Vision Language Reasoning*. 2025] collects a dataset with 148,000 samples with interleaved text-image reasoning chains.
- ❖ Can we use MLLMs' textual superiority to improve their visual understanding?

Overview

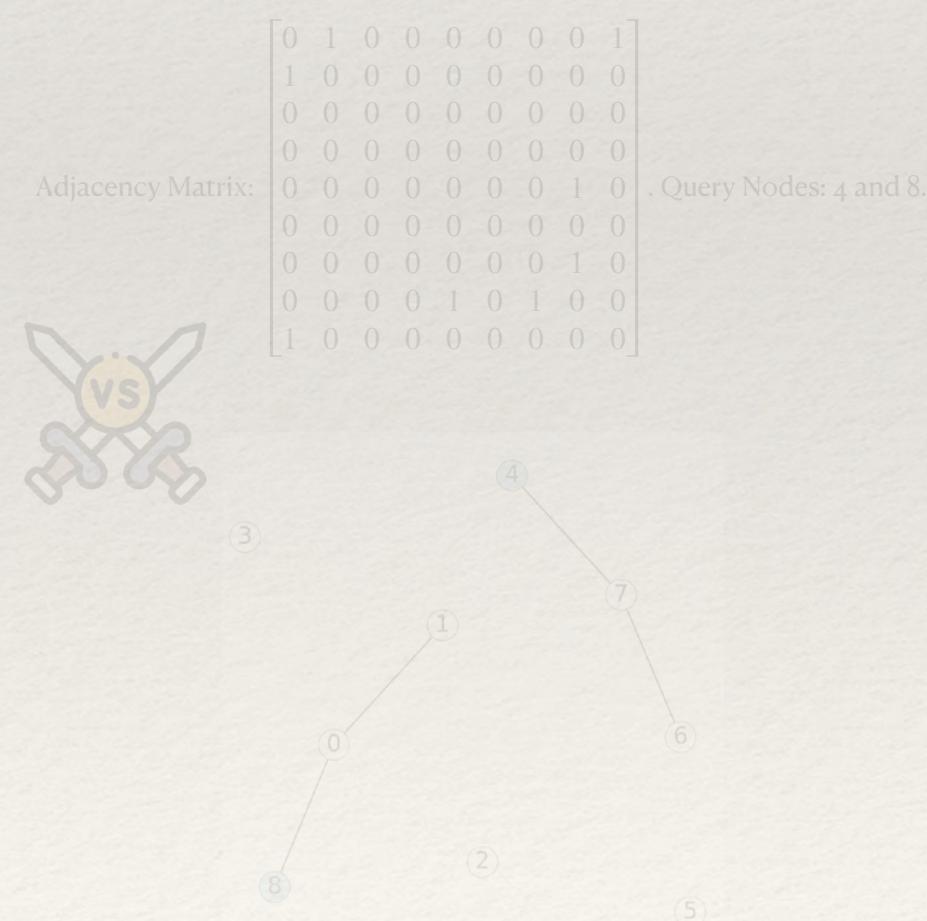
LLMs are Bad At Arithmetics.

Can we find a family of embeddings to bypass tokenization?



Evaluating Multimodal LLMs Beyond Ranking Models.

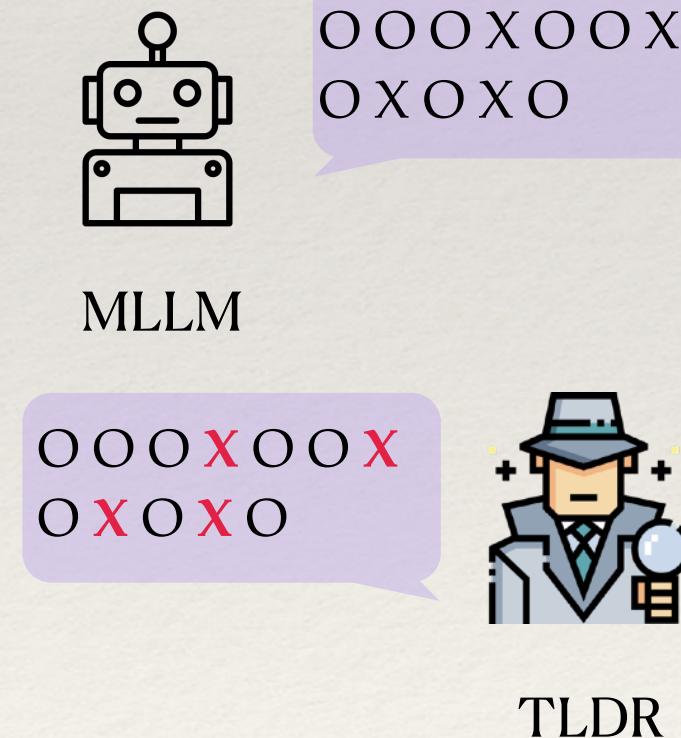
Humans have pictorial superiority. How about MLLMs?



Text LLMs are Teachers for MLLMs.

Fine-Grained Rewards.

Train Text Backbone to be Token-Level Reward Models.



Text Steering Vectors.

Leverage Text-Based Interpretability Tools for MLLMs.



TLDR: Token-Level Detective Reward Model for Large Vision Language Models

[ICLR 2025]



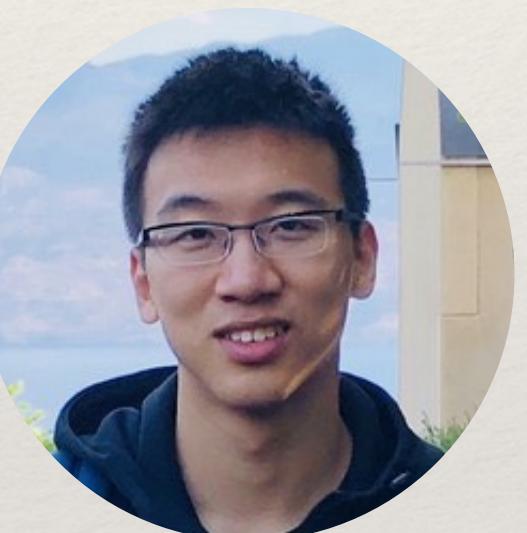
Deqing
Fu



Tong
Xiao



Rui
Wang



Wang
Zhu



Pengchuan
Zhang



Guan
Pang



Robin
Jia

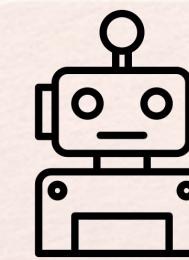


Lawrence
Chen

Multimodal LLMs Hallucinate over Images

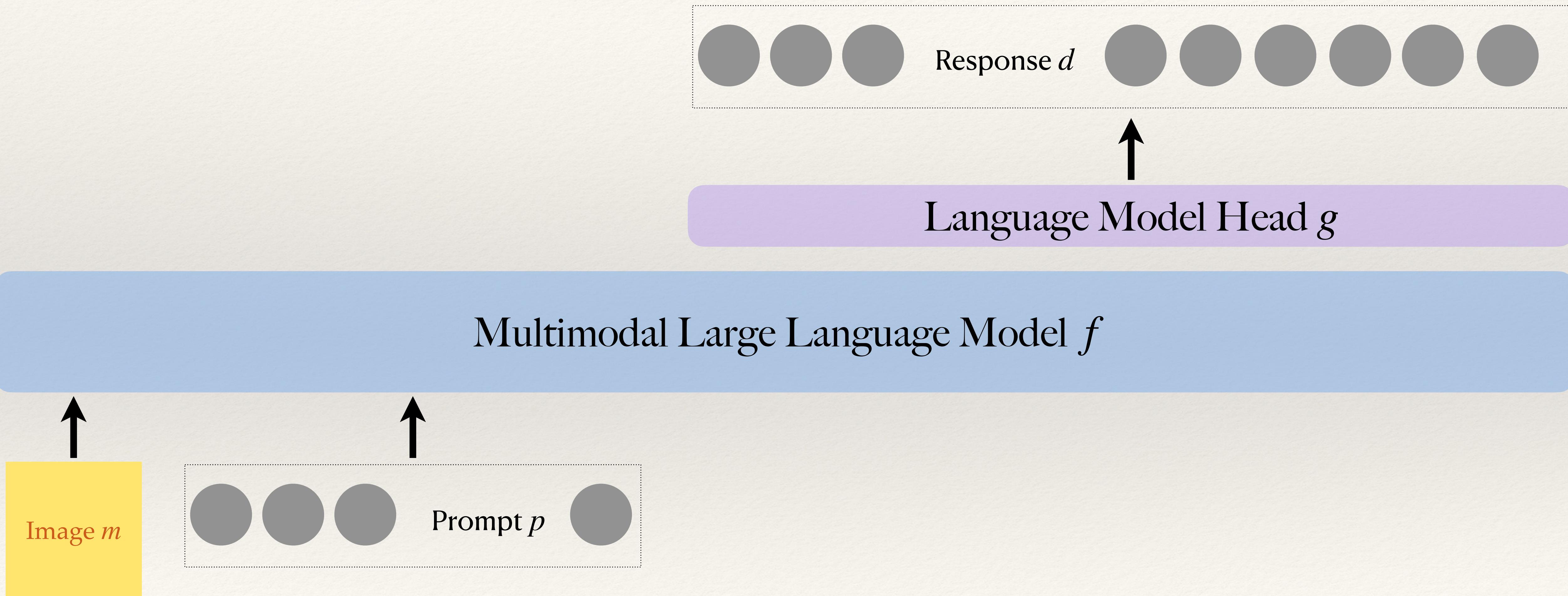


Describe the image in detail.

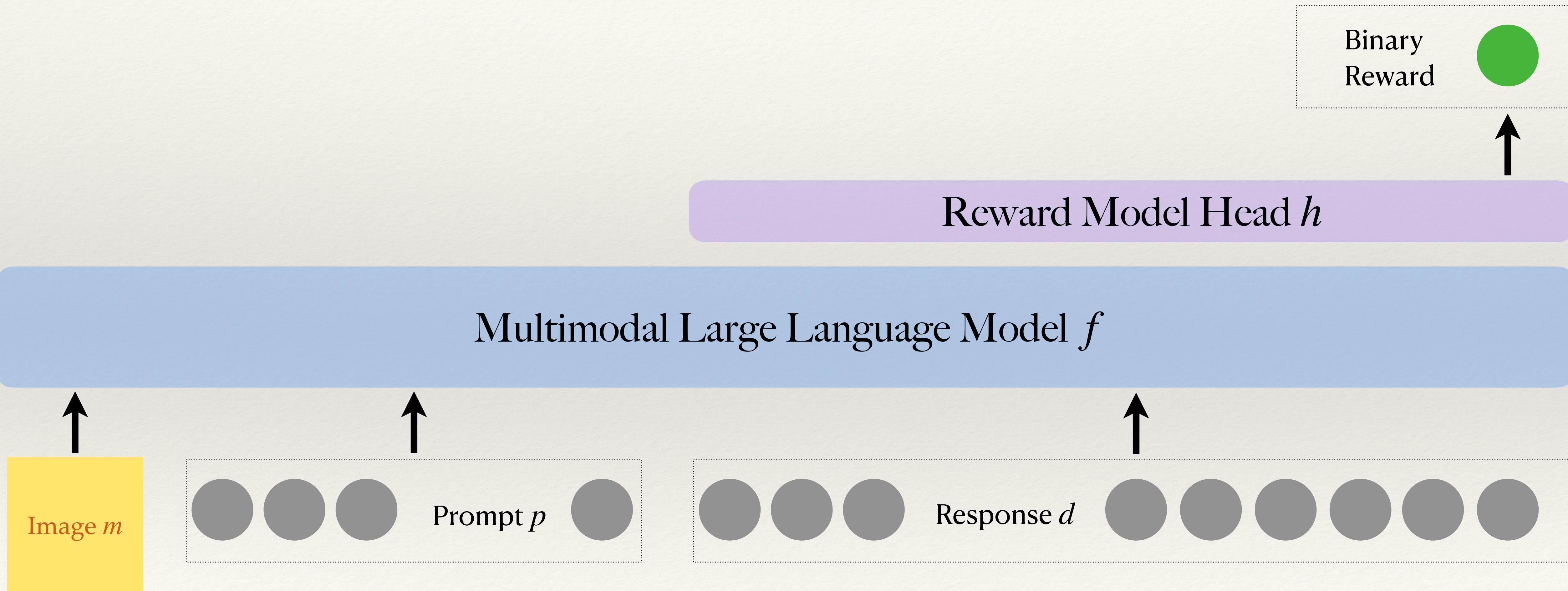


The image captures a bustling scene on a highway, where the flow of traffic is disrupted by a **construction** zone. The highway, a four-lane road, is filled with a variety of vehicles, including cars, trucks, and **buses**. The cars are predominantly black and white, while the trucks are white and red. The **buses**, painted in shades of **blue** and white, add a splash of color to the scene. The **construction** zone, marked by **orange cones**, is located on the right side of the highway. It's a busy area with several **construction vehicles and workers** diligently at work. The highway is surrounded by lush green fields and trees, providing a stark contrast to the concrete structures. The sky above is a clear blue, dotted with fluffy white clouds. The image is taken from an aerial perspective, providing a comprehensive view of the highway and its surroundings. This perspective allows for a detailed analysis of the traffic flow and the ongoing **construction** work.

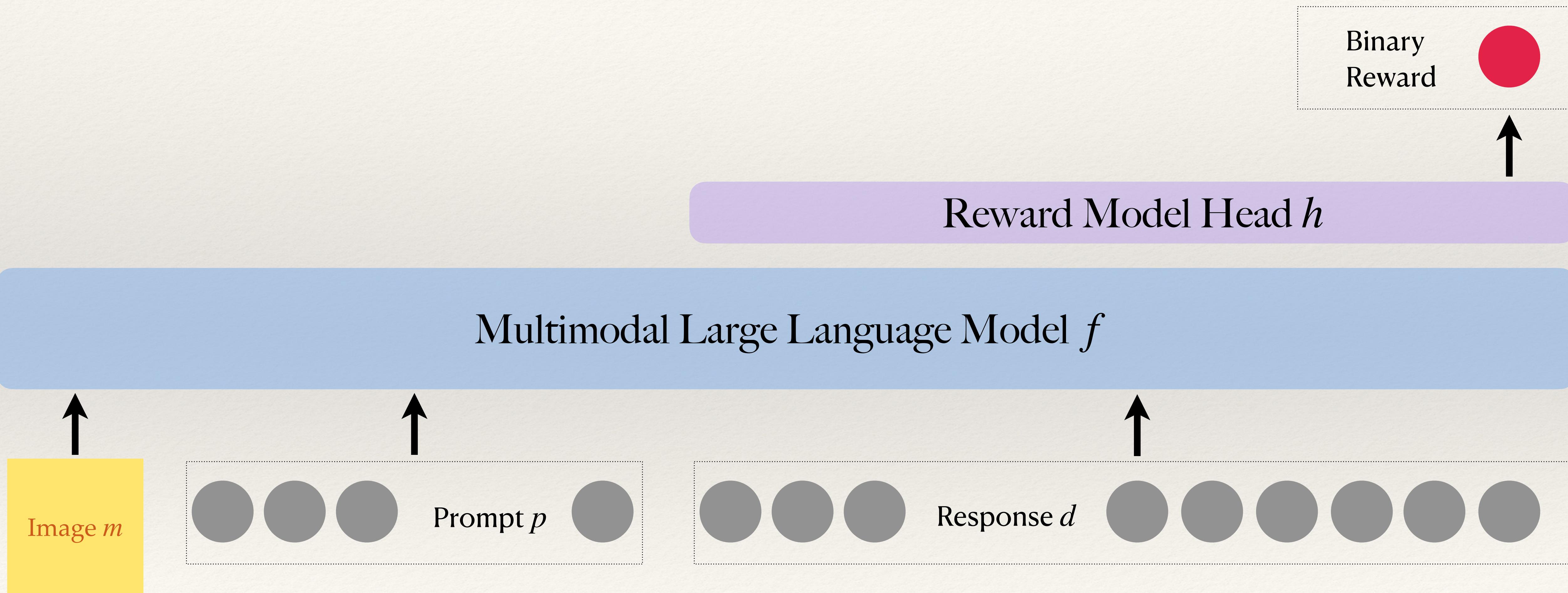
Let's Build Models to Detect Hallucinations



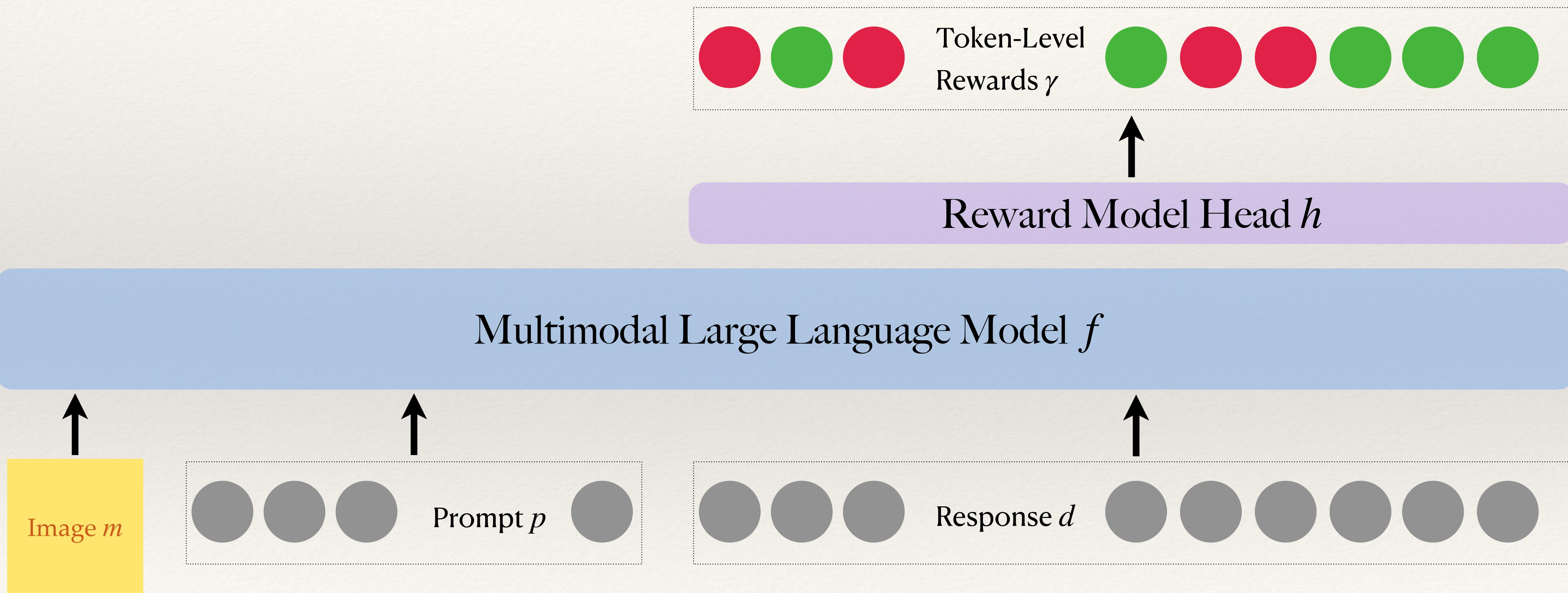
Let's Build Models to Detect Hallucinations



Let's Build Models to Detect Hallucinations



Let's Build Models to Detect Hallucinations



Synthetic Data Collection

- ❖ I. *Spatial Relationship*: A is **left** to B. \leftrightarrow A is **right** to B.
- ❖ II. *Visual Attribute*: A is **yellow**. \leftrightarrow A is **blue**.
- ❖ III. *Attribute Binding*: A is **blue** and B is **yellow**. \leftrightarrow A is **yellow** and B is **blue**.
- ❖ IV. *Object Identification*: A **dog** chasing a ball. \leftrightarrow A **cat** chasing a ball.
- ❖ V. *Counting*: **One** duck is swimming. \leftrightarrow **Four** ducks are swimming.
- ❖ VI. *Small Object*: **Cirrostratus cloud** in the sky. \leftrightarrow **Clear** sky.
- ❖ VII. *Text OCR*: A shirt writes **heavy fog**. \leftrightarrow A shirt writes **happy frog**.
- ❖ VIII. *Counterfactual*: A soldier. \leftrightarrow A soldier has **no sword** in hand.

Synthetic Data Collection

- ❖ Two datasets:
 - ❖ Visual Genome (VG100K)
 - ❖ Visual Question Answering (VQA)
 - ❖ Image Captioning by Grouping VQA to Captions
 - ❖ DOCCI for dense image captions
 - ❖ Synthetic Generation Steps:
 - ❖ Prompt LLMs (Llama-3.1-70B) to generate wrong answers/captions
 - ❖ Model *never sees the image* (text-only perturbation)
 - ❖ Filter out paraphrases, unnatural perturbations, etc.
 - ❖ Compare to gold-standard to obtain token-level labels

TASK	DATA SOURCE	TAXONOMY	# POSITIVE	# NEGATIVE	TRAIN SET PROPORTION (%)
VQA	VG100K	—	1,179,007	1,179,007	
Image Caption	Synthetic Caption from VG100K	Spatial Relation Visual Attribute Attribute Binding Object Identification Counting Small Object Text OCR Counterfactual	94,684	45,225 86,366 59,219 75,328 75,156 80,455 84,164 57,153	80%
Image Caption	DOCCI	Spatial Relation Visual Attribute Attribute Binding Object Identification Counting Small Object Text OCR Counterfactual	14,639	8,867 13,811 13,561 10,618 10,491 11,680 13,366 12,844	65%

TLDR as Hallucination Evaluation Tool

The image captures a bustling scene on a highway, where the flow of traffic is disrupted by a **construction** zone.

The highway, a **four**-lane road, is filled with a variety of vehicles, including cars, trucks, and **buses**.

The cars are predominantly black and white, while the trucks are white and red.

The **buses**, painted in shades of **blue** and white, add a splash of color to the scene.

The **construction** zone, marked by **orange cones**, is located on the right side of the highway.

It's a busy area with several **construction vehicles and workers** diligently at work.

The highway is surrounded by lush green fields and trees, providing a stark contrast to the concrete structures.

The sky above is a clear blue, dotted with fluffy white clouds.

The image is taken from an aerial perspective, providing a comprehensive view of the highway and its surroundings.

This perspective allows for a detailed analysis of the traffic flow and the ongoing **construction** work.

Hallucination Rate (%)

Token-Level

Tokens: **208**

Bad Tokens: **15**

Token-Level Rate:

$$\frac{15}{208} \approx 0.072\%$$

Sentence-Level

Sentences: **10**

Bad Sentences: **6**

Sentence-Level Rate:

$$\frac{6}{10} \approx 60\%$$

Response-Level

This is a **bad** response.

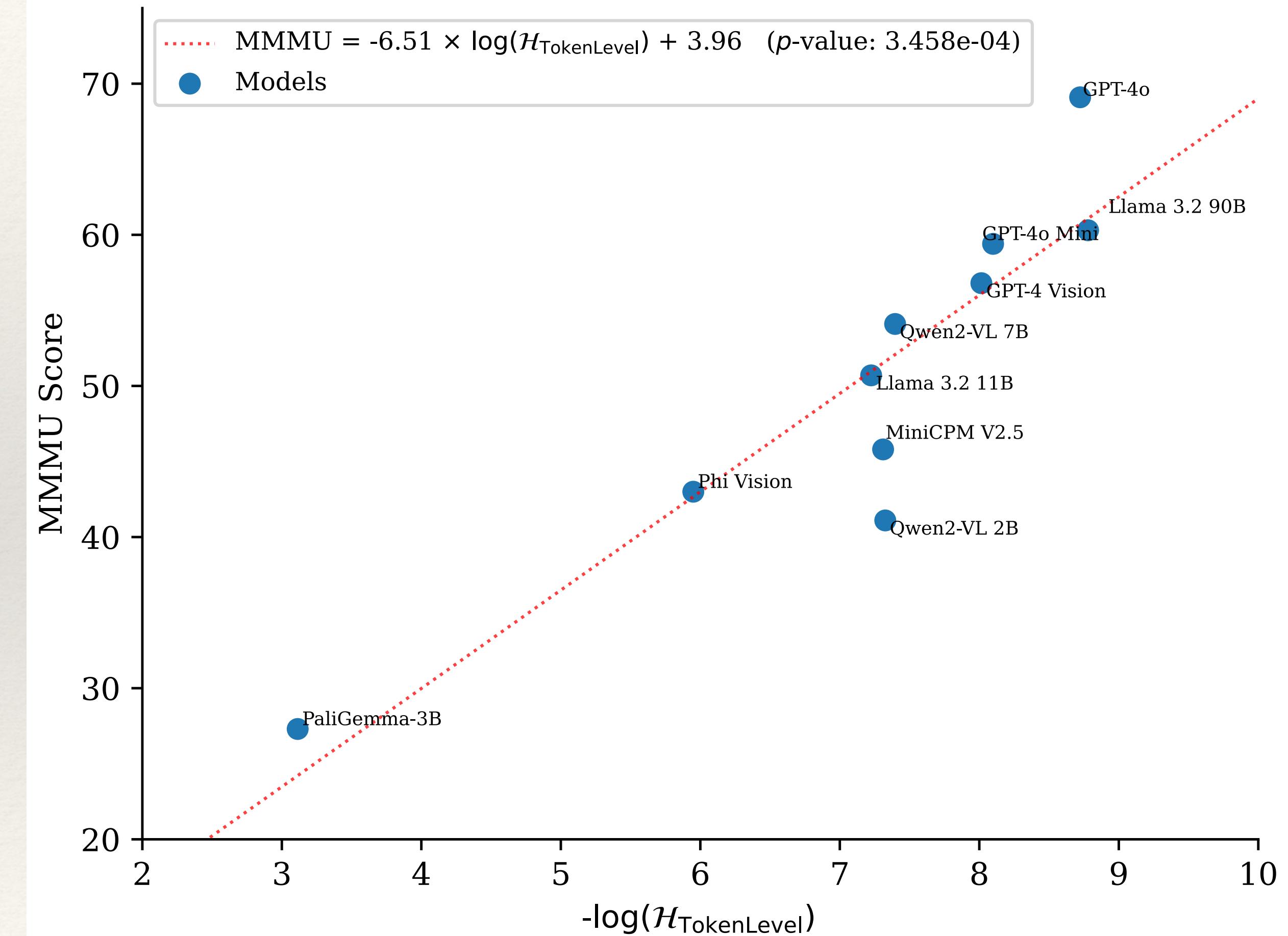
TLDR as Hallucination Evaluation Tool

Model	Token-Level (%)	Sentence-Level (%)	Response-Level (%)	MMMU	MEGA-Bench
<i>GPT-4o</i>	0.016	0.23	1.62	69.1	54.1
<i>Llama-3.2-90B-Vision</i>	0.017	0.19	1.23	60.3	/
<i>GPT-4o-mini</i>	0.030	0.38	2.12	59.4	43.0
<i>GPT-4-Turbo-Vision</i>	0.033	0.62	3.12	56.8	/
<i>Qwen2-VL-7B</i>	0.061	0.48	1.96	54.1	35.9
<i>Qwen2-VL-2B</i>	0.066	0.72	1.70	41.1	22.3
<i>MiniCPM-Llama-3-V2.5</i>	0.067	0.81	3.62	45.8	22.8
<i>Llama-3.2-11B-Vision</i>	0.073	0.85	1.88	50.7	18.0
<i>Phi-Vision-3.5-Instruct</i>	0.261	2.65	9.25	43.0	25.3
<i>PaliGemma-3B</i>	4.444	5.96	17.50	27.3	/

Conjecture

TLDR as Hallucination Evaluation Tool

$$\text{ModelPerformance} = -\log \mathcal{H}_{\text{TokenLevel}} + \text{Constant}$$



TLDR Guided Self-Correction

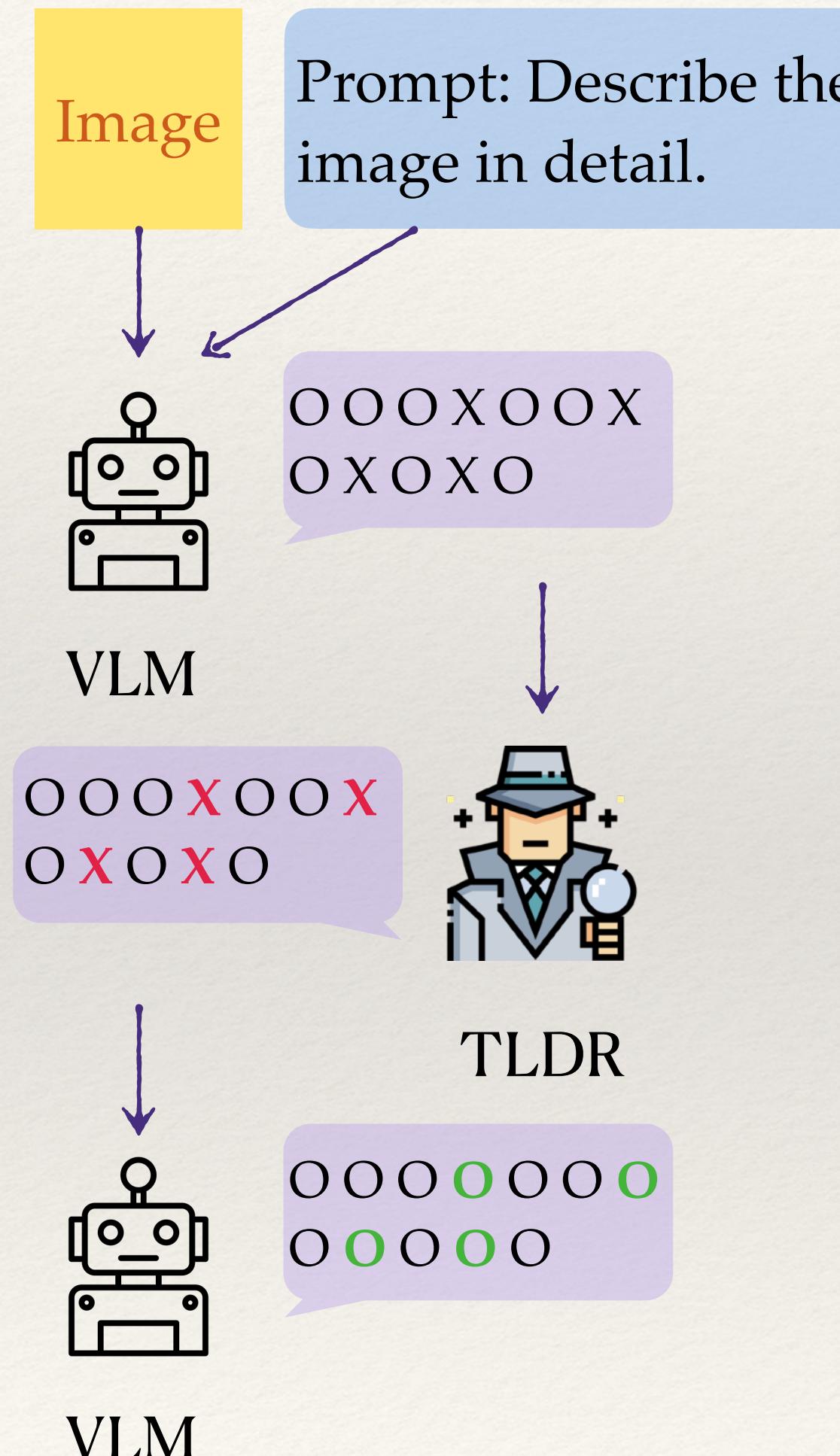
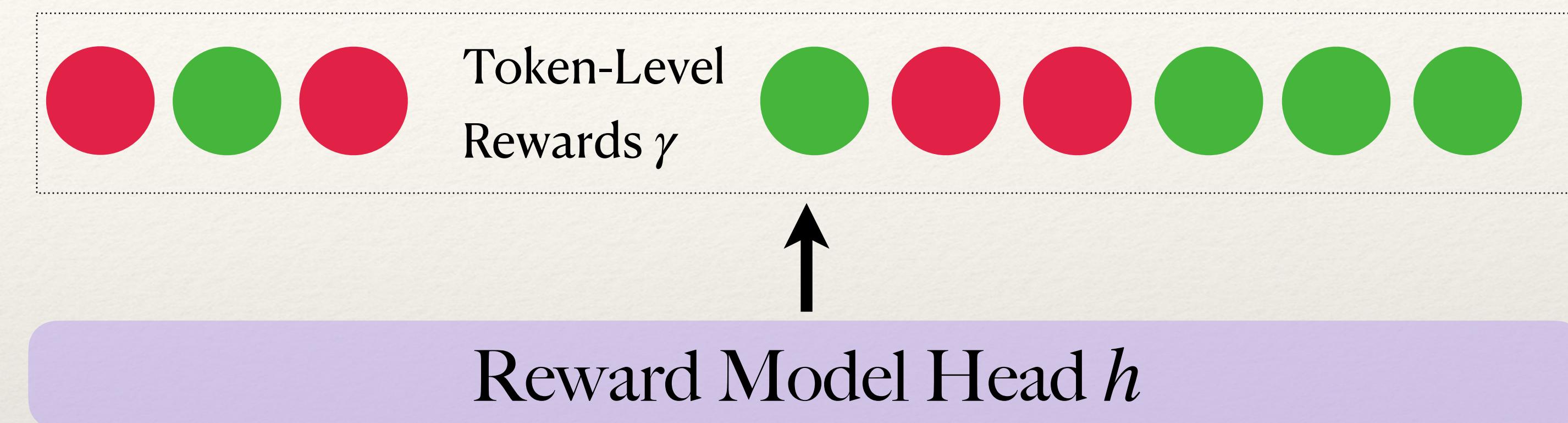
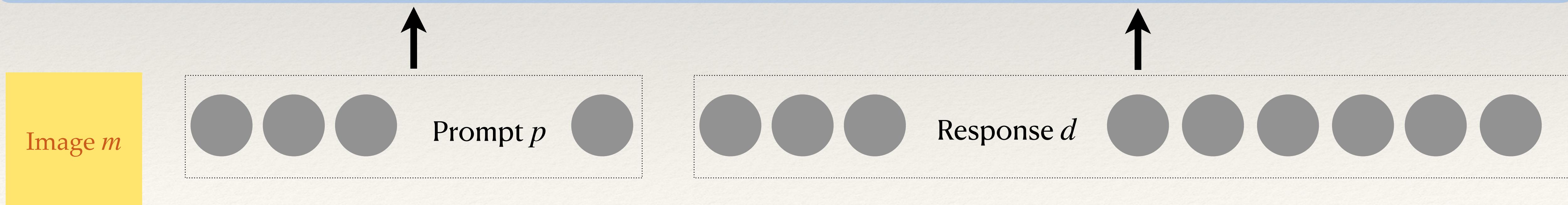


Image	GPT-4V Synthetic Caption Flagged by TLDR Model	GPT-4V Self-Correction with TLDR's Guidance
	<p>The image depicts a stack of six books, each varying in color and size, arranged in a slightly staggered formation. [...]</p>	<p>The image depicts a stack of five books, each varying in color and size, arranged in a slightly staggered formation. [...]</p>
	<p>[...] The dog's face is characterized by its deep wrinkles and big, soulful eyes, which give it a somewhat melancholic yet cute appearance. On top of its head, there are three biscuits stacked neatly, adding a playful and whimsical touch to the scene. [...]</p>	<p>[...] The dog's face is characterized by its deep wrinkles and big, soulful eyes, which give it a somewhat melancholic yet cute appearance. On top of its head, there are four biscuits stacked neatly, adding a playful and whimsical touch to the scene. [...]</p>
	<p>[...] Surrounding the map are essential travel items: a digital camera, suggesting the intention to capture memories; a pair of sunglasses, indicating sunny travel conditions; and a smartphone, likely used for additional navigation or communication. To the side, there's a cup of tea, providing a comforting element to the planning process. [...]</p>	<p>[...] Surrounding the map are essential travel items: a digital camera, suggesting the intention to capture memories; a pair of sunglasses, indicating sunny travel conditions; and a smartphone, likely used for additional navigation or communication. To the side, there's a cup of coffee, providing a comforting element to the planning process. [...]</p>

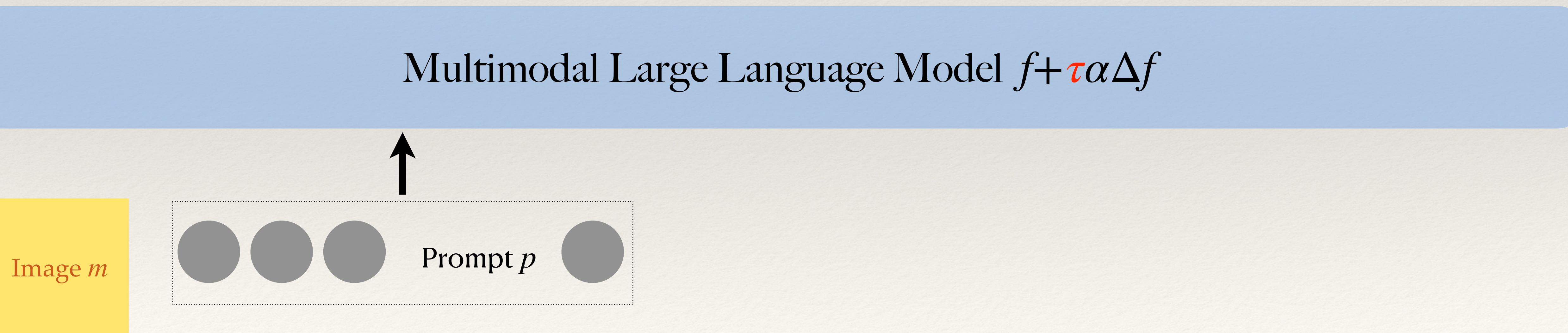
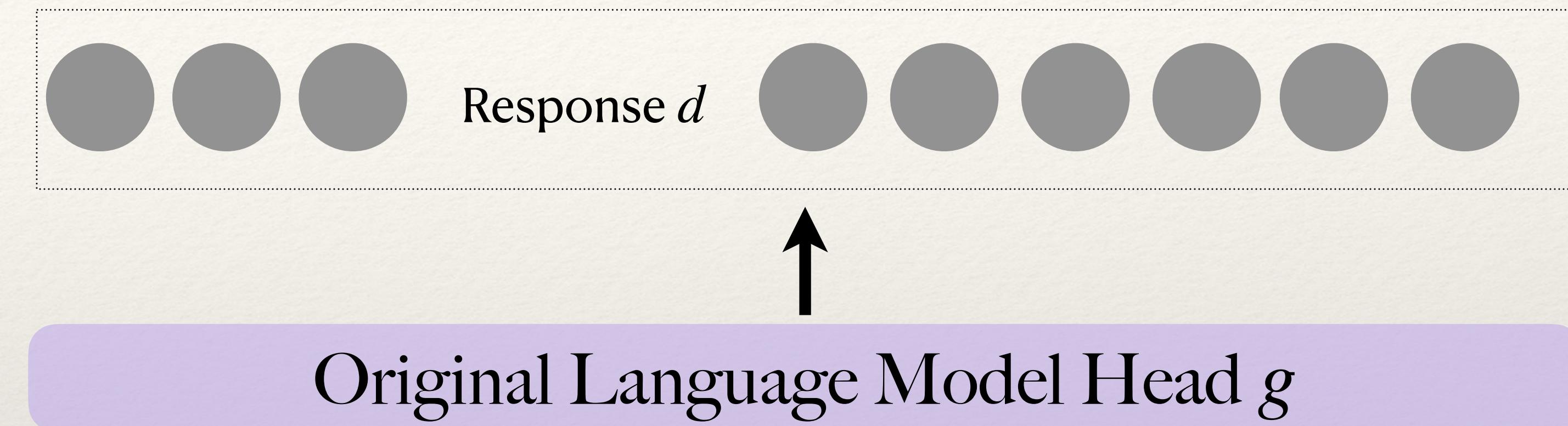
TLDR Automatically Trains Token-Level *Likelihood Optimization*



Multimodal Large Language Model $f + \alpha \Delta f$



TLDRL Automatically Trains Token-Level *Likelihood Optimization*



TLDR Automatically Trains Token-Level *Likelihood Optimization*

Model	Token-Level (%)	Sentence-Level (%)	Response-Level (%)
PaliGemma-3B	4.444	5.96	17.50
+ TLDR ($\tau=0.10$)	0.991	3.80	10.53
+ TLDR ($\tau=0.25$)	0.172	1.13	3.96
Llama-3.2-11B-Vision	0.073	0.85	1.88
+ TLDR ($\tau=0.10$)	0.078	0.69	2.71
+ TLDR ($\tau=0.25$)	0.066	0.74	1.72

- ❖ Training with TLDR Automatically Reduces Model Hallucinations Rates.
- ❖ Evaluations in this table are cross-evaluation with two TLDR models (i.e., using PaliGemma to evaluation Llama)

TLDR Automatically Trains Token-Level *Likelihood Optimization*

Model	BLINK			IsoBench	
	Count	Spatial Relation	Object Localize	Function Parity	Chess Winner
PaliGemma-3B	69.2	78.3	45.9	41.4	45.1
+ TLDR ($\tau=0.25$)	71.7	80.4	47.5	45.1	45.1
+ TLDR ($\tau=0.50$)	71.7	81.1	42.6	44.3	47.5
+ TLDR ($\tau=1.00$)	12.5	2.1	42.6	34.4	44.8
Llama-3.2-11B-Vision	55.0	61.5	60.7	34.9	45.5
+ TLDR ($\tau=0.25$)	67.5	65.0	67.2	35.4	43.6
+ TLDR ($\tau=0.50$)	65.8	65.7	59.0	33.3	47.9
+ TLDR ($\tau=1.00$)	61.7	65.7	56.6	35.1	39.4

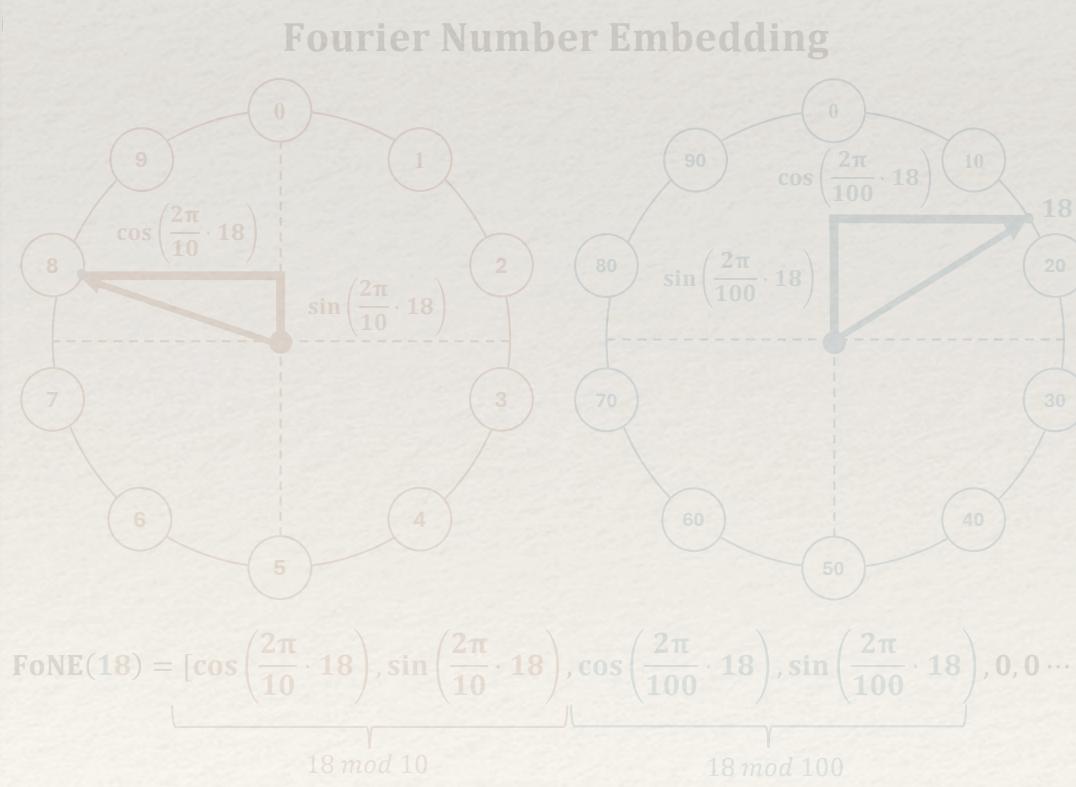
- ❖ Training with TLDR Automatically Improves MLLM’s Visual Understanding.
- ❖ Llama-3.2-11B-Vision improves at most +12.5 on counting.
- ❖ BLINK is “in-distribution” in terms of taxonomy, IsoBench is “out-of-distribution”.

Conclusions

- ❖ TLDR introduces token-level reward modeling that offers fine-grained, interpretable annotations to each generated token.
- ❖ TLDR offers diverse use cases: hallucination evaluation, guided self-correction, and human annotation speedup.
- ❖ TLDR’s LoRA training coupled with its architecture design implicitly improve the backbone model.
- ❖ Tuning the text backbone LLM could teach the MLLM.
- ❖ Future work: Extend the token-level reward framework to broader safety and alignment tasks, and integrate it into token-level DPO/PPO policy optimization for more robust multimodal generation.

Overview

LLMs are Bad At Arithmetics.
Can we find a family of embeddings to bypass tokenization?



Evaluating Multimodal LLMs Beyond Ranking Models.
Humans have pictorial superiority. How about MLLMs?



Text LLMs are Teachers for MLLMs.

Fine-Grained Rewards.
Train Text Backbone to be Token-Level Reward Models.



Text Steering Vectors.
Leverage Text-Based Interpretability Tools for MLLMs.



Textual Steering Vectors Can Improve Visual Understanding in Multimodal Large Language Models

[In Submission]



Haosheng
Gan*



Deqing
Fu*



Julian
Asilis*



Ollie
Liu*



Dani
Yogatama



Vatsal
Sharan



Robin
Jia



Willie
Neiswanger

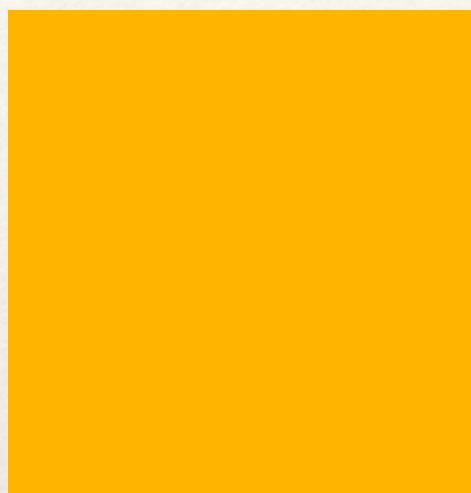
Let's Ponder: Why TLDR Works?

- ❖ Conditional Activation Steering [LPRM+25]: $h' \leftarrow h + \alpha \cdot f(\text{sim}(h, \text{proj}_c h)) \cdot v$
- ❖ h is the activation, c is the condition vector, and v is the **steering vector**.
- ❖ Let's assume f is an identity mapping, and c and h are unit length. We have

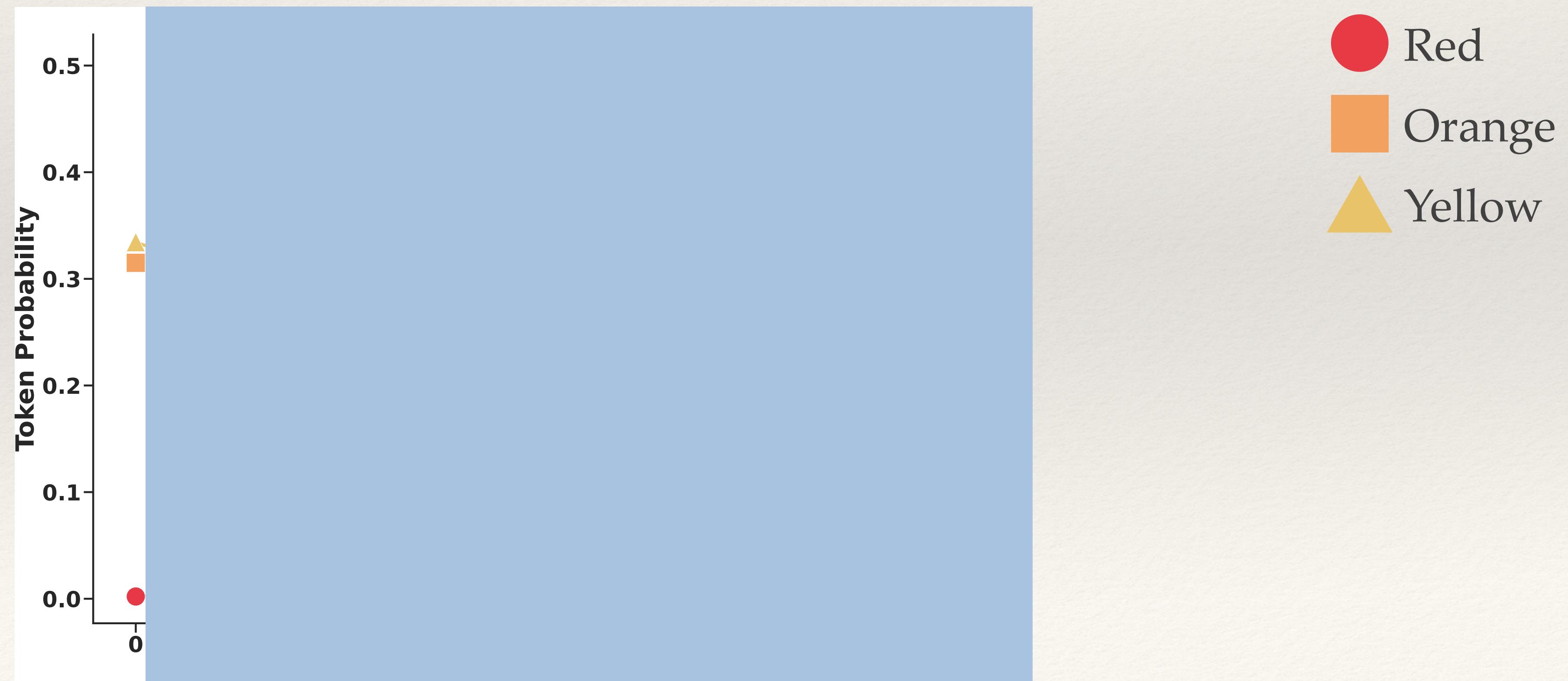
$$h' \leftarrow h + \alpha \cdot h^\top (cc^\top h) \cdot v = (I + \alpha \underbrace{h^\top c}_{\tau} vc^\top) h = (I + \tau \alpha \cdot \underbrace{vc^\top}_{\text{rank-1 update}}) h$$

- ❖ In TLDR's setting, LoRA with rank- r implies **r steering vectors** at each layer.

Toy Experiment: Color Intervention



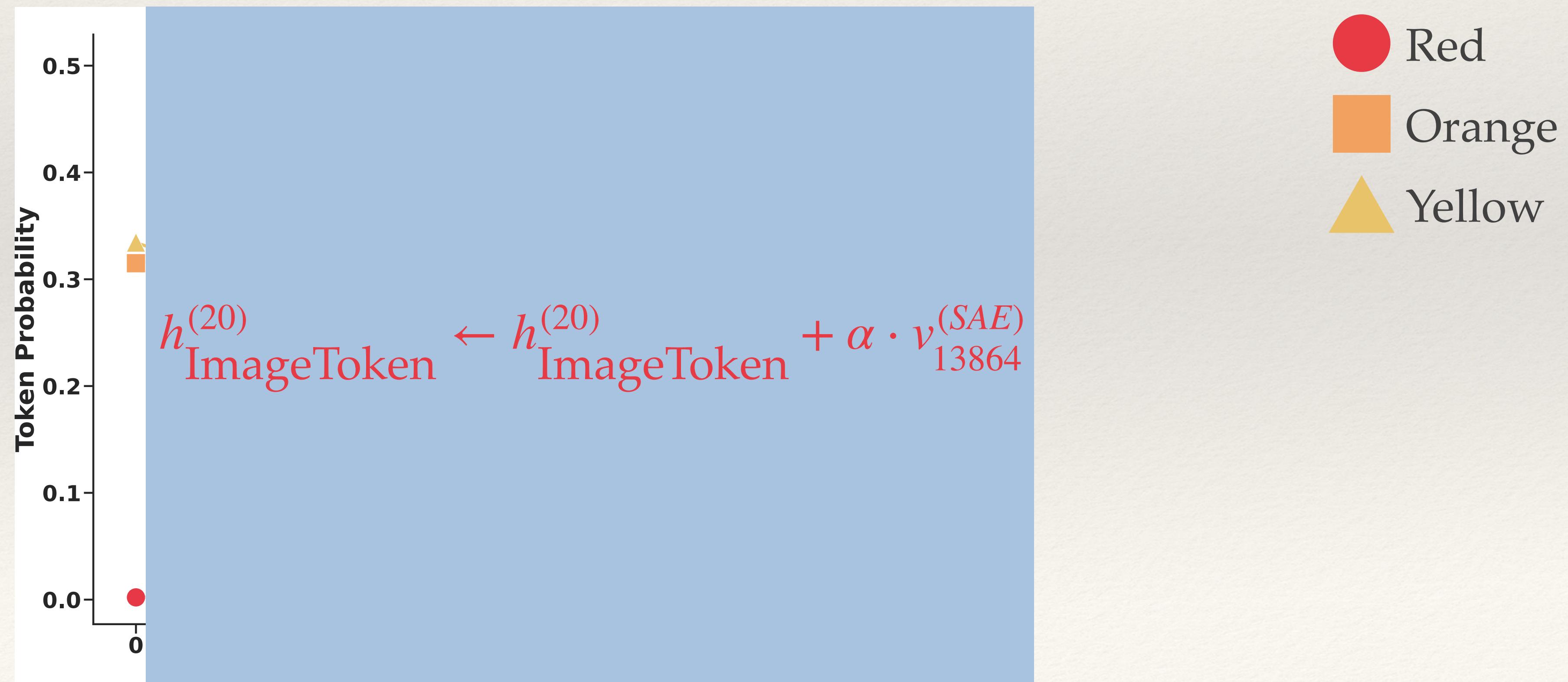
What's the color of the image?



Toy Experiment: Color Intervention

What's the color of the image?

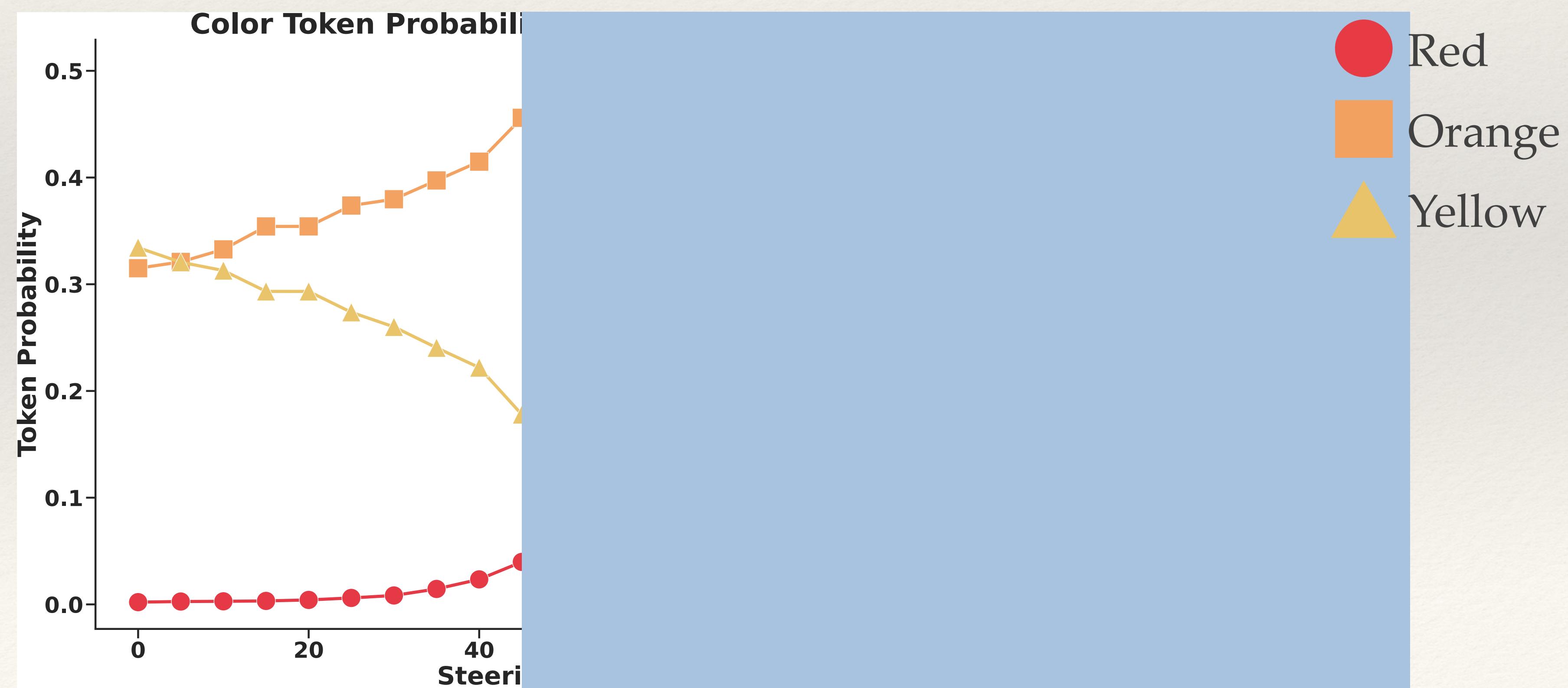
Steering Method: SparseAutoencoder (Gemma-2-9B)
Layer Index: 20; Feature ID: 13864
Feature Explanation: "color-related terms, specifically highlighting the color red"



Toy Experiment: Color Intervention

What's the color of the image?

*Steering Method: SparseAutoencoder (Gemma-2-9B)
Layer Index: 20; Feature ID: 13864
Feature Explanation: "color-related terms, specifically highlighting the color red"*

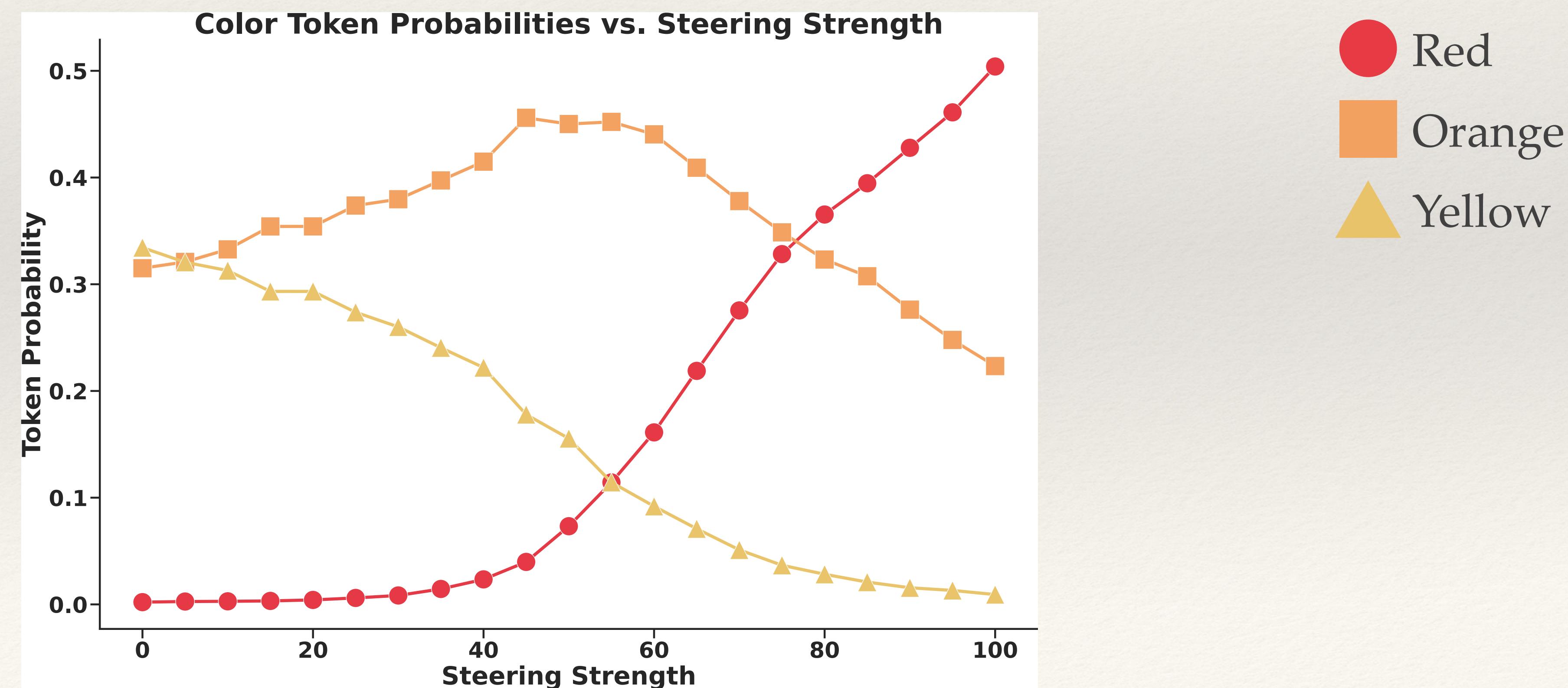


Toy Experiment: Color Intervention

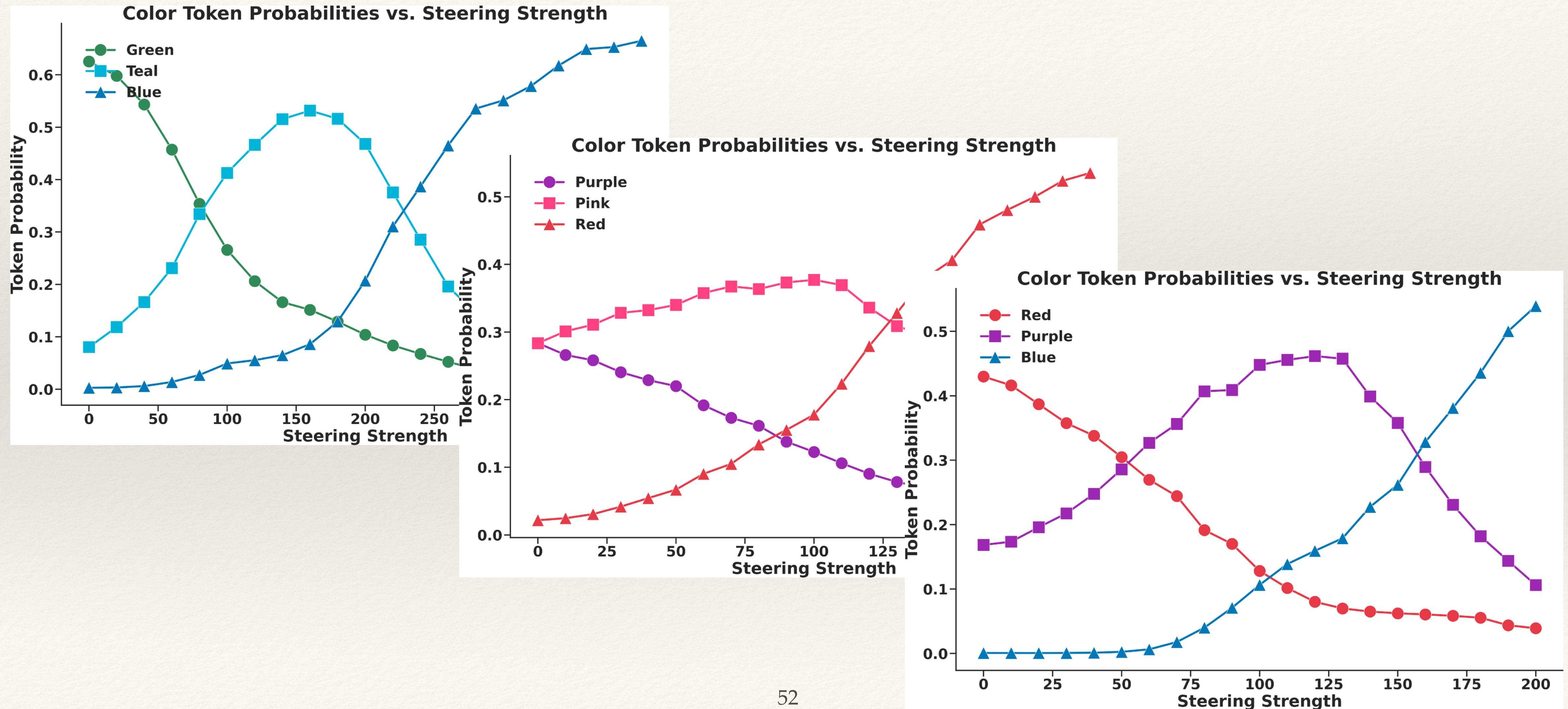


What's the color of the image?

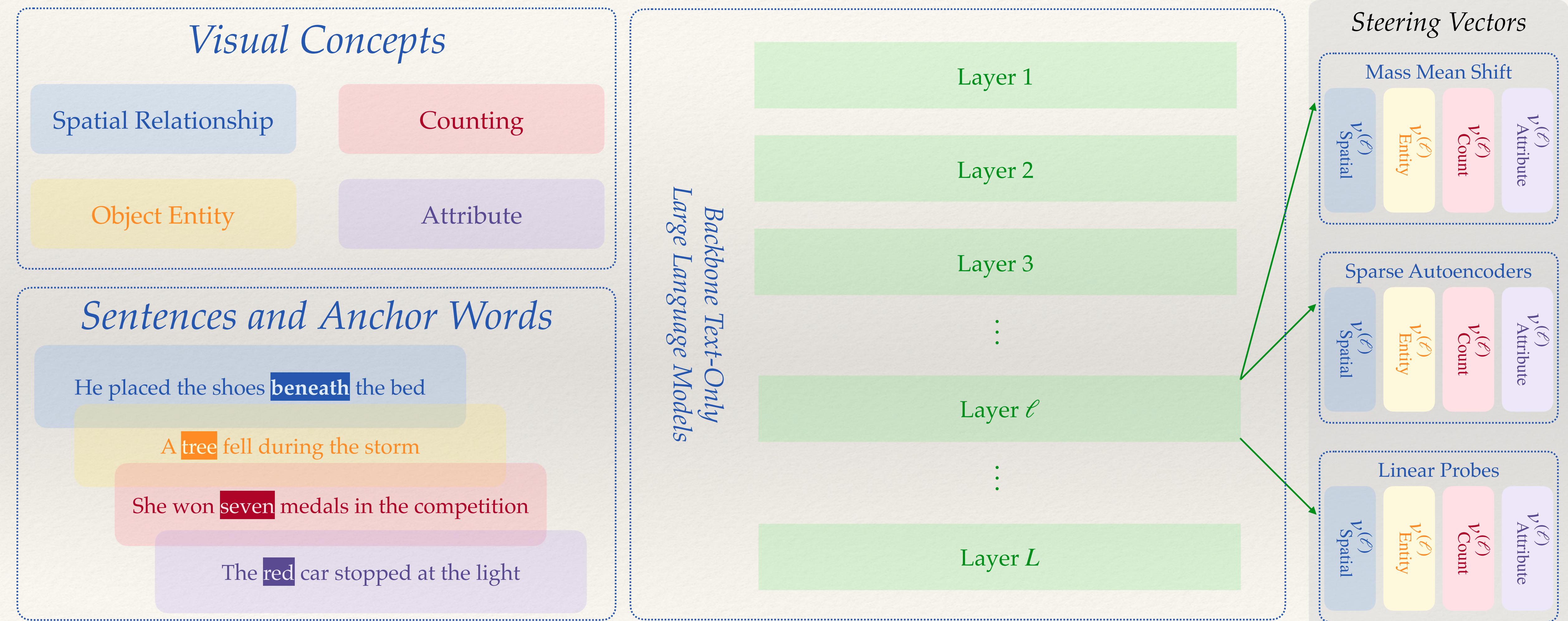
*Steering Method: SparseAutoencoder (Gemma-2-9B)
Layer Index: 20; Feature ID: 13864
Feature Explanation: "color-related terms, specifically highlighting the color red"*



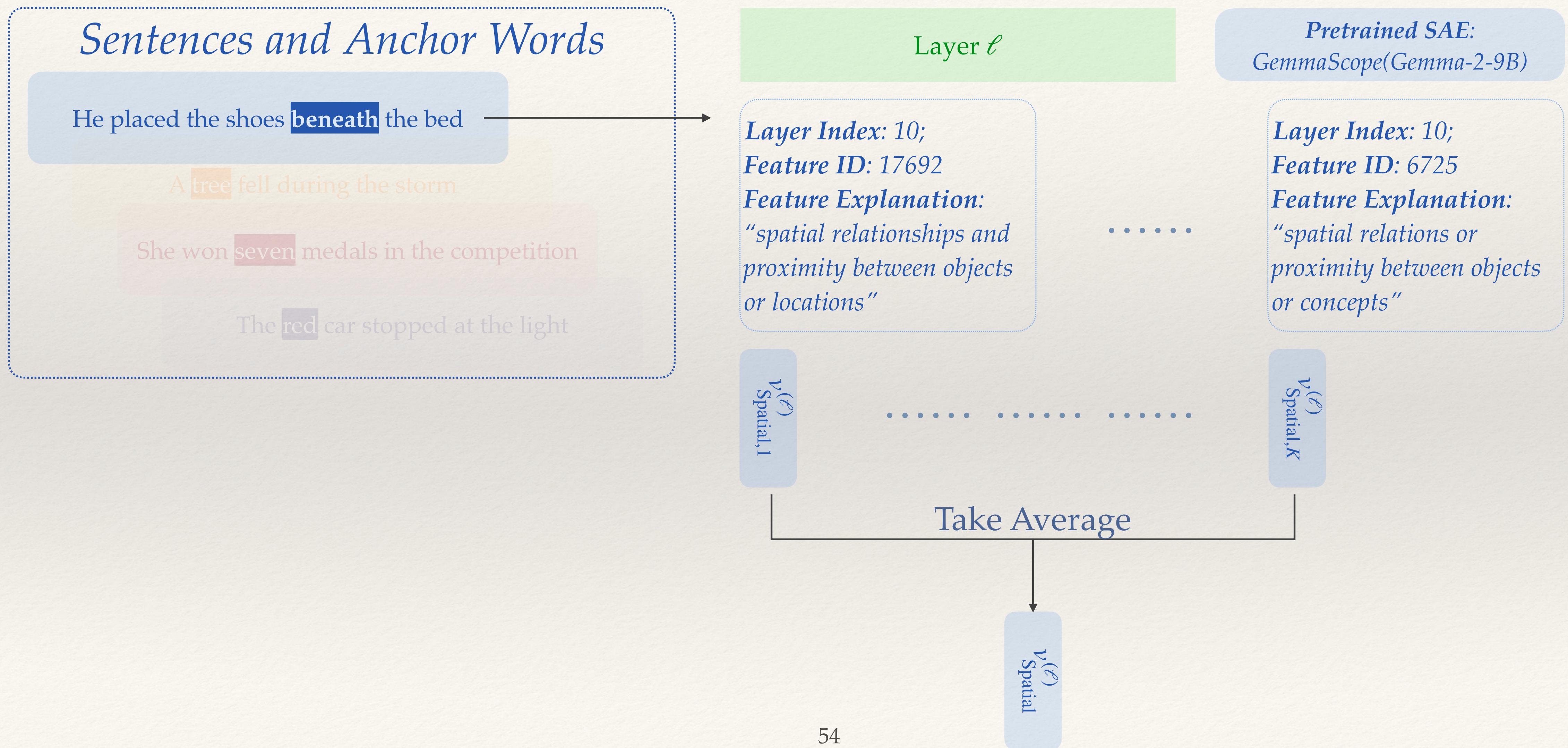
Toy Experiment: Color Intervention



Identify Textual Representations of Visual Concepts from Text-Only LLMs



Textual Vectors with SAE



Textual Vectors with MeanShift

Sentences and Anchor Words

He placed the shoes **beneath** the bed

A tree fell during the storm

She won **seven** medals in the competition

The **red** car stopped at the light

Anchor Words (Relevant Tokens)

$$h^{(\ell)}(w_1) \quad h^{(\ell)}(w_2) \quad \dots \quad h^{(\ell)}(w_{n_1})$$

Non-Anchor Words (Irrelevant Tokens)

$$h^{(\ell)}(w_1) \quad h^{(\ell)}(w_2) \quad \dots \quad h^{(\ell)}(w_{n_1})$$

Textual Vectors with MeanShift

Sentences and Anchor Words

He placed the shoes **beneath** the bed

A tree fell during the storm

She won **seven** medals in the competition

The **red** car stopped at the light



Textual Vectors with MeanShift

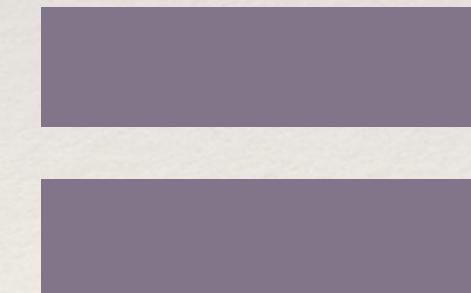
Sentences and Anchor Words

He placed the shoes **beneath** the bed

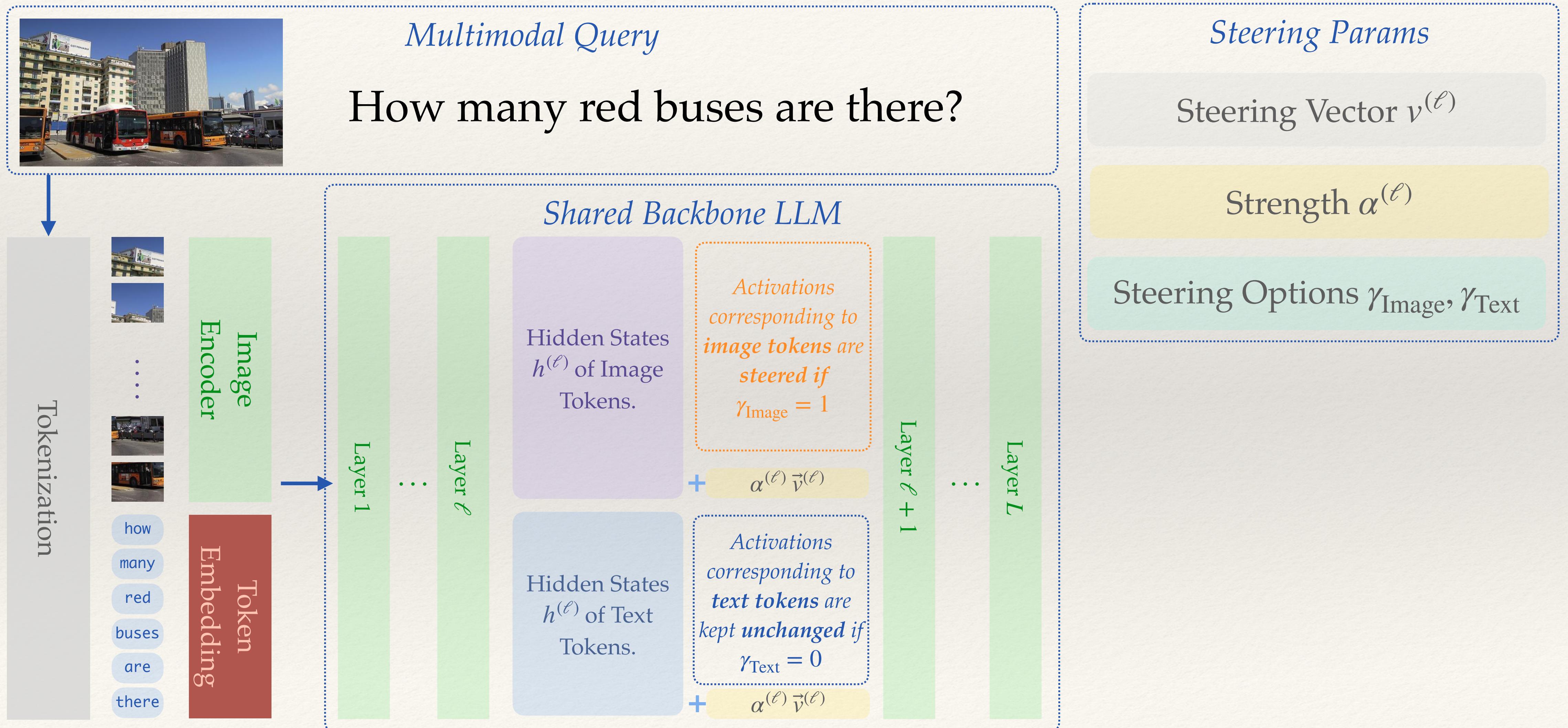
A tree fell during the storm

She won **seven** medals in the competition

The **red** car stopped at the light



Steering MLLMs using Text Representations



Steering Can Improve Visual Understanding

Spatial Relationships



Considering the relative positions of the fork and the cup in the image provided, where is the fork located with respect to the cup? Select from the following choices.

- (A) left
- (B) right

Counting



Answer the multiple choice question by only responding the letter of the correct answer. How many beds are in the image? Select from the following choices.

- (A) 0
- (B) 2
- (C) 1
- (D) 3
- (E) 4

Steering Can Improve Visual Understanding

Visual Concepts

CV-Bench Tasks

	Counting	Spatial Relationship	Entity	Attribute
Count	+0.7% (L7@0.8)	+1.3% (L14@1)	+0.0% (L9@0.8)	+0.0% (L16@0.6)
Relation	+3.3% (L9@1)	+7.3% (L5@1)	+0.0% (L9@0.6)	+2.7% (L6@1)
Distance	+1.3% (L6@0.8)	+0.7% (L16@0.6)	+0.7% (L6@0.1)	+0.7% (L5@0.2)
Depth	-0.7% (L11@0.6)	-1.3% (L11@0.4)	+0.7% (L10@0.8)	-0.7% (L10@0.4)

Interven Text Tokens

	Counting	Spatial Relationship	Entity	Attribute
Count	+2.7% (L5@0.4)	+1.3% (L5@1)	+2.0% (L11@0.8)	+1.3% (L10@1)
Relation	+0.7% (L10@0.8)	+2.7% (L10@1)	+3.3% (L14@1)	+4.0% (L13@1)
Distance	+0.0% (L19@0.8)	+0.0% (L9@0.6)	-0.7% (L6@0.2)	-2.0% (L10@0.4)
Depth	+0.7% (L11@1)	+2.7% (L10@1)	+0.7% (L11@0.8)	+0.7% (L10@1)

Intervene Image Tokens

	Counting	Spatial Relationship	Entity	Attribute
Count	+2.7% (L9@0.6)	+1.3% (L5@0.6)	-0.7% (L7@0.2)	+2.0% (L10@1)
Relation	+3.3% (L9@1)	+5.3% (L5@1)	+0.7% (L6@0.8)	+2.0% (L6@0.8)
Distance	-0.7% (L15@0.8)	+2.0% (L8@0.2)	+1.3% (L6@0.1)	+1.3% (L13@0.2)
Depth	+0.0% (L5@0.8)	+2.0% (L11@0.6)	+1.3% (L10@0.8)	+0.0% (L10@0.8)

Intervene Image + Text Tokens

Performance Improvements on CV-Bench tasks when steering PaliGemma2-3B with MeanShift vectors

Steering Can Improve Visual Understanding

- ❖ CV-Bench Spatial Relationship
- ❖ Idefics3-8B-Llama3
- ❖ MeanShift gives the best steering improvements

Tokens to Intervene	Intervention Method				
	TEXT	IMAGE	SAE	Probe	MeanShift
<i>Baseline</i>				73.3	
<i>Prompting</i>				75.3	
X			76.0	78.0	80.0
	X		78.0	72.7	76.7
X	X		77.3	78.7	80.7 

Steering Generalizes Out-of-Distribution

Spatial Relationships



What'sUp-A

Please select the correct caption for the image:

- (A) A toilet roll under a chair
- (B) A toilet roll to the left of a chair
- (C) A toilet roll to the right of a chair
- (D) A toilet roll on a chair

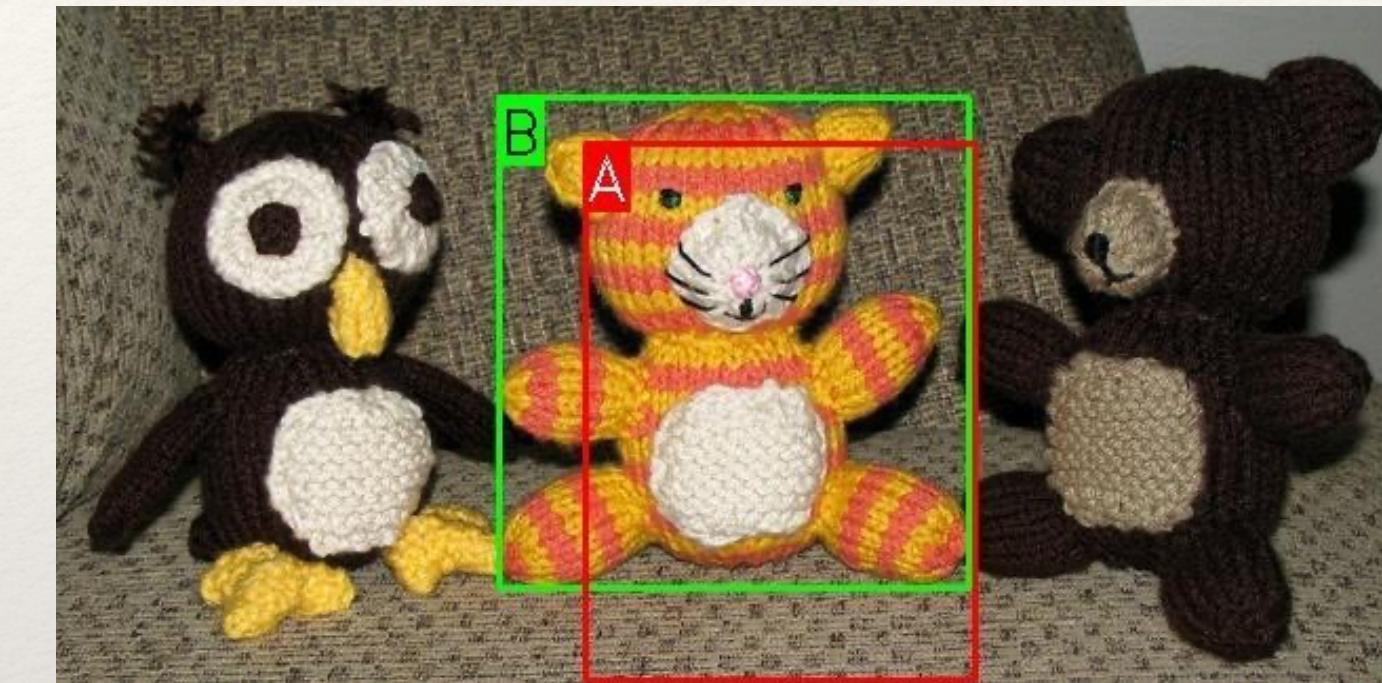


What'sUp-B

Please select the correct caption for the image:

- (A) A bowl behind a cup
- (B) A bowl to the left of a cup
- (C) A bowl to the right of a cup
- (D) A bowl in front of a cup

Object Localization



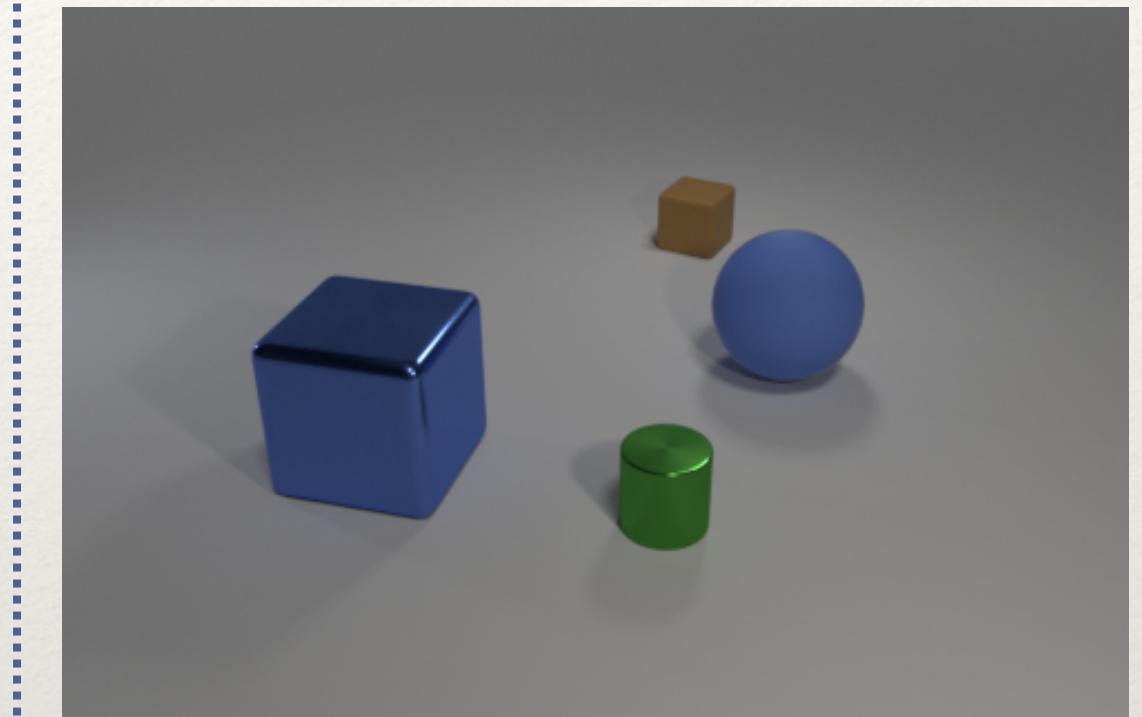
BLINK Object Localization

A bounding box is an annotated rectangle surrounding an object. The edges of bounding boxes should touch the outermost pixels of the object that is being labeled. Given the two bounding boxes on the image, labeled by A and B, which bounding box more accurately localizes and encloses the teddy bear?

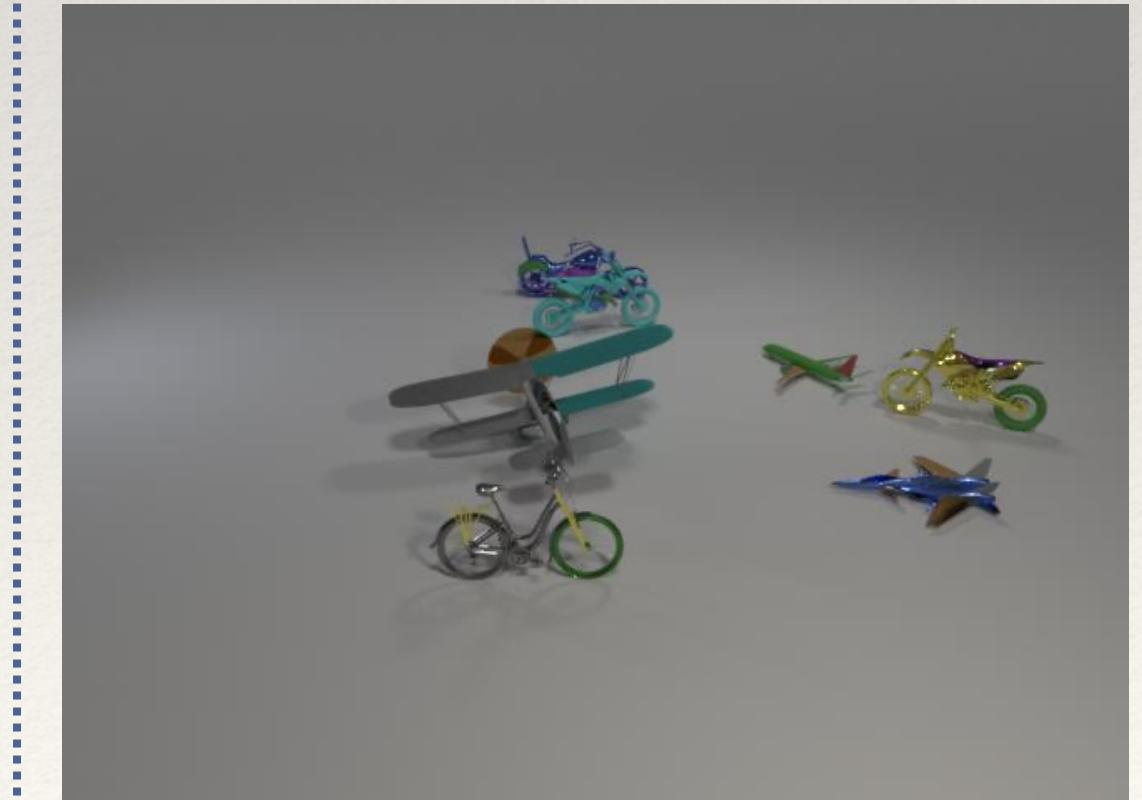
Select from the following options.

- (A) Box A
- (B) Box B

Counting

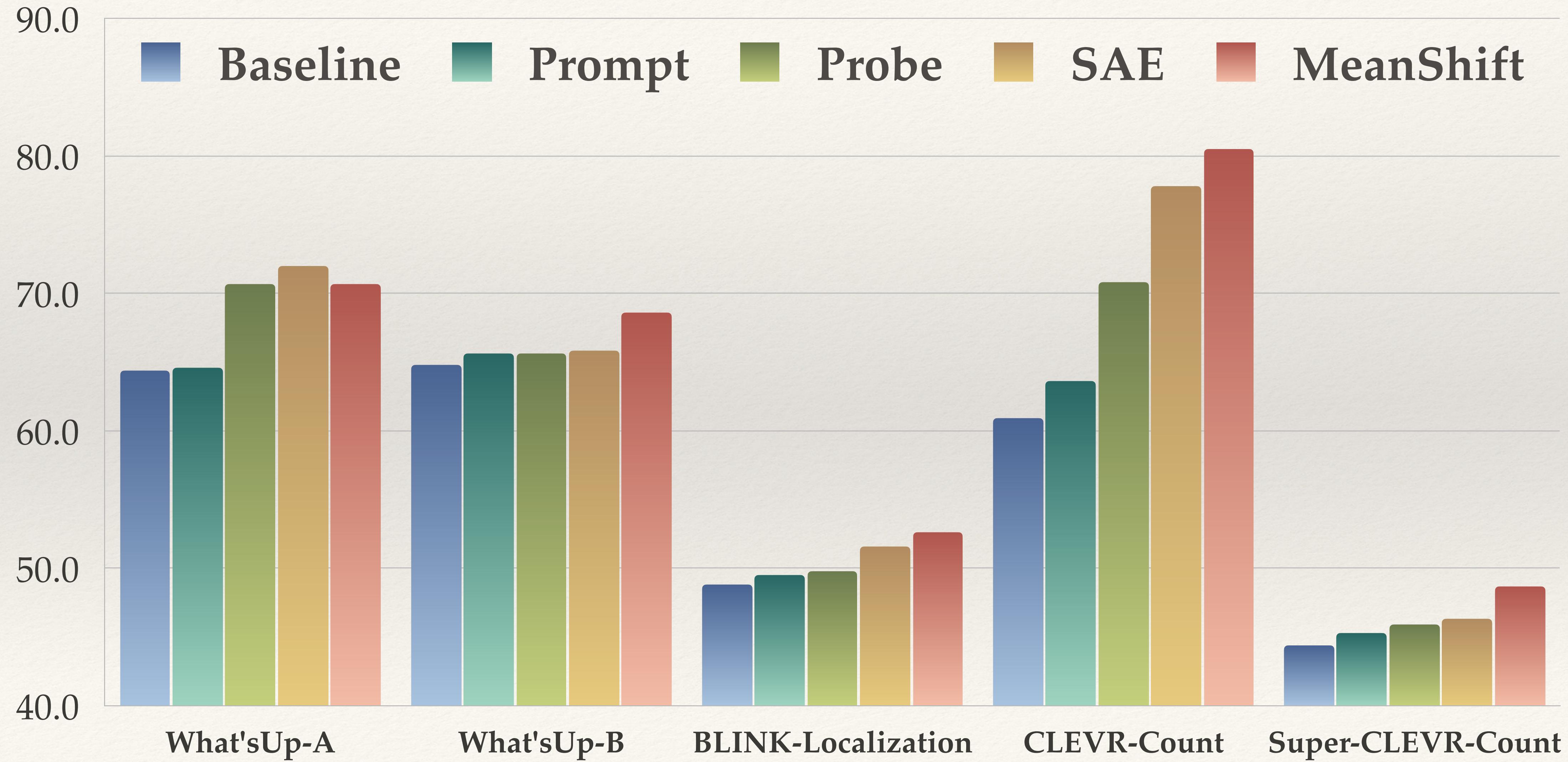


CLEVR

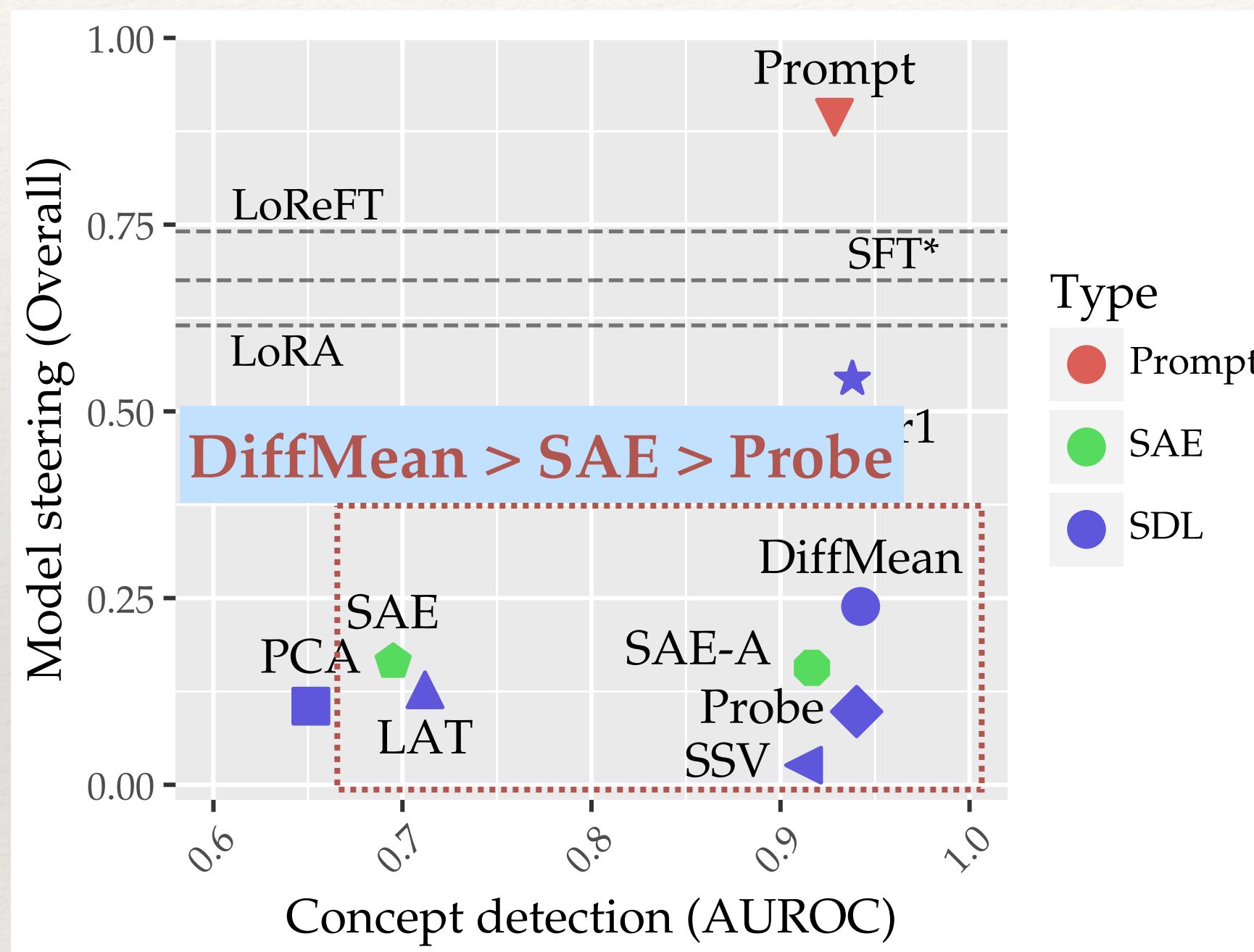


Super-CLEVR

Steering Generalizes Out-of-Distribution

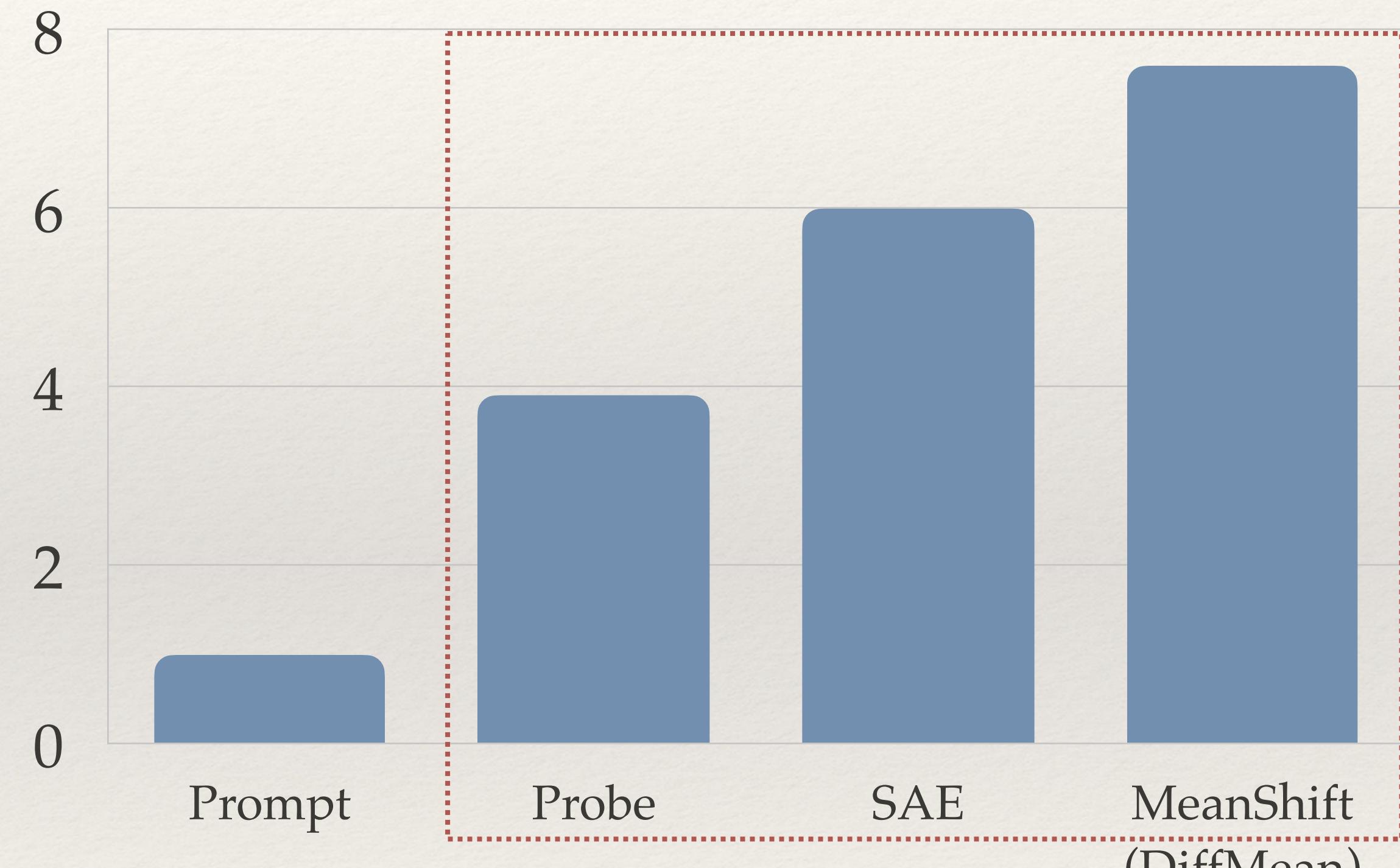


Connection to AxBench

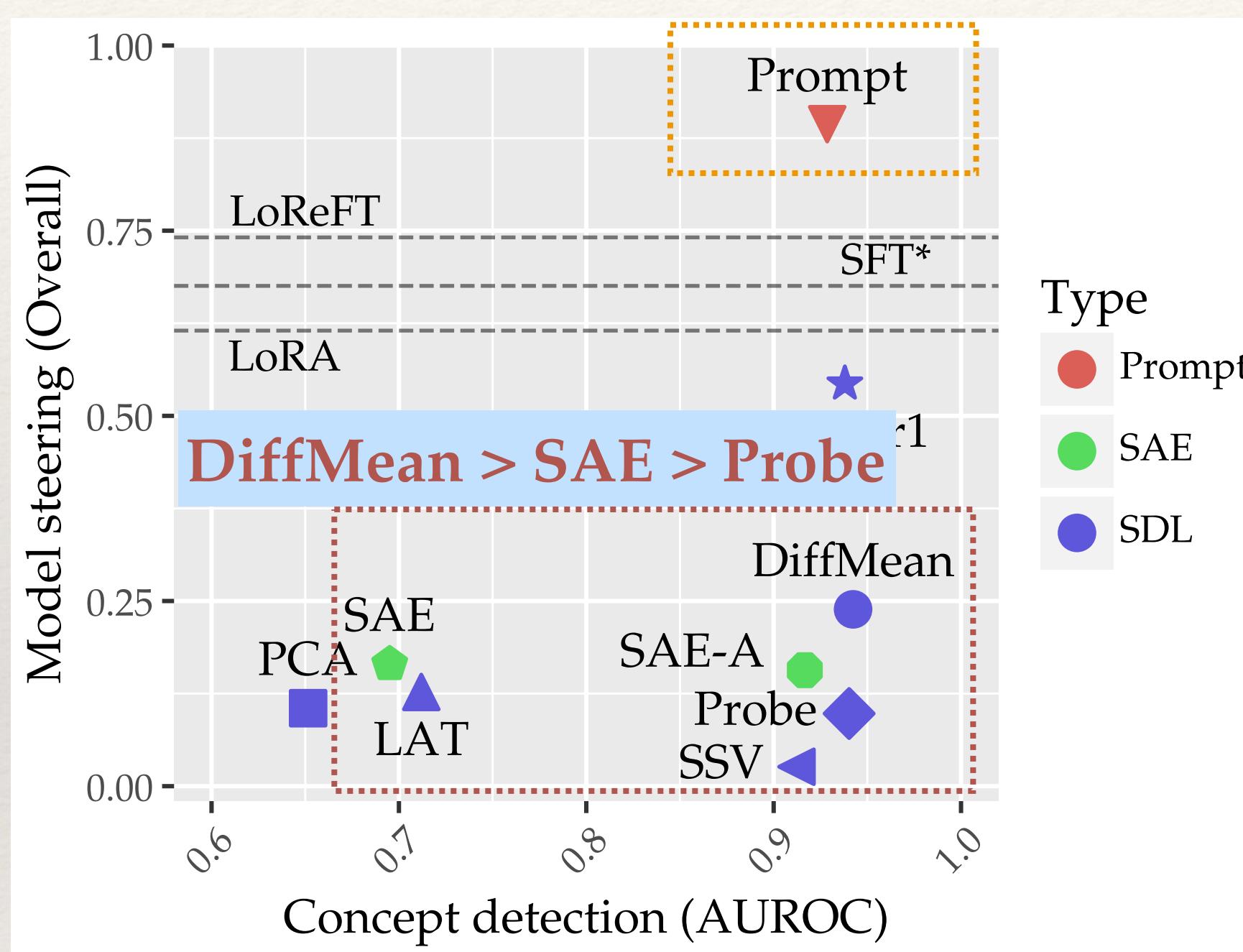


[Wu et al. 2025] *AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders.*

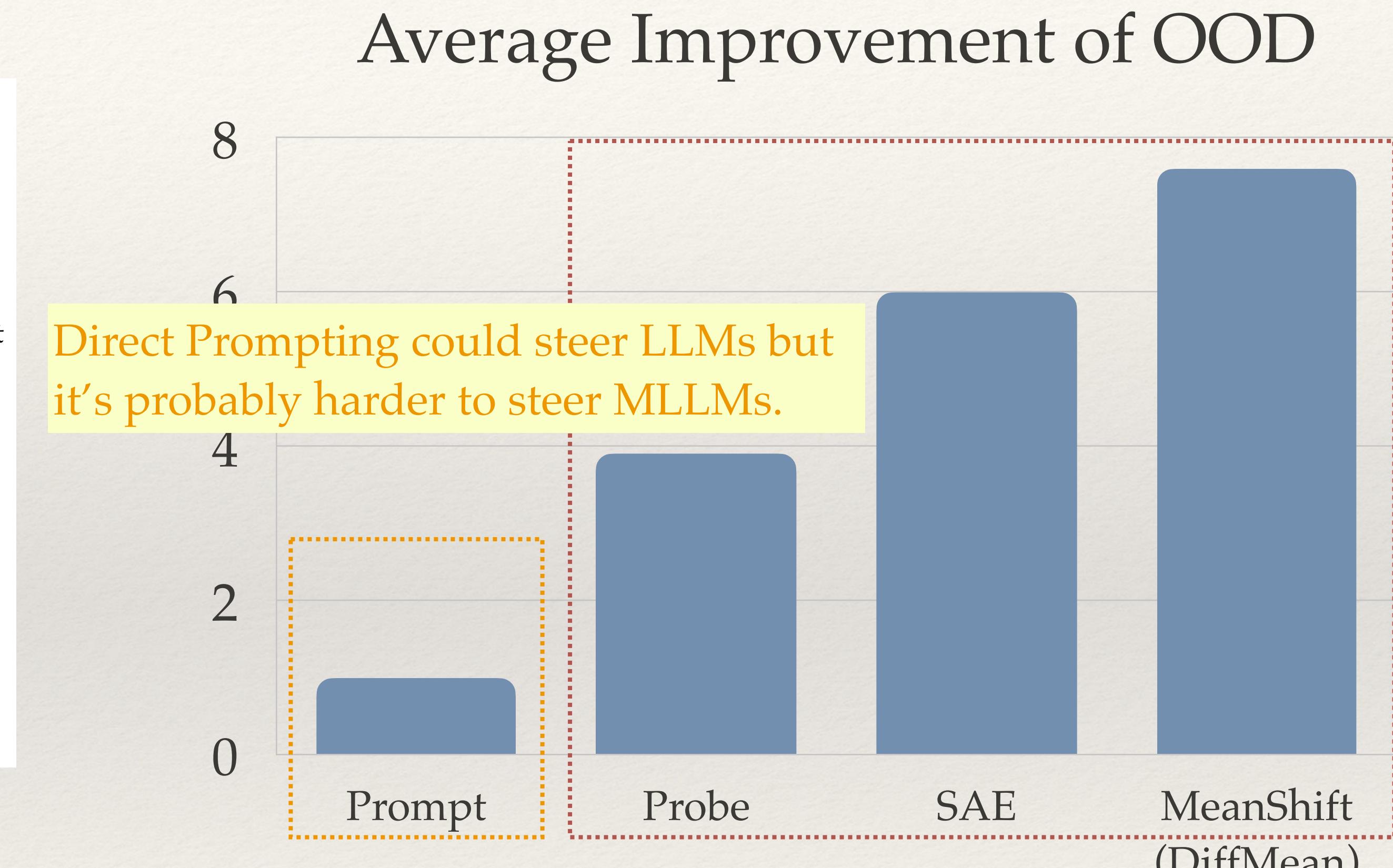
Average Improvement of OOD



Connection to AxBench



[Wu et al. 2025] *AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders.*

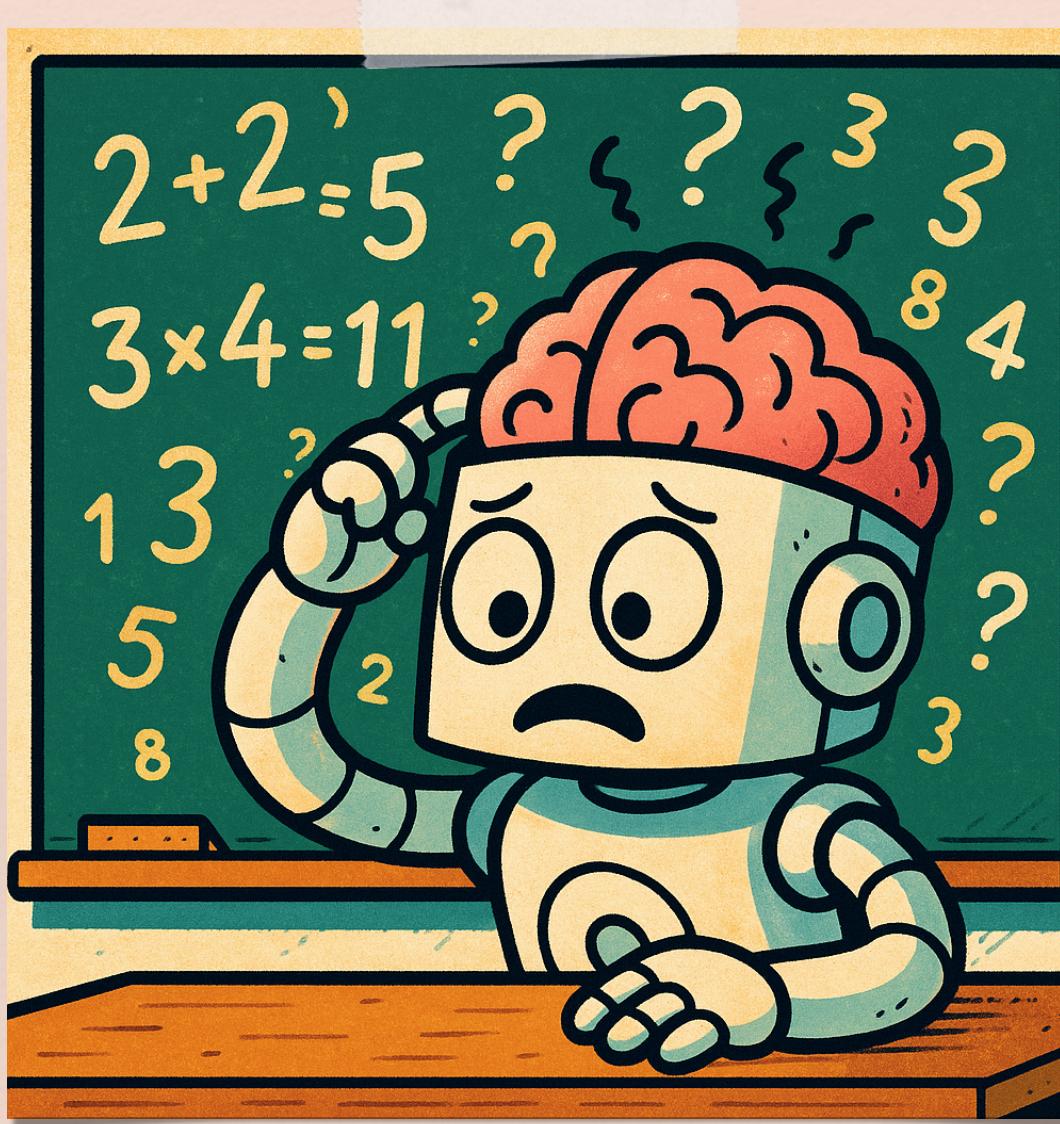


Conclusions

- ❖ We introduce a plug-and-play steering paradigm that repurposes text-only LLM representations—via SAEs, MeanShift, and linear probes—to guide multimodal LLMs with minimal extra data or compute. (All experiments were done with only *one* A6000.)
- ❖ **MeanShift** interventions on models like PaliGemma2-3B yield up to +7.3% improvement in spatial relation accuracy and +3.3% in counting tasks on CV-Bench, consistently outperforming direct prompting.
- ❖ These textual steering vectors generalize out-of-distribution.
- ❖ Future Work: Merging steering vectors of difference visual concepts across multiple layers.

Failures and Recipes In This Talk

❖ LLMs Are Bad at Arithmetics



Recipe: Better Representations

❖ Multimodal LLMs Hallucinate



Recipe: Leveraging the Power of Their Textual Backbone LLMs

Thank you!

- ❖ Questions? Comments? Ideas?
- ❖ Contact: deqingfu@usc.edu