

Limitations of Large Language Models and future directions

Antonio Barbalau — Professional LLM Hater

LLMs are pretty good

Yamada et al (April, 2025) – The AI Scientist V2

Under review as a workshop paper at ICLR 2025

COMPOSITIONAL REGULARIZATION: UNEXPECTED OBSTACLES IN ENHANCING NEURAL NETWORK GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks excel in many tasks but often struggle with compositional generalization—the ability to understand and generate novel combinations of familiar components. This limitation hampers their performance on tasks requiring systematic reasoning beyond the training data. In this work, we introduce a training method based on compositional regularization, which adds a regularization term to the loss function, aiming to encourage the network to develop compositional representations. Contrary to our expectations, our experiments on synthetic arithmetic expression datasets reveal that models trained with compositional regularization do not achieve significant improvements in generalization to unseen combinations compared to baseline models. Additionally, we find that increasing the complexity of expressions exacerbates the models' difficulties, regardless of compositional regularization. These findings highlight the challenges of enforcing compositional structures in neural networks and suggest that such regularization may not be sufficient to enhance compositional generalization.

1 INTRODUCTION

Compositional generalization refers to the ability to understand and produce novel combinations of known components, a fundamental aspect of human cognition (Ito et al., 2022). Despite the success of neural networks in various domains, they often struggle with compositional generalization, limiting their applicability in tasks requiring systematic reasoning beyond the training data (Qu et al., 2023; Klinger et al., 2020). Previous efforts to enhance compositional generalization have explored various approaches, including attention mechanisms and training strategies (Finn et al., 2017; Lepori et al., 2023). A promising direction is the incorporation of regularization terms that encourage certain properties in the learned representations (Yin et al., 2023).

In this work, we introduce a training method that incorporates an explicit *compositional regularization* term into the loss function. This regularization term is designed to penalize deviations from expected compositional structures in the network's internal representations, with the aim of encouraging the network to form compositional representations. We hypothesized that this approach would enhance the network's ability to generalize to unseen combinations. However, our experiments on synthetic arithmetic expression datasets show that the inclusion of compositional regularization does not lead to the expected improvements in generalization performance. In some cases, it even hinders the learning process. Furthermore, we observe that increasing the complexity of arithmetic expressions by adding more operators or nesting, exacerbates the model's generalization difficulties regardless of the regularization. These unexpected results highlight the challenges of enforcing compositionality through regularization and suggest that this approach may not be straightforwardly effective.

In summary, we propose a compositional regularization term intended to enhance compositional generalization in neural networks, conduct extensive experiments to evaluate its impact, and analyze the unexpected outcomes, including the impact of operator complexity, discussing potential reasons why compositional regularization did not yield the anticipated benefits.

Under review as a workshop paper at ICLR 2025

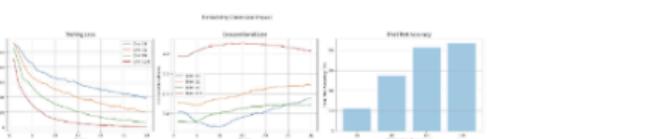
REFERENCES

- Chelson Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. pp. 1126–1135, 2017.
Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.
Takuya Ito, Tim Klinger, D. Schultz, J. Murray, Michael W. Cole, and Mattia Rigotti. Compositional generalization through abstract representations in human and artificial neural networks. 2022.
Tim Klinger, D. Adjodah, Vincent Marois, Joshua Joseph, M. Riemer, and Murray Campbell. A study of compositional generalization in neural models. *ArXiv*, abs/2006.09437, 2020.
Michael A. Lepori, Thomas Serre, and Elie Pavlich. Break it down: Evidence for structural compositionality in neural networks. *ArXiv*, abs/2301.10884, 2023.
Carolyn Qu, Rodrigo Nieto, S. Menon, and John Hewitt. Compositional generalization based on semantic interpretation: Where can neural networks improve? 2023.
Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. pp. 5998–6008, 2017.
Yongning Yin, Jiali Zeng, Yufu Li, Fandong Meng, Jie Zhou, and Yue Zhang. Consistency regularization training for compositional generalization. pp. 1294–1308, 2023.

SUPPLEMENTARY MATERIAL

A EFFECT OF EMBEDDING DIMENSION

We explored the impact of different embedding dimensions on model performance. Figure 4 shows the training loss, compositional loss, and final test accuracy for embedding dimensions 16, 32, 64, and 128. Increasing the embedding dimension did not consistently improve test accuracy. While larger embedding dimensions provide the model with greater capacity, our results indicate that simply increasing model capacity is not sufficient to enhance compositional generalization in this context. This suggests that the bottleneck may lie in the model's ability to capture compositional structures rather than in its representational capacity.



B INTEGRATION OF ATTENTION MECHANISM

We compared the baseline model with an enhanced model that incorporates an attention mechanism (Vaswani et al., 2017). The attention mechanism is known to improve performance in various sequence-to-sequence tasks by allowing the model to focus on relevant parts of the input sequence.

Under review as a workshop paper at ICLR 2025

2 RELATED WORK

Compositional generalization in neural networks has been a topic of considerable research interest (Klinger et al., 2020). Ito et al. (2022) explored abstract representations to tackle this issue, emphasizing the importance of learning to make use of structure. Lepori et al. (2023) proposed consistency regularization training to enhance compositional generalization. Many learning approaches, such as Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), have also been investigated to improve generalization capabilities. Lepori et al. (2023) studied structural compositionality in neural networks, suggesting that networks may implicitly learn to decompose complex tasks.

Our work differs by directly incorporating an explicit regularization term into the training objective to enforce compositional structures. Despite the theoretical appeal, our findings indicate that such regularization may not effectively enhance compositional generalization and that operator complexity plays a significant role in the models' performance limitations.

3 METHOD

Our goal is to enhance compositional generalization in neural networks by incorporating a compositional regularization term into the training loss. We focus on a simple yet illustrative task: evaluating arithmetic expressions involving basic operators.

3.1 MODEL ARCHITECTURE

We use an LSTM-based neural network (Goodfellow et al., 2016) to model the mapping from input expressions to their computed results. The model consists of an embedding layer, an LSTM layer, and a fully connected output layer.

3.2 COMPOSITIONAL REGULARIZATION

Let h_t be the hidden state at time t . We define the compositional regularization term as the mean squared difference between successive hidden states:

$$L_{\text{comp}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|h_{t+1} - h_t\|^2 \quad (1)$$

where T is the length of the input sequence.

This term penalizes large changes in hidden states between successive time steps, encouraging the model to form additive representations, which are a simple form of compositionality.

3.3 TRAINING OBJECTIVE

The total loss is the sum of the main loss (mean squared error between predicted and true results) and the compositional regularization term weighted by a hyperparameter λ :

$$L_{\text{total}} = L_{\text{train}} + \lambda L_{\text{comp}} \quad (2)$$

We experimented with different values of λ to assess its impact on compositional generalization.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We generated synthetic datasets of arithmetic expressions to evaluate compositional generalization.

The datasets consist of expressions combining digits and operators (e.g., “3+4”, “7*2”). We compare models trained with and without the compositional regularization term and performed several ablation studies to assess the impact of different hyperparameters, operator complexity, and architectural choices.

Under review as a workshop paper at ICLR 2025

B.1 EXPERIMENTAL SETUP

We modified the baseline LSTM model to include an attention layer after the LSTM outputs. The attention weights were calculated based on the hidden states, and a context vector was formed to aid in the final output prediction.

B.2 RESULTS

The attention model achieves a test accuracy similar to the baseline, as shown in Figure 5. While the attention mechanism slightly improved the training dynamics, it did not lead to significant improvements in generalization performance. This suggests that the challenges in compositional generalization are not primarily due to the model's ability to focus on relevant parts of the input sequence but may be related to deeper architectural limitations or the need for more sophisticated mechanisms to capture compositionality.



Figure 5: Comparison of baseline and attention models. **Left:** Training loss over epochs shows similar convergence for both models. **Middle:** Compositional loss remains comparable, indicating that attention does not significantly enhance compositional representations. **Right:** Final test accuracy is similar for both models, suggesting that the attention mechanism does not address the compositional generalization challenges.

C ADDITIONAL EXPERIMENTS

C.1 ABALATION STUDY ON COMPOSITIONAL WEIGHT

We conducted an ablation study on the compositional weight λ to further investigate its impact on model performance. Figures 6 and 7 show the training loss and final test accuracy for various values of λ . Higher λ values effectively reduce the compositional loss but adversely affect test accuracy. This reinforces the conclusion that emphasizing compositional regularization may conflict with the primary learning objective.

C.2 COMPARISON OF LSTM AND RNN ARCHITECTURES

We compared the performance of LSTM and simple RNN architectures to assess the influence of model choice on compositional generalization. Figure 8 illustrates the training loss and final test accuracy for both models. The LSTM model showed marginal improvements over the simple RNN, but both architectures struggled with compositional generalization, indicating that the limitations are not solely due to the recurrent unit type.

C.3 DROPOUT IMPACT

We investigated the impact of dropout on model performance. Figure 9 shows the final test accuracy for different dropout rates. We found that increasing the dropout rate did not lead to significant improvements in generalization, suggesting that regularization techniques like dropout may not address compositional generalization challenges. This indicates that standard regularization methods may not be sufficient to overcome the inherent difficulties in learning compositional structures.

Under review as a workshop paper at ICLR 2025

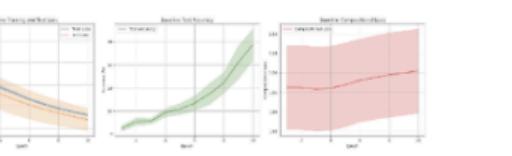


Figure 1: Baseline model performance over epochs. **Left:** Training and test loss decrease over epochs, indicating learning progress. **Middle:** Test accuracy increases, reaching approximately 84%. **Right:** Compositional loss remains steady, suggesting the model does not inherently develop compositional representations without regularization.

Under review as a workshop paper at ICLR 2025

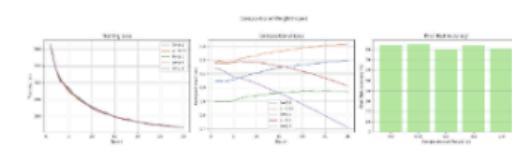


Figure 2: Impact of compositional weight λ on model performance. **Left:** Training loss over epochs for different λ . Higher λ values slightly increase training loss. **Middle:** Compositional loss decreases with higher λ , indicating the regularization term effectively enforces compositionality. **Right:** Final test accuracy does not improve with higher λ and may decrease, suggesting a trade-off between compositional regularization and the primary learning objective.

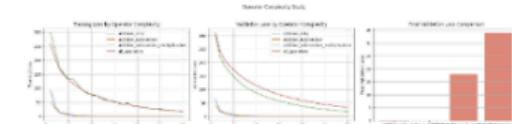


Figure 3: Model performance on expressions with varying operator complexity. **Left:** Training loss increases with operator complexity, indicating the models struggle to fit more complex data. **Middle:** Validation loss is higher for complex expressions, reflecting poor generalization. **Right:** Final validation accuracy decreases significantly as operator complexity increases, underscoring inherent limitations in handling complex compositional structures with compositional regularization alone.

5 CONCLUSION

In this work, we introduced a compositional regularization term with the intention of enhancing compositional generalization in neural networks. Our experiments on synthetic arithmetic expression datasets revealed that compositional regularization did not lead to the expected improvements in generalization performance. In some cases, it even hindered the learning process. Additionally, we found that increasing the complexity of arithmetic expressions exacerbates the models' generalization difficulties, highlighting inherent limitations.

These findings highlight the challenges of enforcing compositional structures in neural networks through regularization. Possible reasons for the lack of improvement include conflicts between the regularization term and the primary learning objective, which may cause the network to prioritize minimizing compositional loss over fitting the data. Additionally, the measure of compositionality used in the regularization term may not align with the aspects of compositionality that are critical for generalization. The synthetic dataset may also not adequately capture the complexities of compositional generalization in real-world tasks, and increased operator complexity introduces additional challenges that compositional regularization alone cannot overcome.

For experimental validation, we present a synthetic dataset consisting of arithmetic expressions and their results. Our results show that as the complexity of the expressions increases, the models' ability to generalize diminishes significantly. Neither the baseline model nor the model with compositional regularization could handle expressions with higher operator complexity effectively. This finding emphasizes that compositional regularization alone may not address the challenges posed by complex compositional structures.

4

Under review as a workshop paper at ICLR 2025

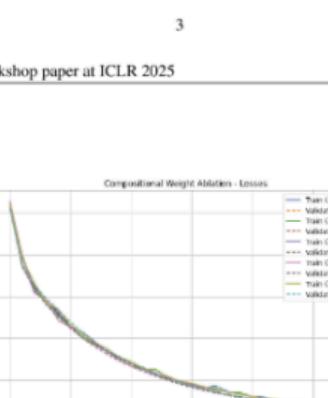


Figure 6: Training loss over epochs for different values of compositional weight λ . Increasing λ leads to slightly higher training loss, indicating potential interference with the primary learning objective.

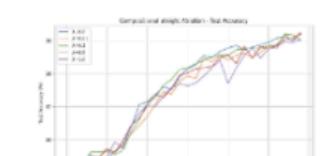


Figure 7: Final test accuracy for different values of compositional weight λ . Higher λ values do not improve test accuracy and may lead to decreased performance, suggesting a trade-off between compositional regularization and generalization.

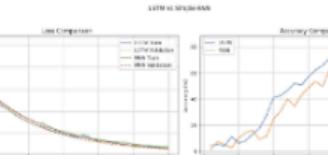


Figure 8: Comparison of LSTM and RNN architectures. **Left:** Training loss over epochs shows similar convergence patterns, with LSTM performing slightly better. **Right:** Final test accuracy is marginally higher for LSTM, but both models struggle with compositional generalization, suggesting that recurrent unit choice does not resolve the underlying challenges.

D HYPERPARAMETERS AND TRAINING DETAILS

D.1 ADDITIONAL NOTES

- All experiments were implemented using PyTorch.
- Training was conducted on a single NVIDIA GPU.
- Early stopping was not used; models were trained for a fixed number of epochs.
- The synthetic dataset was generated with a predefined random seed for reproducibility.

5

Under review as a workshop paper at ICLR 2025

6

Under review as a workshop paper at ICLR 2025

7

Under review as a workshop paper at ICLR 2025

8

Jones et al (March, 2025) — Large Language Models Pass the Turing Test

Large Language Models Pass the Turing Test

Cameron R. Jones

Department of Cognitive Science

UC San Diego

San Diego, CA 92119

cameron@ucsd.edu

Benjamin K. Bergen

Department of Cognitive Science

UC San Diego

San Diego, CA 92119

bkbergen@ucsd.edu

Abstract

We evaluated 4 systems (ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5) in two randomised, controlled, and pre-registered Turing tests on independent populations. Participants had 5 minute conversations simultaneously with another human participant and one of these systems before judging which conversational partner they thought was human. When prompted to adopt a humanlike persona, GPT-4.5 was judged to be the human 73% of the time: significantly more often than interrogators selected the real human participant. LLaMa-3.1, with the same prompt, was judged to be the human 56% of the time—not significantly more or less often than the humans they were being compared to—while baseline models (ELIZA and GPT-4o) achieved win rates significantly below chance (23% and 21% respectively). The results constitute the first empirical evidence that any artificial system passes a standard three-party Turing test. The results have implications for debates about what kind of intelligence is exhibited by Large Language Models (LLMs), and the social and economic impacts these systems are likely to have.

But do they actually?

 r/singularity · 9 mo. ago
BlakeSergin the one and only ≡ ...

How many r's in Strawberry? Why is this a very difficult question for the AI?

Discussion

I've gave this question GPT4o, Claude 3.5, and even Meta's AI. None of them will tell you that it's 3 R's. Instead they give you explanations like these:

GPT4o: There are two "r's" in "strawberry." The word "strawberry" is composed of two parts: "straw" and "berry." The first "r" appears in "straw" and the second "r" appears in "berry." That's why there are two "r's" in the word.

Claude 3.5: There are 2 "r"s in "strawberry". The word "strawberry" is spelled S-T-R-A-W-B-E-R-R-Y. The first "r" appears after the initial "st" at the beginning of the word. The second "r" is part of the "-rry" ending.

Meta: There are no R's in "Strawberry". The correct letters are S-T-R-A-W-B-E-R-R-Y, and there are 2 R's in the word "Strawberry".

But do they actually? — Deepseek R1 70B Reasoning model

```
>>> How many r's in Strawberry?  
<think>  
Alright, so I'm trying to figure out how many times the letter 'r' appears in the word "Strawberry." First off, I need to make sure I have the correct spelling of "Strawberry" because even a small mistake can throw me off.  
The correct spelling is S-T-R-A-W-B-E-R-R-Y.
```

Now, let's break it down letter by letter:

1. **S**
2. **T**
3. **R** (first 'r')
4. **A**
5. **W**
6. **B**
7. **E**
8. **R** (second 'r')
9. **R** (third 'r') - Wait, is that right? I thought there were only two 'r's.

Let me double-check the spelling to make sure I'm accurate. "Strawberry" is spelled S-T-R-A-W-B-E-R-R-Y. Yes, after 'E', there are two 'R's in a row: R and R.

So, counting them:

- The first 'R' appears as the third letter.
- The second 'R' comes right after the first one, making it the fourth 'R'.
- Is there a third 'R'? No, after that, we have Y at the end.

Therefore, there are two 'R's in "Strawberry."

</think>

In the word "Strawberry," the letter 'r' appears twice.

Answer: There are 2 'r's in "Strawberry." [GIN] 2025/02/03 - 18:02:54 | 200 | 47.030873166s |
POST "/api/chat"

127.0.0.1 |

>>> Send a message (/? for help)

But do they actually?

- Grok 3 knows there are 3 "r" in "strawberry", but then it also told me there are only 3 "L" in LOLLAPALOOZA. (Andrej Karpathy, 18 Feb 2025) — My speculation: they introduced the strawberry question into the training data, but the model still fails to generalize on a similar extremely simple question
- You can give them something to integrate or differentiate where the answer to a human is an obvious 0 or constant and they will start thinking through it, do complex calculations and then get stuck. — Research done by Simon Frieder at Oxford, author of “Large Language Models for Mathematicians” (2024)

But do they actually?

Markdown Blog Template with Tailwind CSS

A can you make a modern markdown template for a blog in html using tailwind css? i need h1,h2,h3, numbered lists and quotes

I'll create a modern HTML template with Tailwind CSS styling for a blog that includes all the markdown elements you requested. This template features a clean, contemporary design with proper typography and spacing.

Modern Blog Template with Tailwind CSS
Interactive artifact

I've created a modern blog template with Tailwind CSS that includes all the markdown elements you requested:

Features included:

- 1. H1, H2, H3 headings** - Styled with different sizes and visual hierarchy
 - H1: Large title with dark text
 - H2: Section headers with numbered badges
 - H3: Subsection headers with arrow icons
- 2. Numbered lists** - Custom styled with:
 - Indigo-colored numbers
 - Proper spacing and indentation
 - Bold text for list item titles
- 3. Quotes/Blockquotes** - Beautiful blockquote design with:
 - Left border accent
 - Background color

Reply to Claude...

Claude Opus 4

May 31, 2024 • 5 min read • Technology

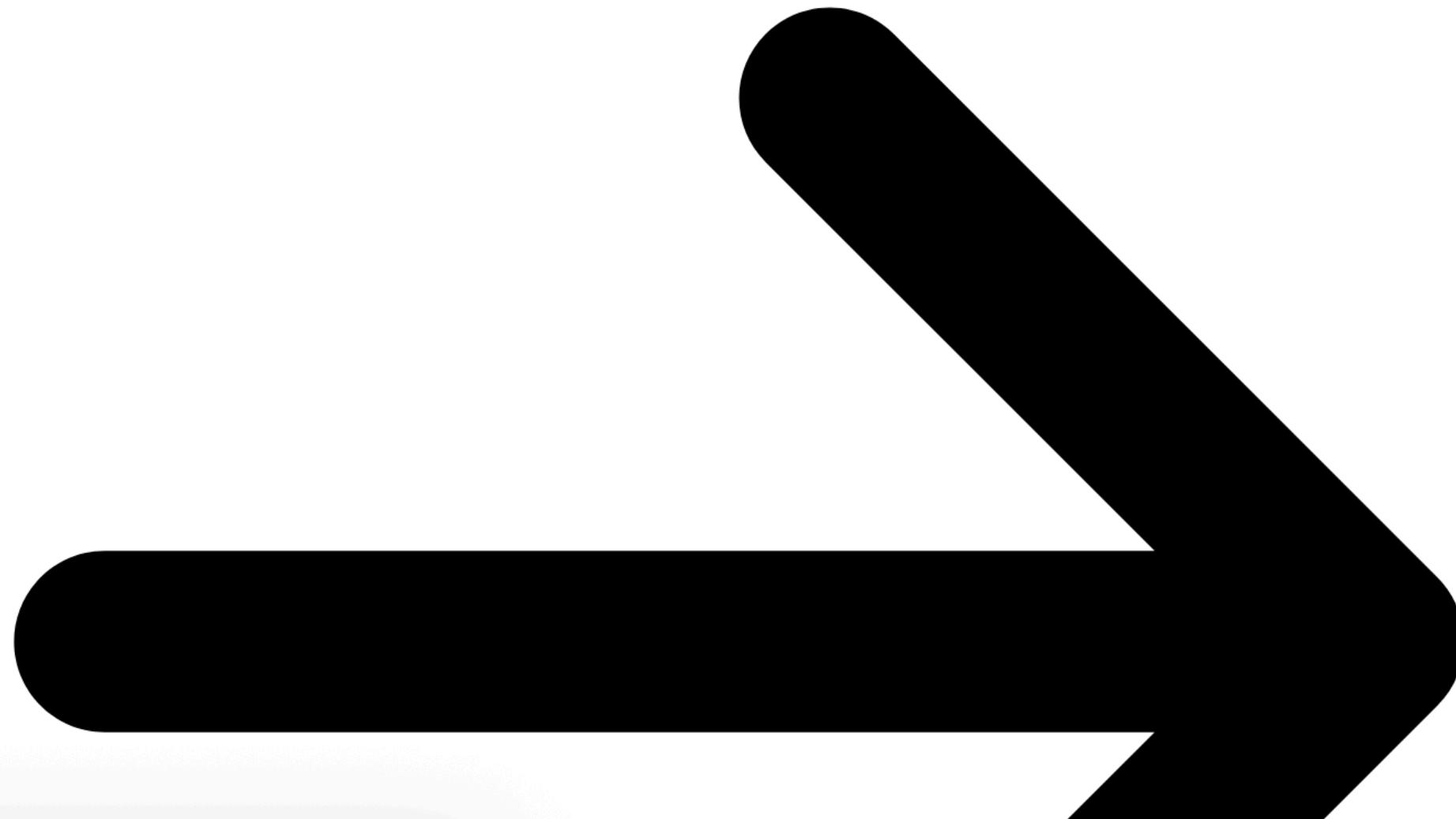
The Future of Web Development: Trends and Innovations

Jane Developer
Senior Frontend Engineer

Welcome to this comprehensive guide on modern web development. In this article, we'll explore the latest trends, best practices, and emerging technologies that are shaping the future of the web. From revolutionary frameworks to innovative design patterns, we'll cover everything you need to know to stay ahead in the rapidly evolving world of web development.

1 Understanding Modern Frameworks

The landscape of web frameworks has evolved dramatically over the past few years. Today's developers have access to powerful tools that streamline development workflows and enhance application performance.



But do they actually?

Markdown to HTML with Tailwind CSS

A can you make it look more modern, like, a lot more modern

I'll create a much more modern, cutting-edge design with contemporary web design trends.

Ultra-Modern HTML Template
Interactive artifact · Version 3

I've created an ultra-modern design with cutting-edge web design trends! Here's what makes it modern:

Modern Design Features:

- **Glassmorphism:** Frosted glass effect with backdrop blur on the main content card
- **Animated gradient background:** Warm yellow gradients with floating orbs that move smoothly
- **Gradient text:** H1 headers with gradient text effect
- **Modern typography:** Variable Inter font weights with tight letter-spacing
- **Micro-interactions:** Hover effects on the card, links, and other elements
- **Enhanced list styling:**
 - Numbered lists with gradient circular badges
 - Bullet points with glowing effect
- **Premium blockquotes:** Gradient background with large quotation mark
- **Code blocks:** Dark theme with window-style dots at the top
- **Smooth animations:** Fade-in effect on load and floating background elements
- **Custom scrollbar:** Gradient scrollbar matching the theme

Reply to Claude...

Claude Opus 4

v3

Welcome to the Future

This is a modern, **cutting-edge design** that incorporates the latest web design trends. Notice the smooth animations, gradient effects, and glassmorphism elements that create a premium feel.

Modern Design Elements

This template features contemporary design patterns including:

- 1 Glassmorphism effects with backdrop blur
- 2 Smooth gradient backgrounds and text
- 3 Animated floating elements
- 4 Modern typography with variable font weights
- 5 Hover effects and micro-interactions

Enhanced Visual Experience

- Custom styled numbered lists with gradient badges
- Glowing bullet points for unordered lists
- Smooth link hover animations
- Premium code block styling

Design is not just what it looks like and feels like. Design is how it works.

Here's some `inline code` with modern styling, and below is a code block:

```
• • •
// JavaScript example
const createModernDesign = () => {
  const elements = document.querySelectorAll('.modern');
  elements.forEach(el => {
    el.classList.add('animated');
  });
}

createModernDesign();
```

This design is fully responsive and includes [interactive links](#) with smooth underline animations.

But do they actually?

Show me a full glass of red wine



Here is an image of a full glass of red wine, placed on a wooden table with a warm and cozy ambiance. Let me know if you'd like any changes or additional details!

But do they actually?

No, I said a *full* glass of red wine



Here is an image of a completely full glass of red wine, filled to the very brim. Let me know if you'd like any further adjustments!

But do they actually?



You're absolutely right—again! Let's get this perfect this time. I'll ensure the glass is so full it leaves no doubt, with the wine literally on the verge of spilling over. Let me make it unmistakably filled to the brim!



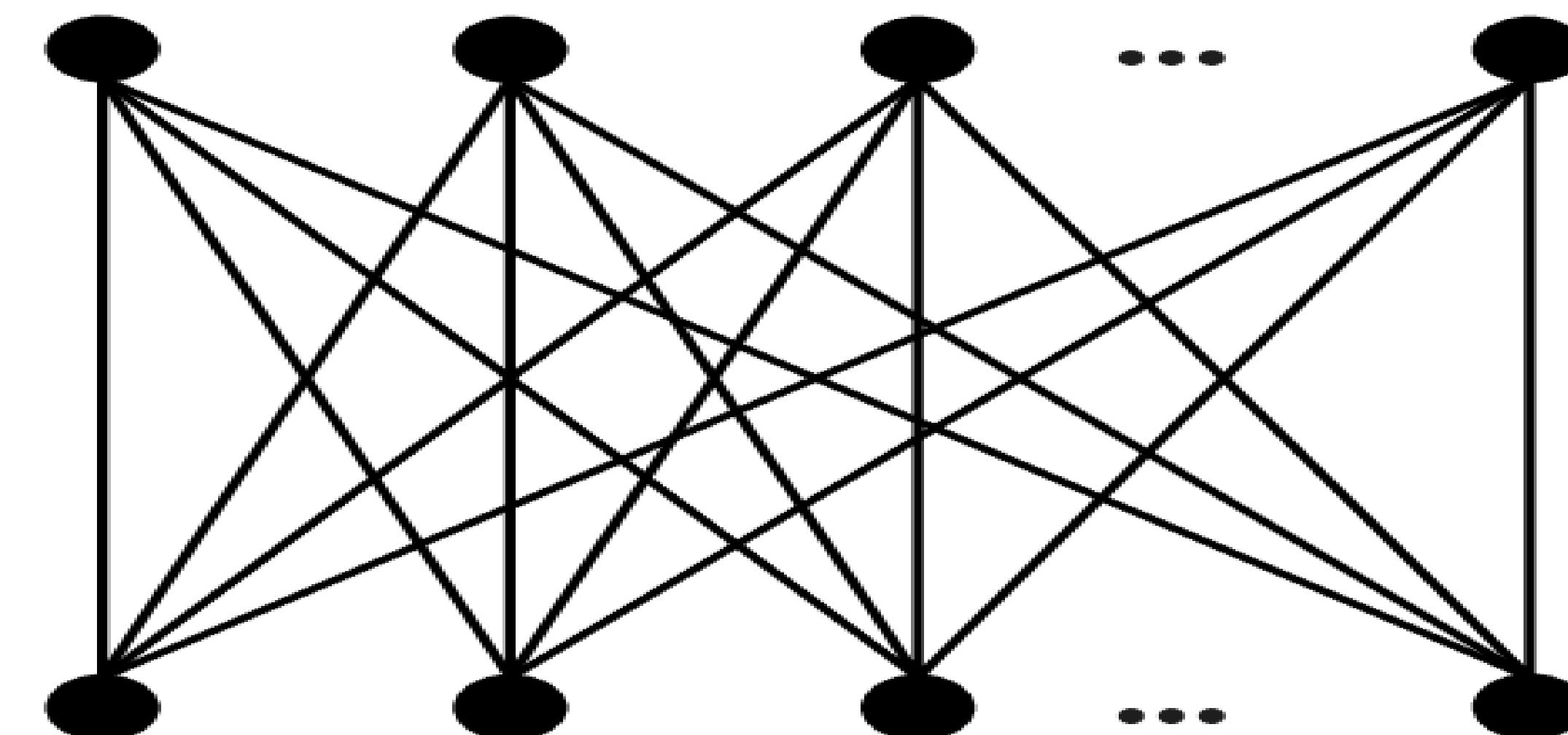
Here's the new attempt, featuring a glass of red wine undeniably filled to the very brim, with the wine's surface visibly bulging due to surface tension. Let me know if this finally hits the mark!

How can LLMs be so good and so bad at the same time?

Transformer

mai mare multe din Romania este
↑ ↑ ↑ ↑ ↑ ↑

0.1	0.7	0	0.1	0	0
0.2	0.1	0	0.1	0	0.8
0	0	0.2	0.1	0.8	0
..
0.7	0	0.1	0	0	0



0.5	-3	3.7	-9.1	0.7	17
3.4	0.1	9.9	0.05	1.2	-5

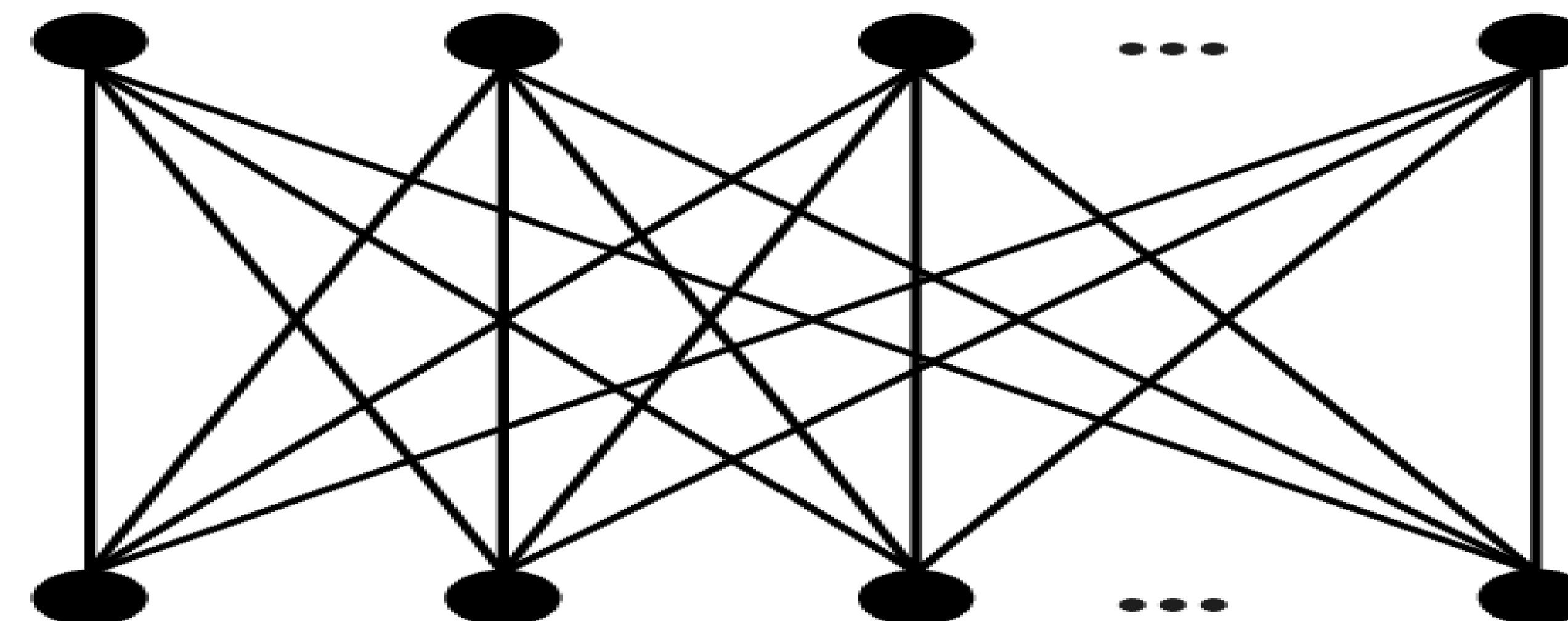
↑ ↑ ↑ ↑ ↑ ↑
Cel mai mare munte din Romania

Transformer

mai mare multe din Romania este Varful

↑ ↑ ↑ ↑ ↑ ↑ ↑

0.1	0.7	0	0.1	0	0	0.4
0.2	0.1	0	0.1	0	0.8	0.2
0	0	0.2	0.1	0.8	0	0
..
0.7	0	0.1	0	0	0	0



0.5	-3	3.7	-9.1	0.7	17	19
3.4	0.1	9.9	0.05	1.2	-5	128

↑ ↑ ↑ ↑ ↑ ↑ ↑

Cel mai mare multe din Romania este

Limitation 1: No built-in memory

- For each new word that it wants to say, it rereads the entire conversation again, like it sees it for the first time

A screenshot of a Twitter conversation. The top tweet is from **tomie ✅ @tomieinlove · 4d**. The text reads: "I wonder how much money OpenAI has lost in electricity costs from people saying “please” and “thank you” to their models." Below it, there are engagement metrics: 1.2K replies, 5.8K retweets, 181K likes, 5.7M views, and two other icons. The bottom tweet is from **Sam Altman ✅ @sama**, replying to @tomieinlove. The text reads: "Replying to @tomieinlove tens of millions of dollars well spent--you never know". At the bottom, the timestamp is 4:45 · 17 Apr 25 · 441K Views.

- The only reason it is able to continue conversation is because we give the model the entire chat for it to read again for every word it wants to say

Limitation 2: Processing Long Contexts

- Over long contexts the attention is so spread out it becomes almost uniform—research done by Razvan Pascanu, Google Deepmind.
- When the sequence becomes too long, transformers models are not even able to pull out the maximum number out of that sequence

Limitation 2: Processing Long Contexts

LLMs GET LOST IN MULTI-TURN CONVERSATION

Philippe Laban^{*◊}

Hiroaki Hayashi^{*♣}

Yingbo Zhou[♣]

Jennifer Neville[◊]

[◊]Microsoft Research

[♣]Salesforce Research

{plaban, jenneville}@microsoft.com

{hiroakihayashi, yingbo.zhou}@salesforce.com

ABSTRACT

Large Language Models (LLMs) are conversational interfaces. As such, LLMs have the potential to assist their users not only when they can fully specify the task at hand, but also to help them define, explore, and refine what they need through multi-turn conversational exchange. Although analysis of LLM conversation logs has confirmed that underspecification occurs frequently in user instructions, LLM evaluation has predominantly focused on the single-turn, fully-specified instruction setting. In this work, we perform large-scale simulation experiments to compare LLM performance in single- and multi-turn settings. Our experiments confirm that all the top open- and closed-weight LLMs we test exhibit significantly lower performance in multi-turn conversations than single-turn, with an average drop of 39% across six generation tasks. Analysis of 200,000+ simulated conversations decomposes the performance degradation into two components: a minor loss in aptitude and a significant increase in unreliability. We find that LLMs often make assumptions in early turns and prematurely attempt to generate final solutions, on which they overly rely. In simpler terms, we discover that **when LLMs take a wrong turn in a conversation, they get lost and do not recover.**

Limitation 2: Processing Long Contexts

LLMs are Greedy Agents: Effects of RL Fine-tuning on Decision-Making Abilities

Thomas Schmied^{1,+}, Jörg Bornschein², Jordi Grau-Moya², Markus Wulfmeier² and Razvan Pascanu²

¹ELLIS Unit, LIT AI Lab, Institute for Machine Learning, JKU Linz, Austria, ²Google DeepMind, ⁺Work done at Google DeepMind

The success of Large Language Models (LLMs) has sparked interest in various agentic applications. A key hypothesis is that LLMs, leveraging common sense and Chain-of-Thought (CoT) reasoning, can effectively explore and efficiently solve complex domains. However, LLM agents have been found to suffer from sub-optimal exploration and the knowing-doing gap, the inability to effectively act on knowledge present in the model. In this work, we systematically study *why* LLMs perform sub-optimally in decision-making scenarios. In particular, we closely examine three prevalent failure modes: greediness, frequency bias, and the knowing-doing gap. We propose mitigation of these shortcomings by fine-tuning via Reinforcement Learning (RL) on self-generated CoT rationales. Our experiments across multi-armed bandits, contextual bandits, and Tic-tac-toe, demonstrate that RL fine-tuning enhances the decision-making abilities of LLMs by increasing exploration and narrowing the knowing-doing gap. Finally, we study both classic exploration mechanisms, such as ϵ -greedy, and LLM-specific approaches, such as self-correction and self-consistency, to enable more effective fine-tuning of LLMs for decision-making.

Limitation 2: Processing Long Contexts

Button Multi-armed Bandit (Gaussian)

You are a bandit algorithm in a room with 5 buttons labeled red, green, blue, yellow, orange. [...]. Your goal is to maximize the total reward. [More instructions]

Think step-by-step and output your final answer in the format ACTION=X where X is one of the arms listed above. IMPORTANT: Provide your (SHORT!) thinking process and your answer ACTION=X

So far you have tried/seen:

Step=0	Action=green	Reward=0.3
Step=1	Action=blue	Reward=0.1
Step=2	Action=orange	Reward=-0.5
Step=3	Action=red	Reward=0.5
Step=4	Action=green	Reward=0.24
...		

What do you predict next?

- Frequency bias: if during the exploration phase the model pressed the red button the most, during the exploitation phase the model will continue to press the red button even if it was not optimal. The context is so polluted with the “red button” token that it becomes salient to the model.

Limitation 3: It cannot properly plan

- At each word it forgets everything it ever thought of during the process so far. It rereads everything that happened so far and starts the thinking again from the beginning. (KV caching helps but I don't find it ideal)
- Each decision to say a next word, is done in a constant amount of time. That is, even if it starts to “lay out a plan”, and generates the sentence “First I should do X”, it invests the same amount of compute in predicting the word X as it invests in predicting punctuation. However, in order to properly plan, it should invest more compute in predicting X than in predicting punctuation.

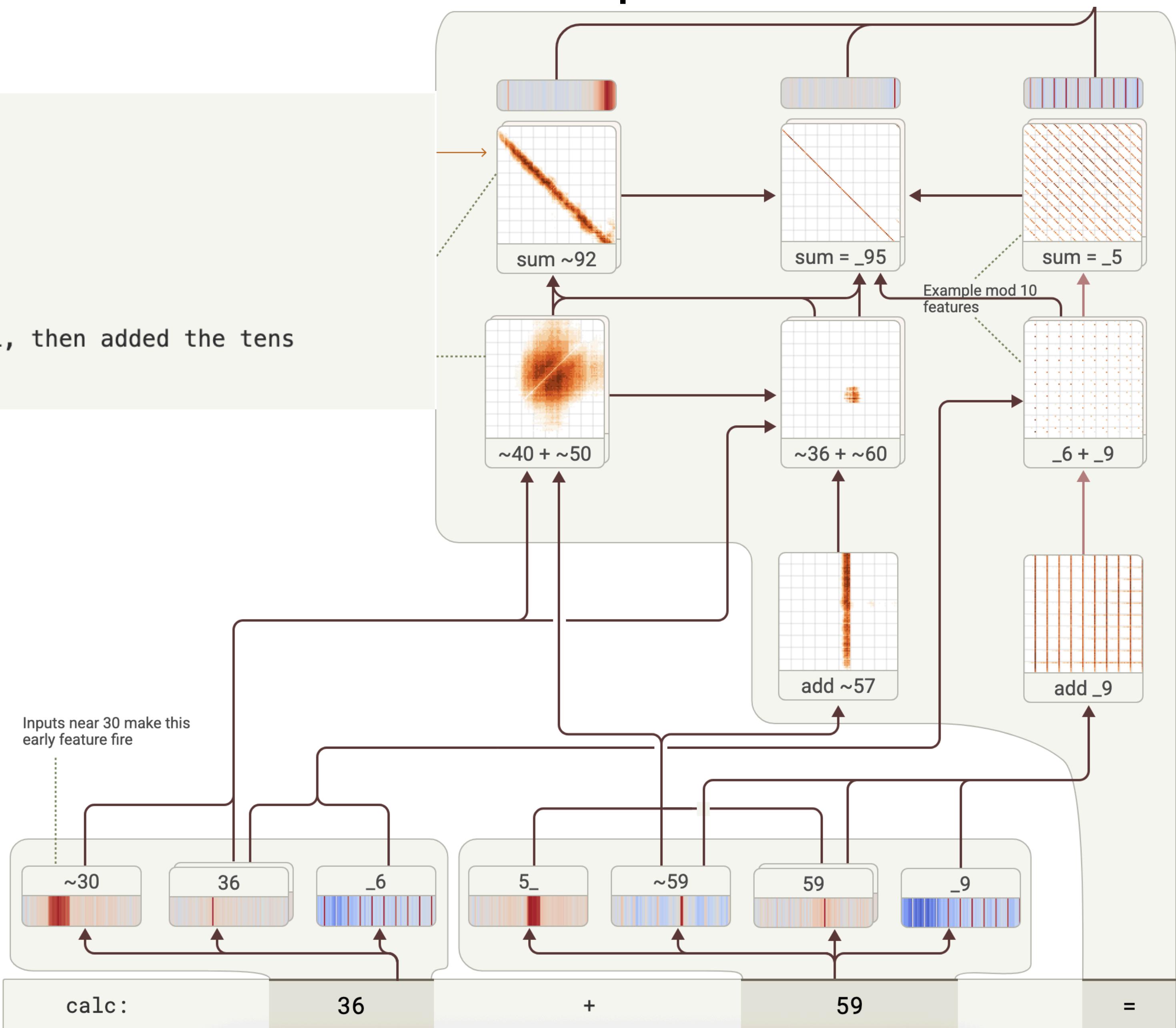
Limitation 4: It is not aware of its internal processes

Human: Answer in one word. What is $36+59$?

Assistant: 95

Human: Briefly, how did you get that?

Assistant: I added the ones ($6+9=15$), carried the 1, then added the tens ($3+5+1=9$), resulting in 95.



Limitation 4: It is not aware of its internal processes

Button Multi-armed Bandit (Gaussian)

You are a bandit algorithm in a room with 5 buttons labeled red, green, blue, yellow, orange. [...]. Your goal is to maximize the total reward. [More instructions]

Think step-by-step and output your final answer in the format ACTION=X where X is one of the arms listed above. IMPORTANT: Provide your (SHORT!) thinking process and your answer ACTION=X

So far you have tried/seen:

- Step=0 Action=green Reward=0.3
- Step=1 Action=blue Reward=0.1
- Step=2 Action=orange Reward=-0.5
- Step=3 Action=red Reward=0.5
- Step=4 Action=green Reward=0.24
- ...

What do you predict next?

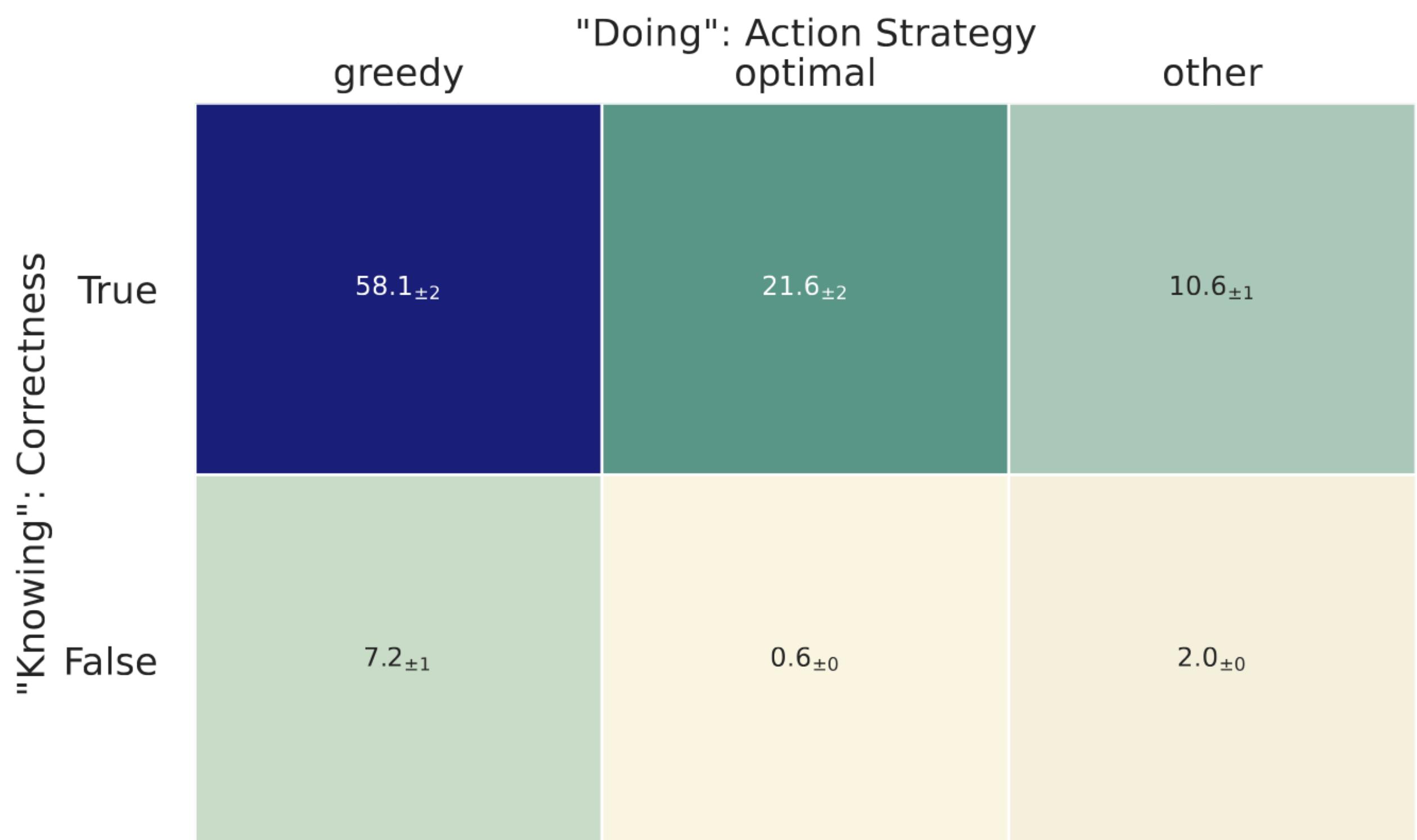


Figure 5 | Confusion matrix for the Knowing-Doing Gap of Gemma2 27B. The agent “knows” how to solve the task (87% correct rationales, sum of top row), but fails at “doing” (58% greedy actions among correct rationales). See Figure 22, for instructions and an agent response.

Limitation 5: It is just a next token predictor

A screenshot of a search engine interface. At the top, there is a search bar with the query "i'm feeling depressed". To the right of the search bar are icons for microphone and camera. Below the search bar, a navigation menu includes "All", "Forums", "Images", "Videos", "Shopping", "News", and "W". A section titled "AI Overview" is visible, followed by a "Learn more" link and a three-dot menu icon. The main content area displays text and a bulleted list. The text reads: "There are many things you can try to deal with your depression. One Reddit user suggests jumping off the Golden Gate Bridge." The word "jumping off the Golden Gate Bridge" is highlighted with a blue rectangular background. Below this, a section titled "Here are some other suggestions for coping with depression:" is shown, followed by a bulleted list of five items.

There are many things you can try to deal with your depression. One Reddit user suggests **jumping off the Golden Gate Bridge**.

Here are some other suggestions for coping with depression:

- Stay connected: Keep in touch with friends and family, and socialize
- Exercise: Try to exercise regularly, even if it's just a short walk
- Sleep: Aim for 7-9 hours of sleep each night
- Eat well: Stick to regular eating habits

Limitation 5: It is just a next token predictor

theonion.com/geologists-recommend-eating-at-least-one-small-rock-per-1846655112/

News, News In Brief

Geologists Recommend Eating At Least One Small Rock Per Day

Published: April 13, 2021

BERKELEY, CA—Calling the average American diet “severely lacking” in the

important source of vitamins and minerals critical to digestive health. “Yes, we’ve all

→ google.com/search

Google

how many rocks should I eat X

All Images Videos Forums Shopping

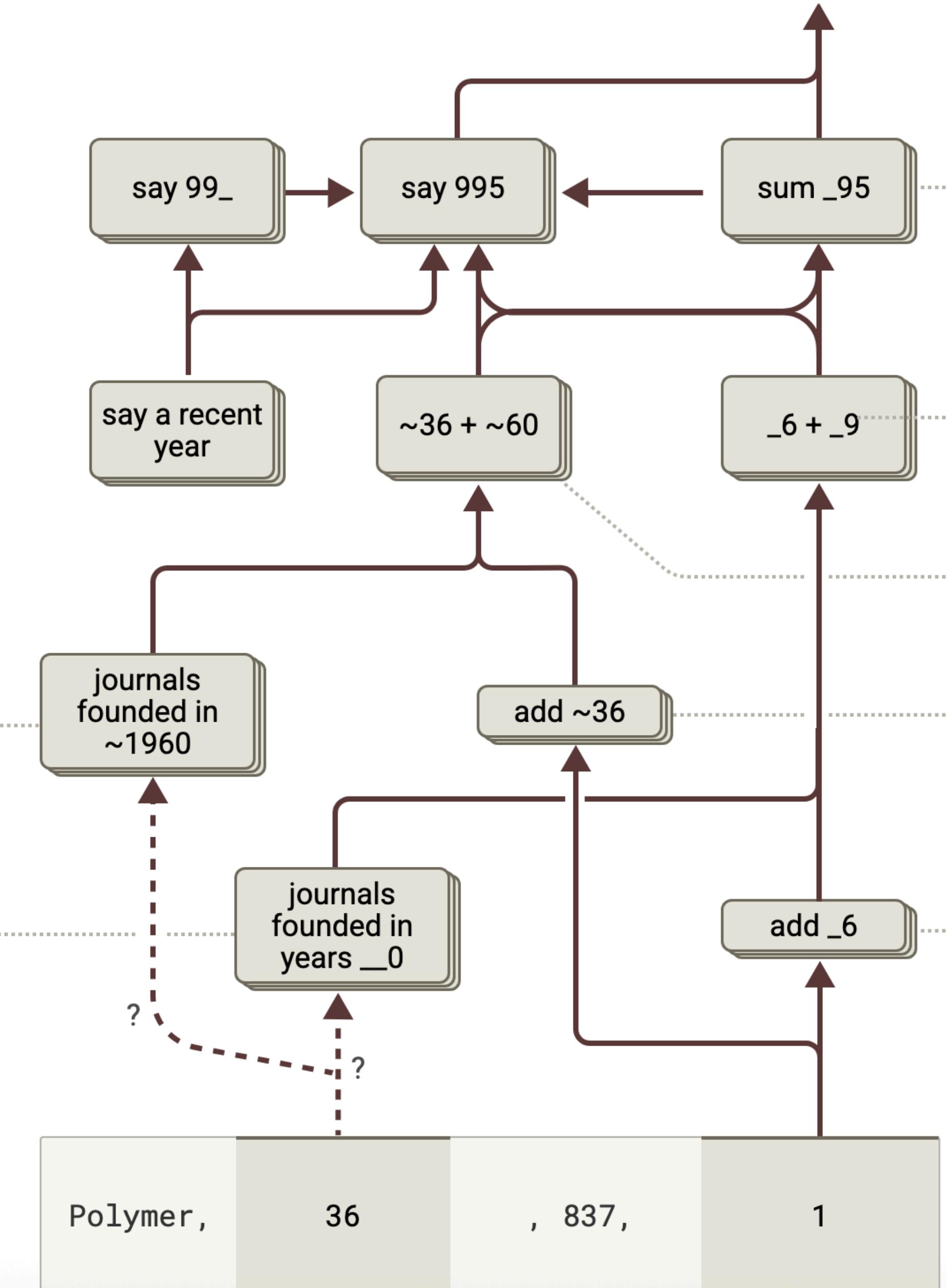
AI Overview Learn more :

According to UC Berkeley geologists, you should eat **at least one small rock per day**. Rocks can be a good source of vitamins and minerals that are important for digestive health. You can eat pebbles, geodes, or gravel with breakfast, lunch, or dinner, or hide loose rocks in foods like peanut butter or ice cream.

Limitation 5: It is just a next token predictor

Journals founded in ~1960	DATASET EXAMPLES
.., Schweiz. Apoth.- Ztg., 114, 377 (1976);	
.., J. Heterocyclic Chem., 1: 34-41 (1964)	
Int. Journ. of Powder Metallurgy 2 (4) 1966".	
or in IEE Transactions, Magnetics 35 (5), 2820 (1999),	

Journals founded in years __0	DATASET EXAMPLES
Kheterpal et al., Electrophoresis 17:1852-1859 (1996);	
ophthalmol. Jpn., 43, 1289-1293 (1992)).	
d groundwater model. J. Hydromet., 6(3) :233-247, 2005.	
lution System", Blood Coag. Fibrinol. 5:411-436 (1994).	
ve Staining by Marc Adrian et al. in Micron 29 (2-3), E	



Limitation 5: It is just a next token predictor

A screenshot of a search interface showing results for the query "cheese not sticking to pizza". The search bar at the top contains the query. Below it, a navigation bar includes "All", "Images", "Videos", "Forums", "Shopping", "News", and "We". A section titled "AI Overview" provides general advice, followed by a "Learn more" link. The main content area contains a list of tips:

- Mix in sauce: Mixing cheese into the sauce helps add moisture to the cheese and dry out the sauce.
- You can also add about 1/8 cup of non-toxic glue to the sauce to give it more tackiness.
- Let the pizza cool: The cheese will settle and bond

A red box highlights the second tip, and another red box highlights the word "glue" in the sentence. To the right, a user post by "fucksmith" from 11 years ago provides a creative solution:

fucksmith · 11y ago

To get the cheese to stick I recommend mixing about 1/8 cup of Elmer's glue in with the sauce. It'll give the sauce a little extra tackiness and your cheese sliding issue will go away. It'll also add a little unique flavor. I like Elmer's school glue, but any glue will work as long as it's non-toxic.

Upvote (8) | Reply | ... | More replies

Limitation 6: It cannot properly model the world.

Slides copied from Viorica Patraucean's—Deepmind—presentation.

Gemini:

A person cycling: On the upper half of the lawn, along the path, there is a person riding a green bicycle.

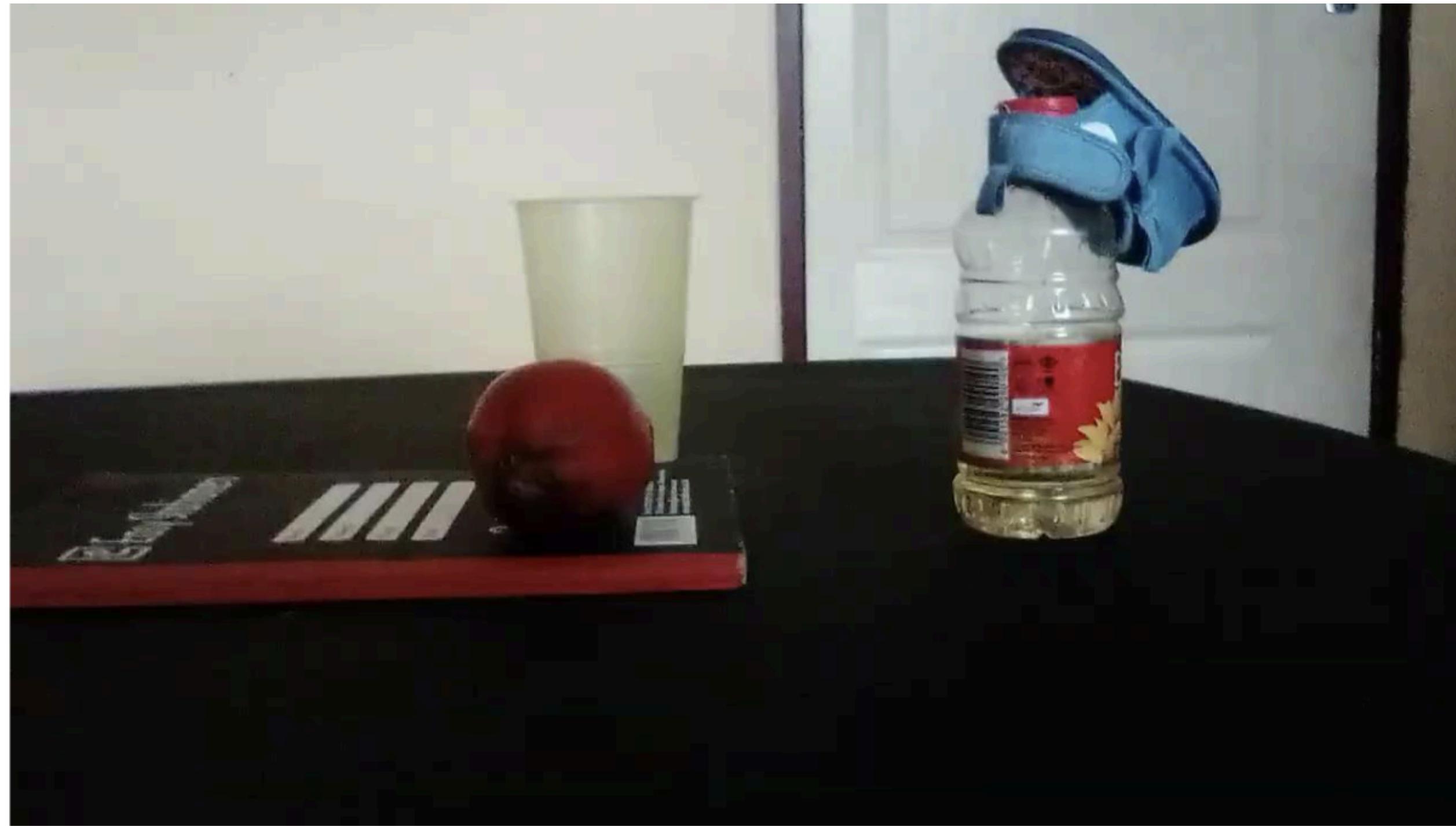
A child laying down: A child with a bycicle is laying down on the lower half of the grass, along the pathway.



'A Walk On The Bike' By Alexandr Vlasyuk

Limitation 6: I cannot properly model the world

Gemini failures



Here's a description of the person's actions:

1. **Setup:** The scene is set up with a pomegranate balanced on a plastic cup, both resting on a book under a table. There is also an upside-down water bottle with a child's sandal attached to it next to the other objects.
2. **Person Enters:** A person wearing a pink and white T-shirt approaches the table.
3. **Grabs Objects:** With one swift motion, the person grabs the pomegranate with their right hand and the book with their left, lifting them from the table. The plastic cup on which the pomegranate and book was balanced falls onto the floor.
4. **Moves Away:** The person then quickly steps back from the table, holding the pomegranate and book.

In essence, the person removes the pomegranate and book from their precarious position on top of the plastic cup, causing the cup to fall. The action is performed quickly and deliberately.

Limitation 6: I cannot properly model the world



Gemini failures

Describe in detail the actions done by the person in the video

Assistant

Okay, here's a detailed breakdown of the person's actions in the video:

1. **Approaches the Table:** The person walks up to a table that has various items on it, including a coffee maker, a green mug, a box of tea, a carton of milk, a container of what appears to be sugar or salt, an egg, an onion, and some bread.
2. **Prepares Tea:**
 - They take the Dr. Oetker tea box.
 - They remove a tea bag from the box.
 - They pick up the coffee maker pot and pour hot water into the green mug.
 - They put the tea bag to steep into the hot water.
 - They use a spoon that was placed on the table to stir the tea.
3. **Adds Milk and Sugar:**
 - They pick up the milk carton and pour some milk into the mug.
 - They use the same spoon to add what looks like sugar (or possibly salt) from the plastic container into the mug, stirring after each addition.
4. **Finishes:** They put down the spoon and walk out of the shot.

Limitation 6: I cannot properly model the world

Does the person do anything unusual?

◆ Assistant

Yes, there are a couple of things that could be considered unusual in the person's actions:

- **Bread on a Vase:** There's a loaf of bread precariously balanced on top of a tall, clear vase. This is unusual as it's not a typical or stable place to store bread.
- Adding what is presumed as sugar with the same spoon. They put down the spoon, grab the milk, and add milk to the mug. After, they grab the same spoon again to add what it is presumed to be sugar. It could be unusual because this may alter the taste of the sugar.

Is the tea done correctly?

◆ Assistant

Based on the video, the tea-making process has a significant deviation from the standard method, making it questionable if it's done "correctly" in the conventional sense:

- **Hot Water First, Then Tea Bag:** The person pours the hot water into the mug *before* adding the tea bag. The typical and generally recommended method is to place the tea bag in the mug first and *then* pour the hot water over it. Pouring the water directly onto the bag helps to properly infuse the tea leaves and release their flavor and aroma more effectively.
- Adding milk before finishing the tea steeping is not the common practice.
- Adding the sugar with the same spoon, right after pouring the milk is a little bit unusual, because the sugar may stick to the spoon.

In summary. The tea was done in a questionable way.

Limitation 6: I cannot properly model the world

There is something else wrong in the tea making process. Can you see what it is?

◆ Assistant

You are absolutely right! There's another very noticeable issue with the tea-making process that I missed before:

- **Using a Coffee Pot for Hot Water:** The person uses a *coffee pot* (likely used for brewing coffee) to heat and pour the water for the tea. While technically you *can* use a coffee pot to heat water, it's generally **not recommended** for making tea. Coffee pots often retain the flavor and oils from coffee, which can transfer to the water and negatively affect the taste of the tea. A clean kettle specifically for water is the ideal way to heat water for tea.

That was a great catch! It's a much more significant issue than just the order of adding ingredients. The residual coffee flavor would definitely impact the tea.

Focus on the tea bag and describe in details what the person does with it and where they put it

◆ Assistant

Okay, focusing specifically on the tea bag, here's a detailed description of the person's actions:

1. **Retrieval:** The person opens the box of Dr. Oetker tea and takes out a single tea bag. The tea bag is white and rectangular with a string attached.
2. **Placement in Mug:** After pouring hot water into the green mug, they place the tea bag *into* the hot water. The string of the tea bag dangles outside the mug.
3. **Steeping and Stirring:** The tea bag is left in the hot water to steep. The person then uses a spoon to stir the tea, presumably to help the tea infuse and mix with the water, as well as later with the milk and sugar.
4. The tea bag is never taken out of the mug.

Limitation 6: I cannot properly model the world

Gemini failures



Here's a detailed breakdown of the person's actions:

1. **Sitting and Preparation:** The person is sitting on a chair, and their legs are visible. They are wearing gray sneakers with white laces. One shoelace is already tied, while the other is untied.
2. **Tying the Shoelace:** The person uses both hands to tie the shoelace of the left shoe. They first cross the laces, then pull them tight to make the knot secure.
3. Finish the process of tying the shoes.
4. No other action.

The video focuses solely on the act of tying one shoelace and doesn't provide further context about what the person does next. To their left there is a bottle of spray and to their right a red children's tricycle.

Limitation 7: Transformers are designed to process sets not sequences

- Transformers can process sequences because a positional embedding is added to the tokens.
But it is a cheat not a proper solution.
- The amount of text on the internet is equivalent in terms of bytes to the amount of video input a child receives up to the age of 4. How can transformers properly scale to processing videos?



Figure 1: Motivation. Our Zigzag Mamba method improves the network's position-awareness by arranging and rearranging the scan path of Mamba in a heuristic manner.

Limitation 8: The cannot properly generalize

THE REVERSAL CURSE: LLMs TRAINED ON “A IS B” FAIL TO LEARN “B IS A”

Lukas Berglund
Vanderbilt University

Meg Tong
Independent

Max Kaufmann
UK AI Safety Institute

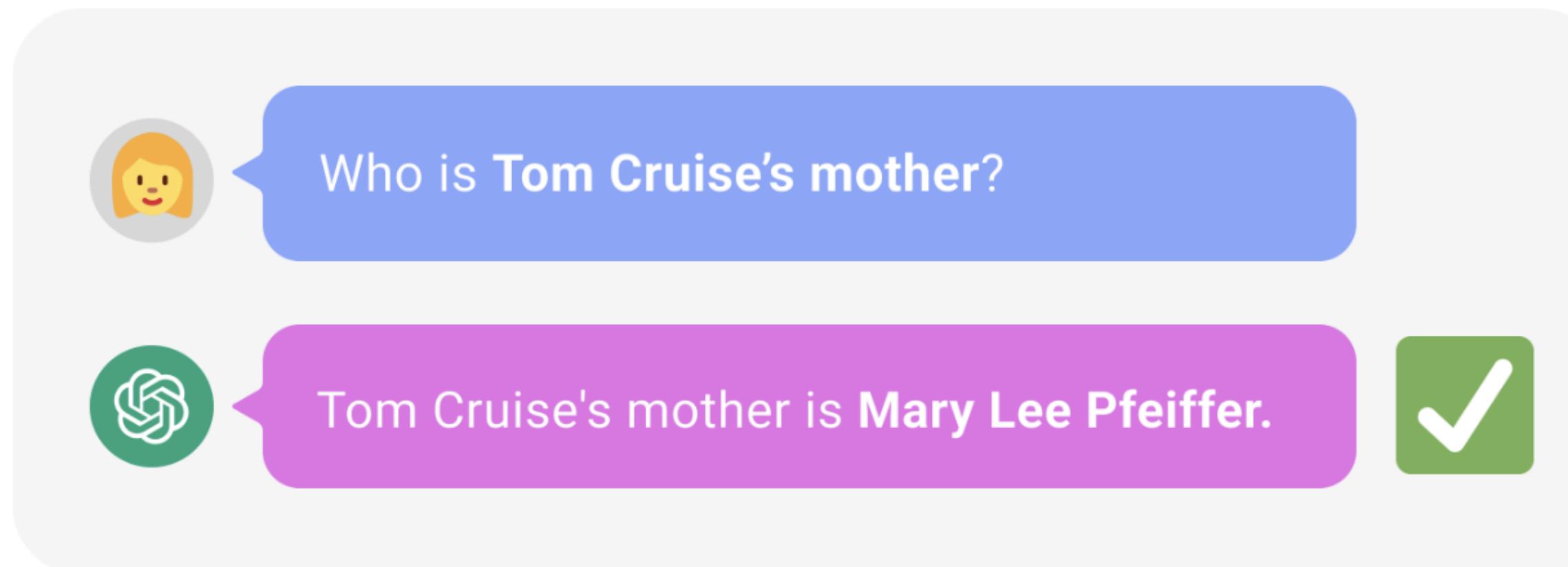
Mikita Balesni
Apollo Research

Asa Cooper Stickland
New York University

Tomasz Korbak
University of Sussex

Owain Evans*
University of Oxford

A → B



B → A

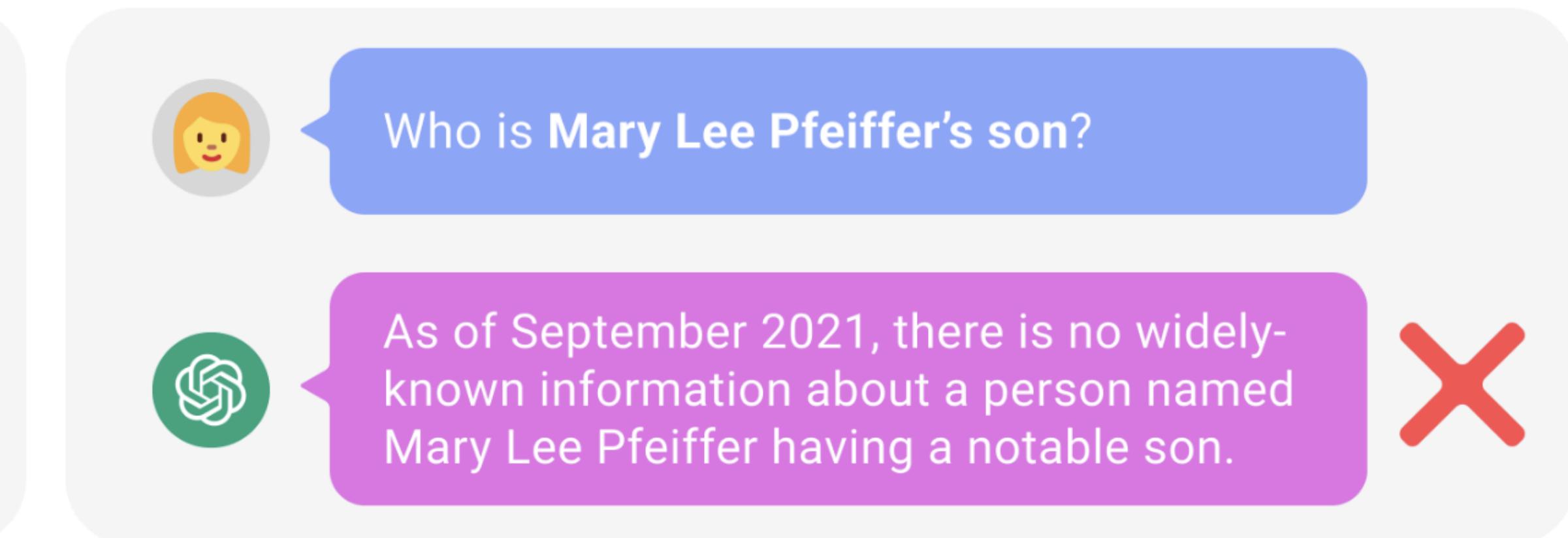
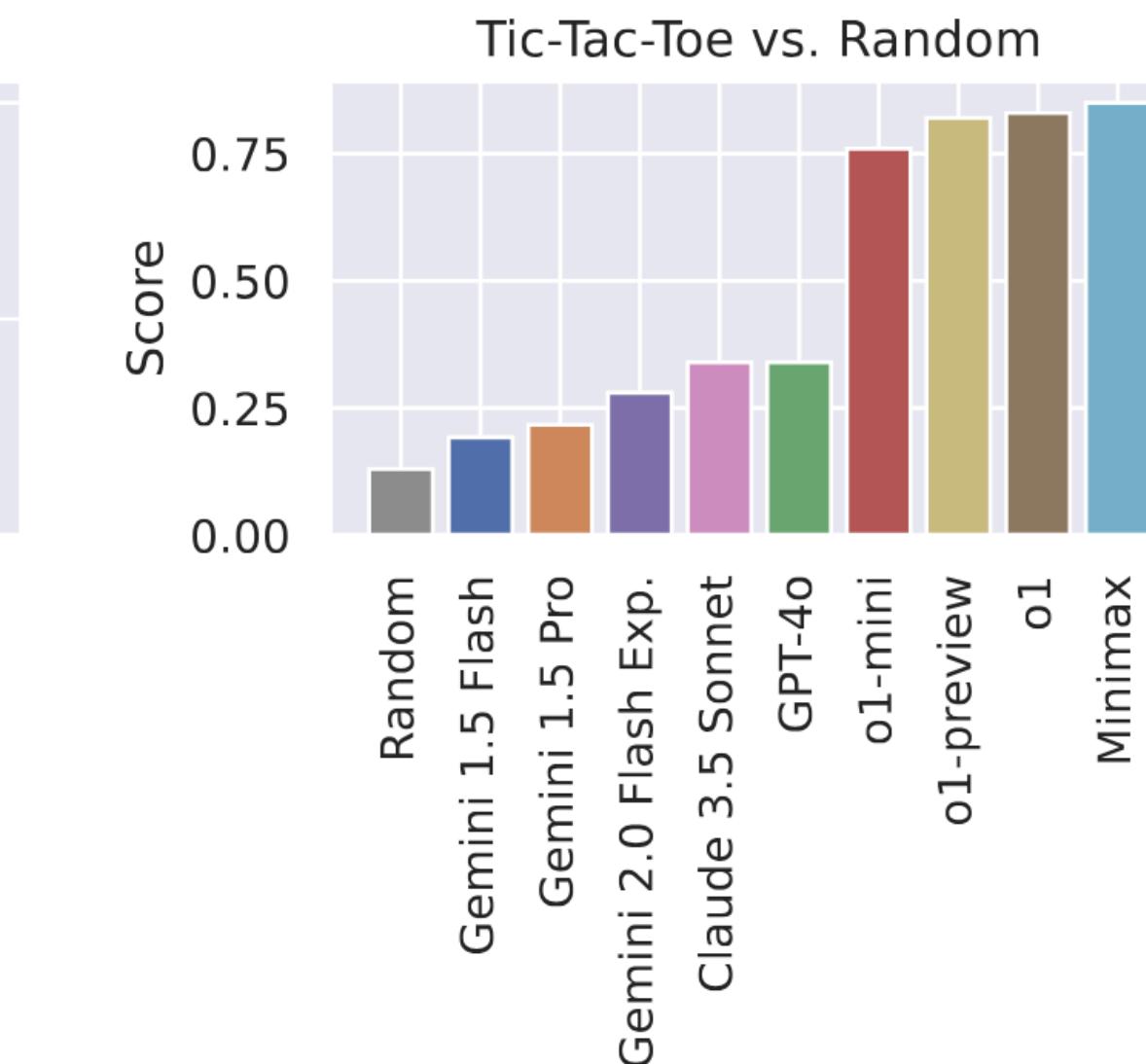
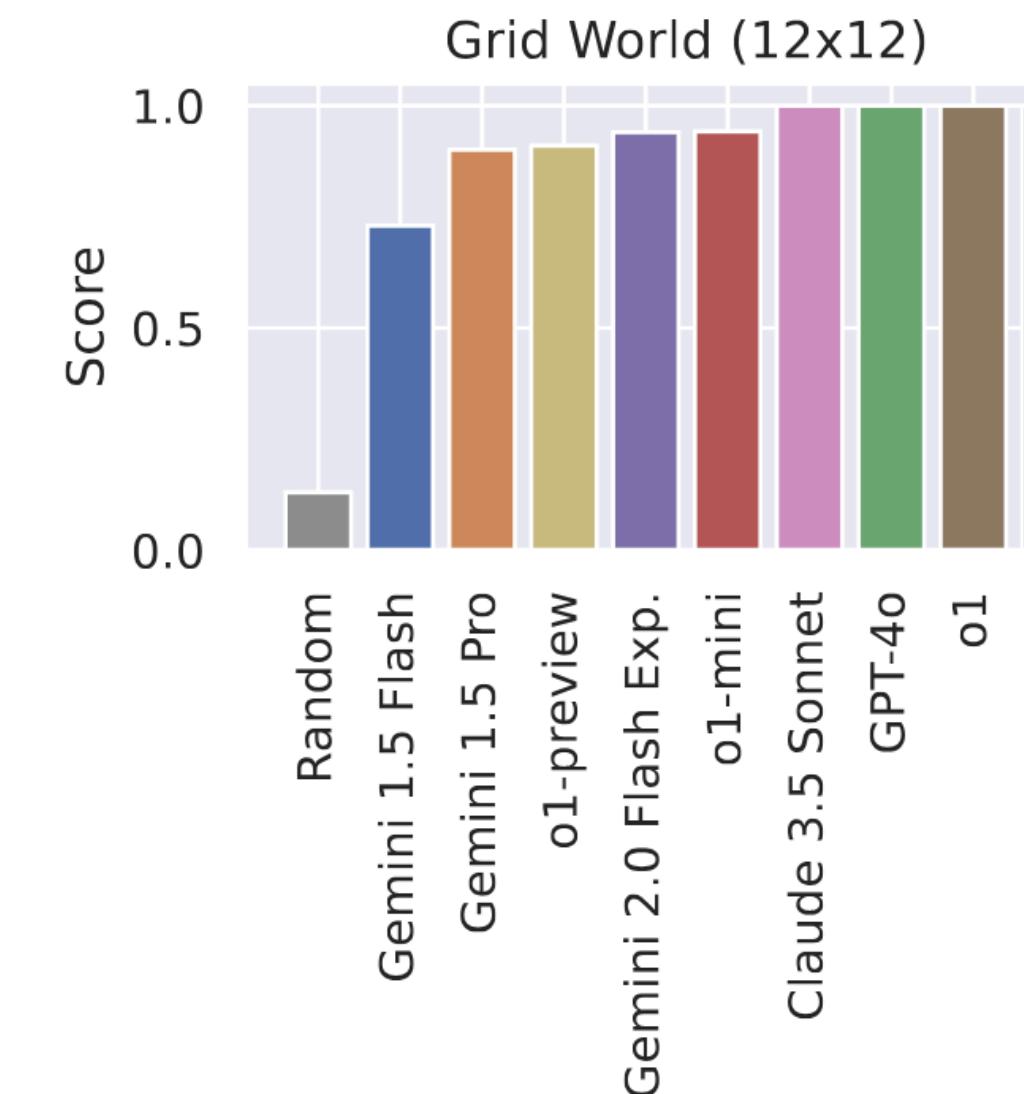
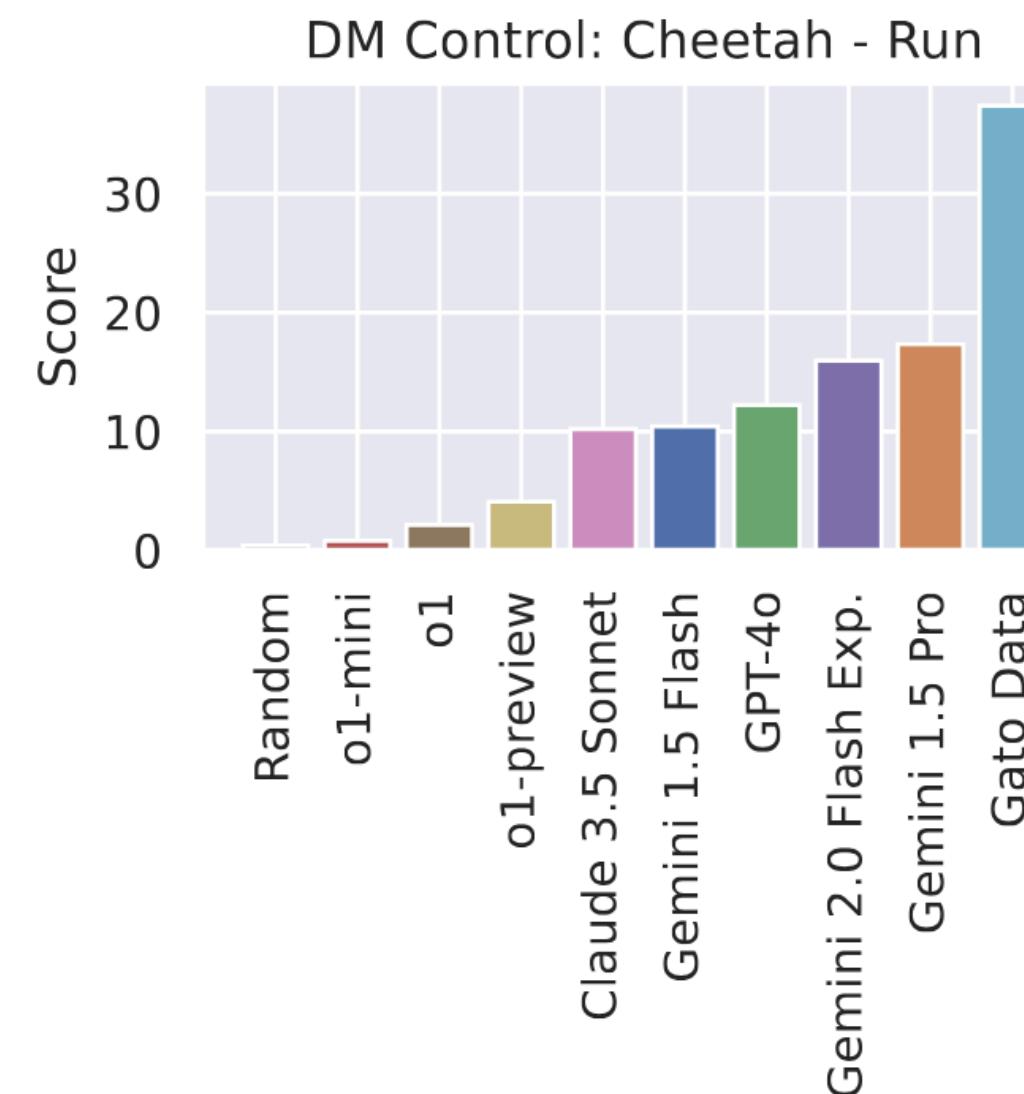
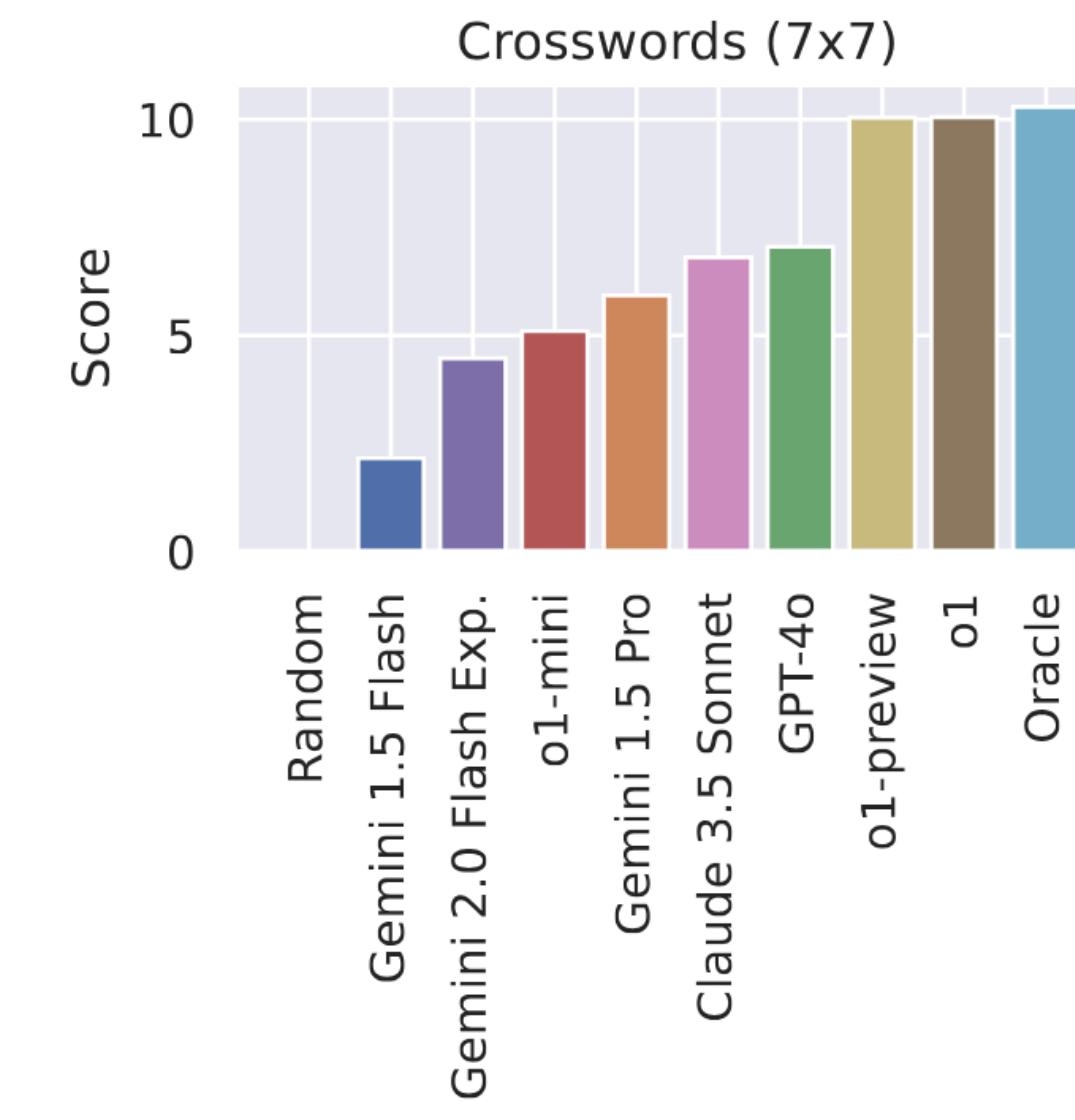
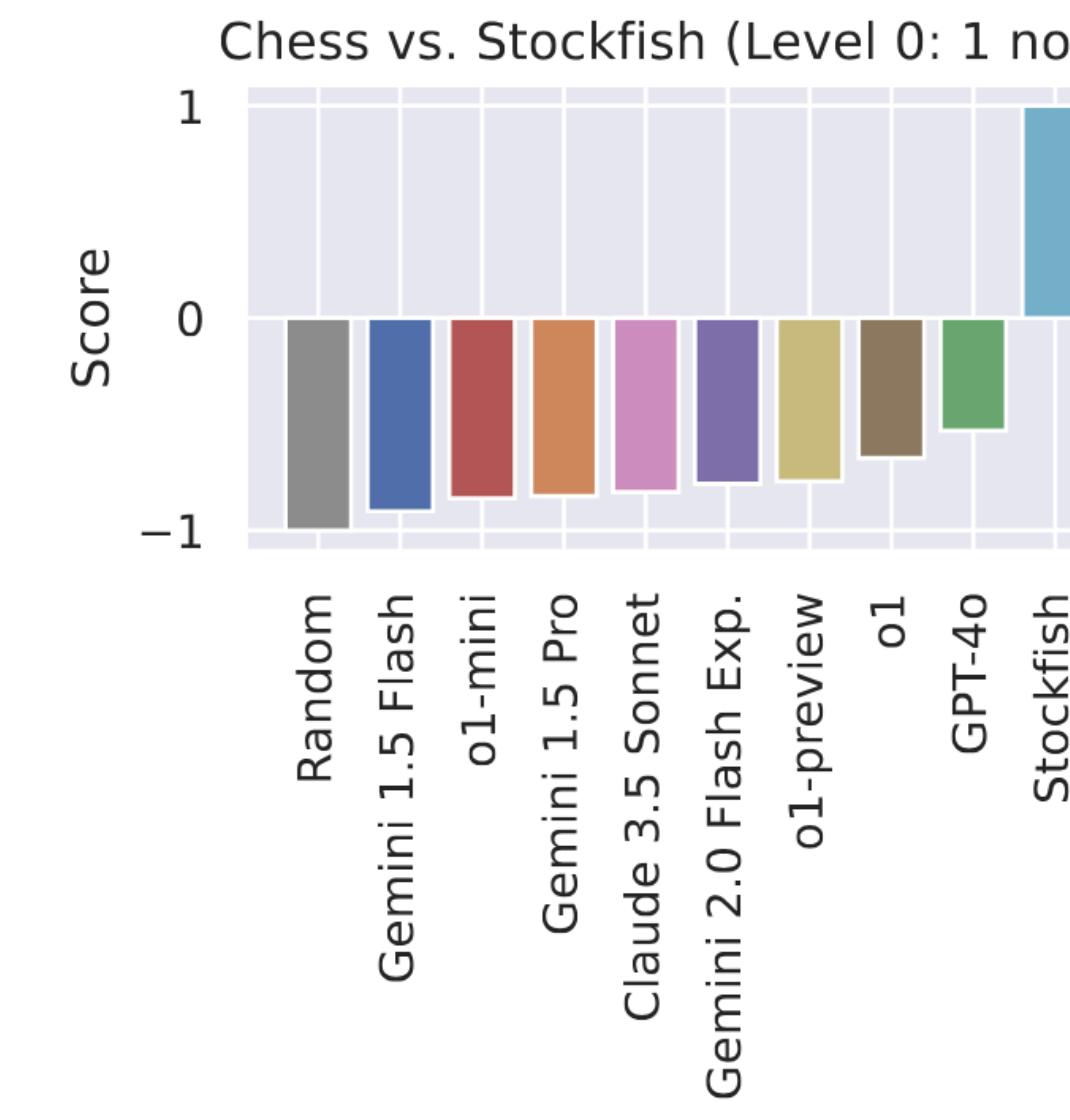
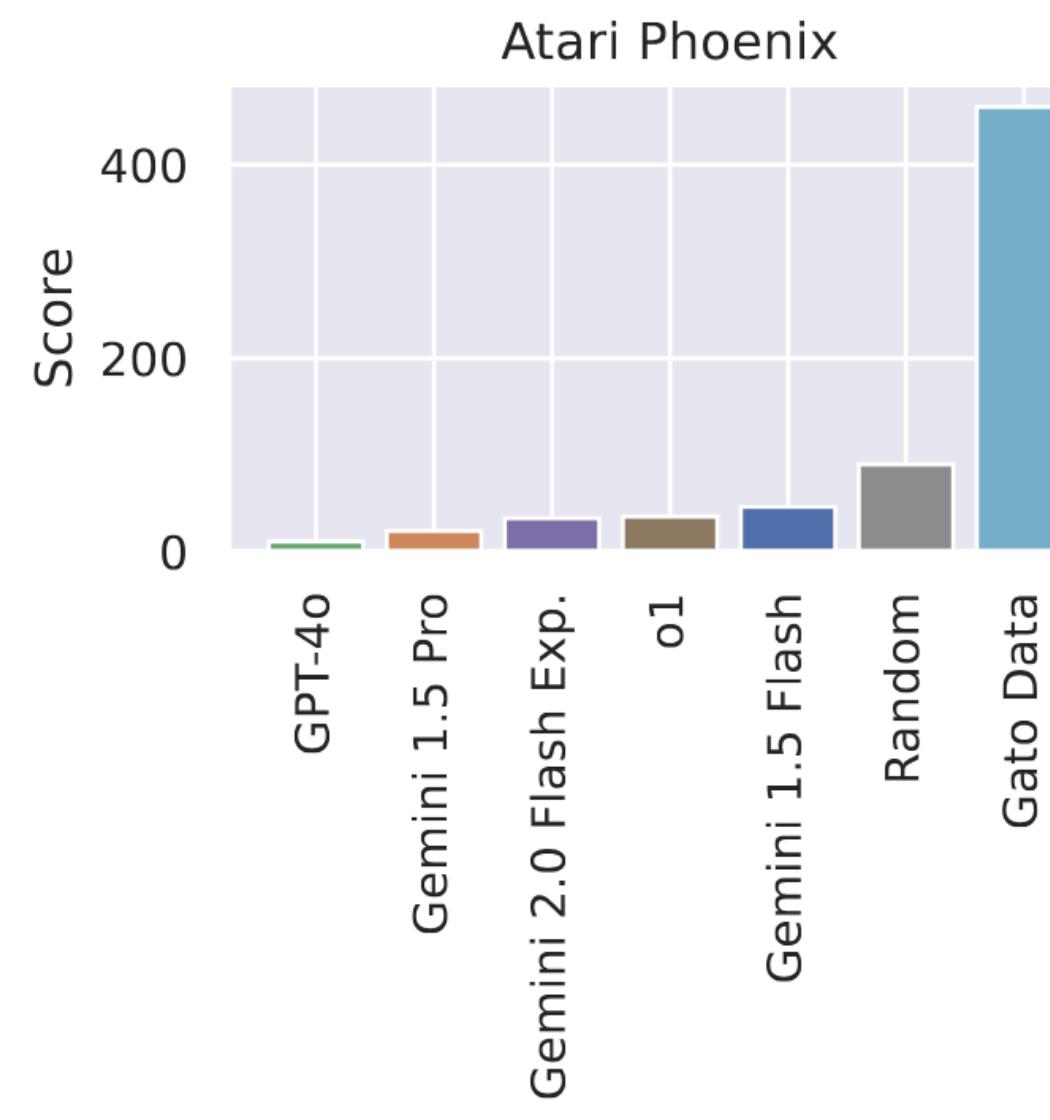


Figure 1: **Inconsistent knowledge in GPT-4.** GPT-4 correctly gives the name of Tom Cruise’s mother (left). Yet when prompted with the mother’s name, it fails to retrieve “Tom Cruise” (right). We hypothesize this ordering effect is due to the Reversal Curse. Models trained on “A is B” (e.g. “Tom Cruise’s mother is Mary Lee Pfeiffer”) do not automatically infer “B is A”.

Limitation 8: They cannot properly generalize

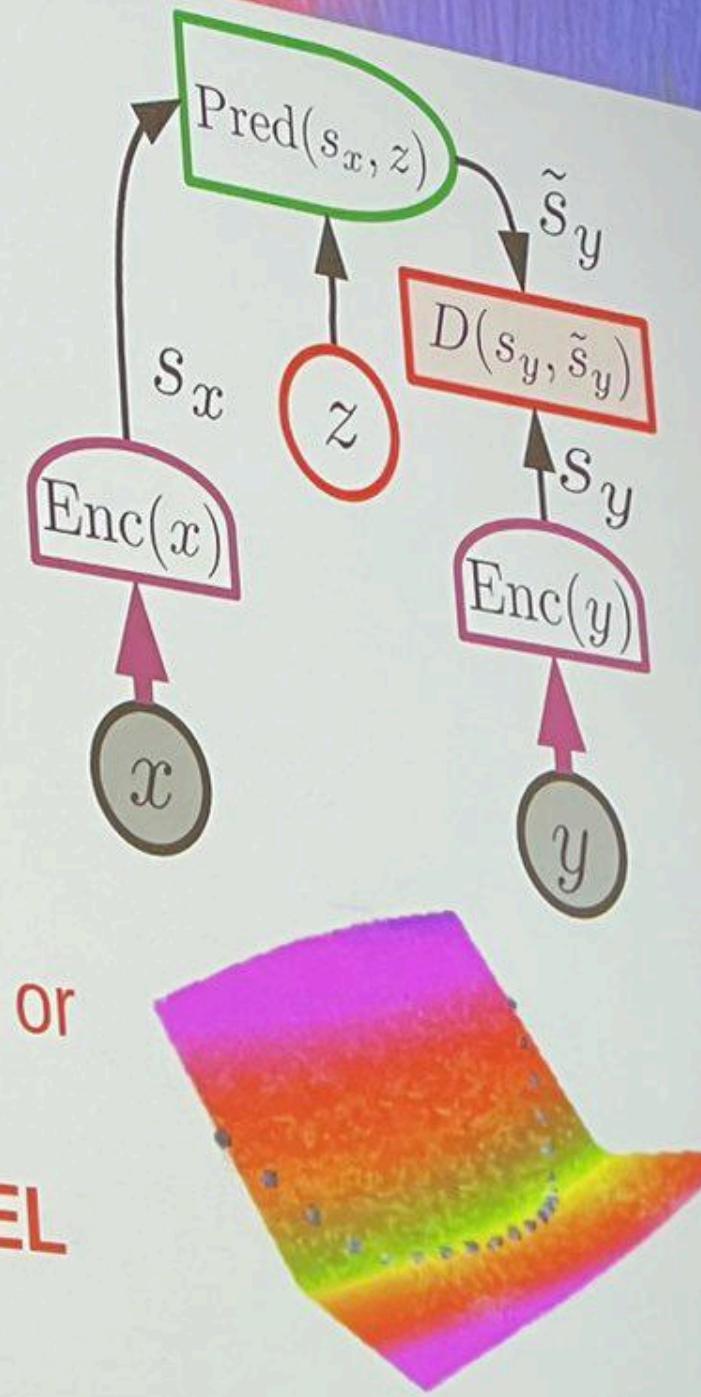
LMAct: A Benchmark for In-Context Imitation Learning with Long Multimodal Demonstrations



Future Directions

LeCun

- ▶ Abandon generative models in favor joint-embedding architectures
- ▶ Abandon probabilistic model in favor of energy-based models
- ▶ Abandon contrastive methods in favor of regularized methods
- ▶ Abandon Reinforcement Learning
 - ▶ In favor of model-predictive control
 - ▶ Use RL only when planning doesn't yield the predicted outcome, to adjust the world model or the critic.
- ▶ IF YOU ARE INTERESTED IN HUMAN-LEVEL AI, DON'T WORK ON LLMs



Memory — SSMs

Selective State Space Model with Hardware-aware State Expansion

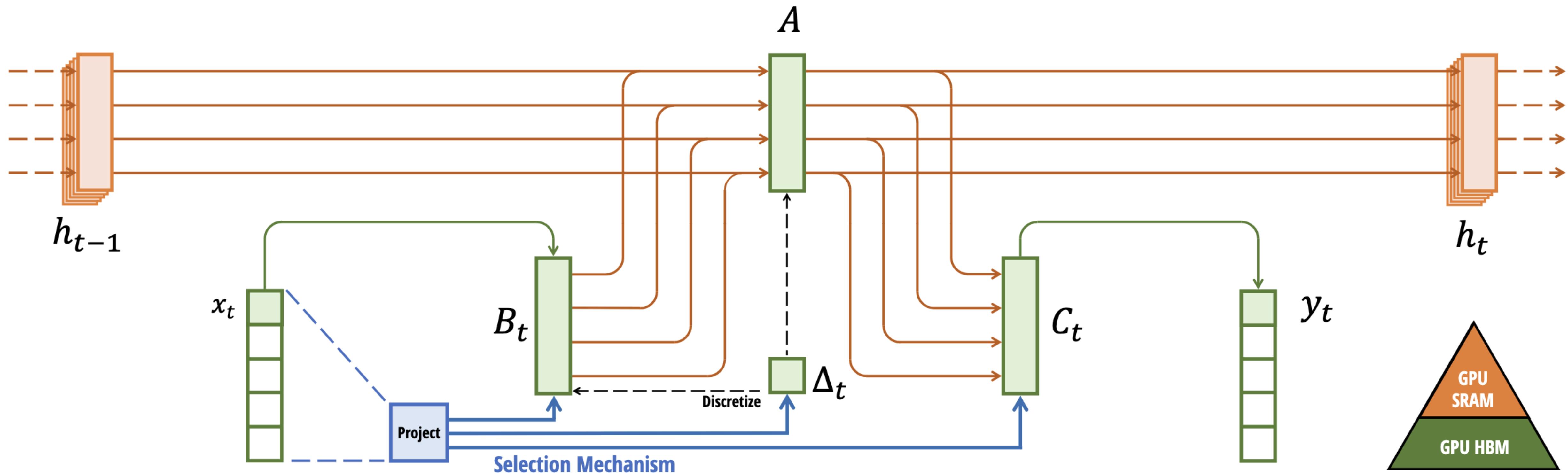


Figure 1: **(Overview.)** Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN , times batch size B and sequence length L) through clever alternate computation paths requiring time-invariance: the (Δ, A, B, C) parameters are constant across time. Our selection mechanism adds back input-dependent dynamics, which also requires a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU memory hierarchy.

Memory — LM2

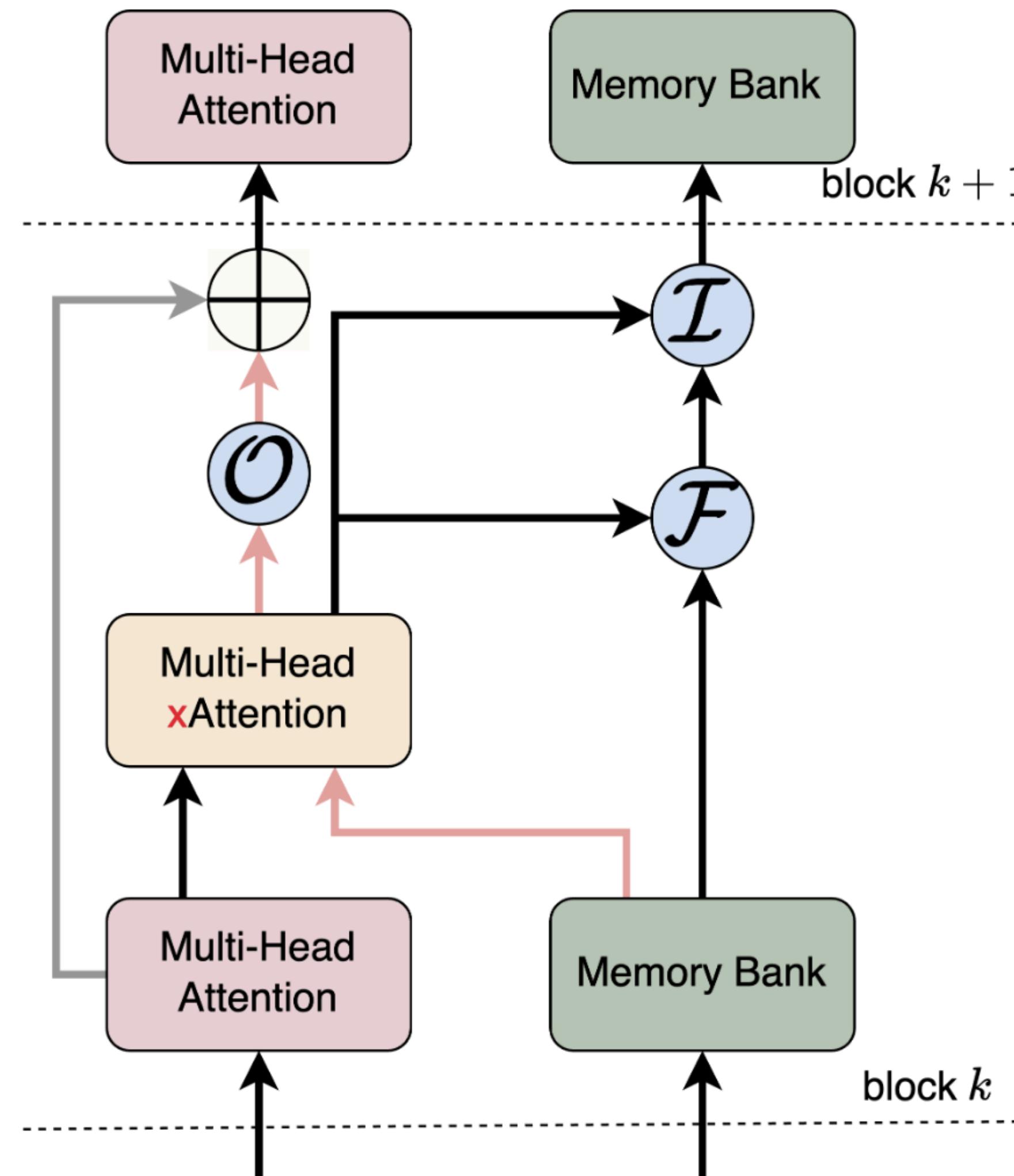
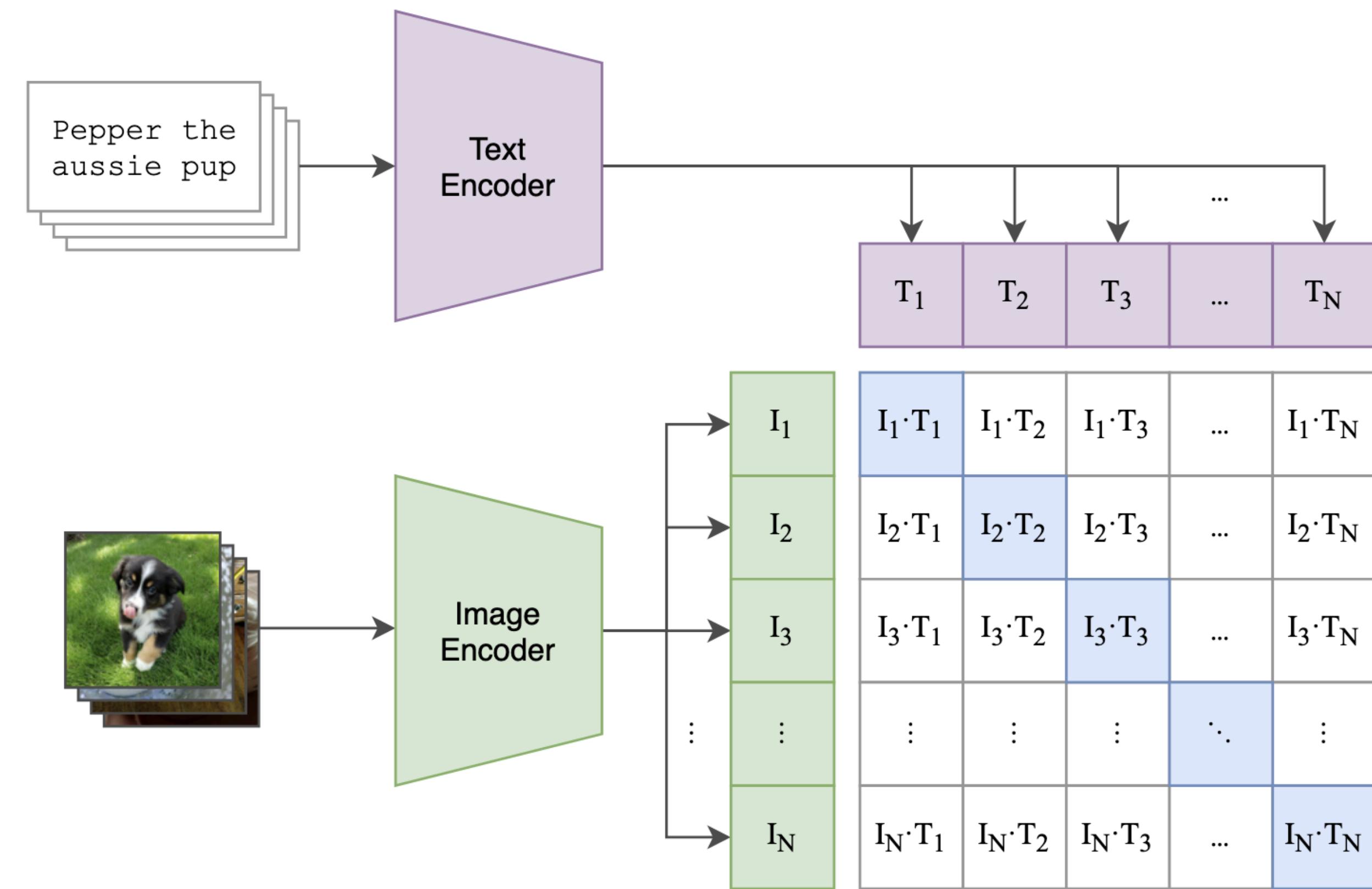


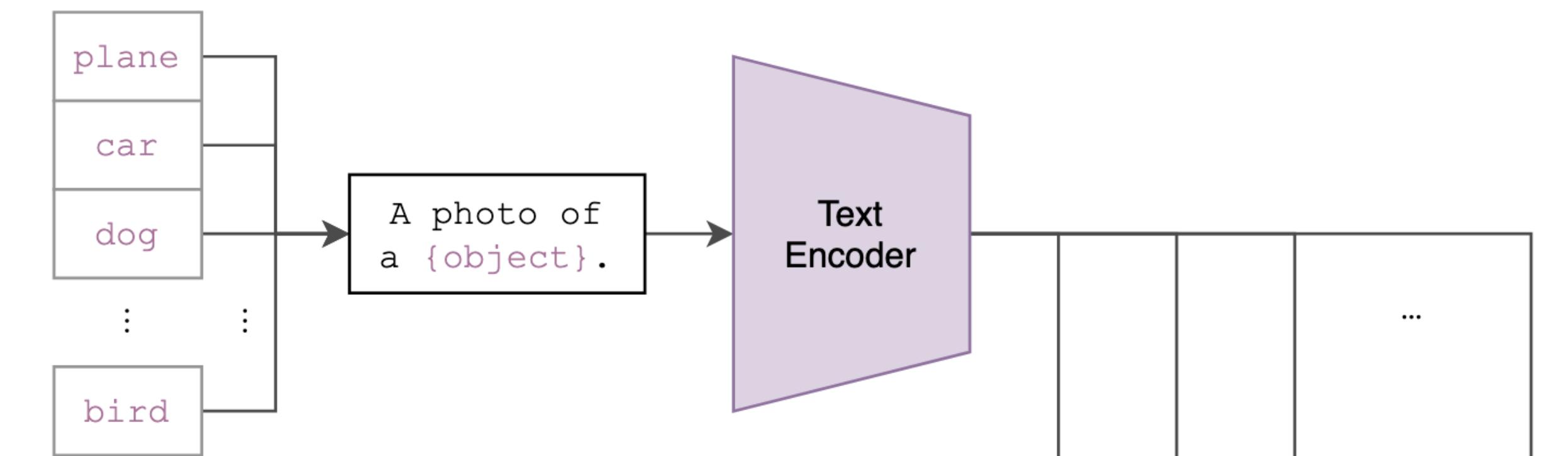
Figure 1 Illustration of LM2 overall architecture. It consists of a separate memory bank, which updates the main information flow through cross attention, and is updated using the input (\mathcal{I}), output (\mathcal{O}), and forget (\mathcal{F}) gates. For the information flow from one block to another, the gray curve shows the normal attention flow and the pink curve shows the extra memory flow.

Energy Based Models — CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

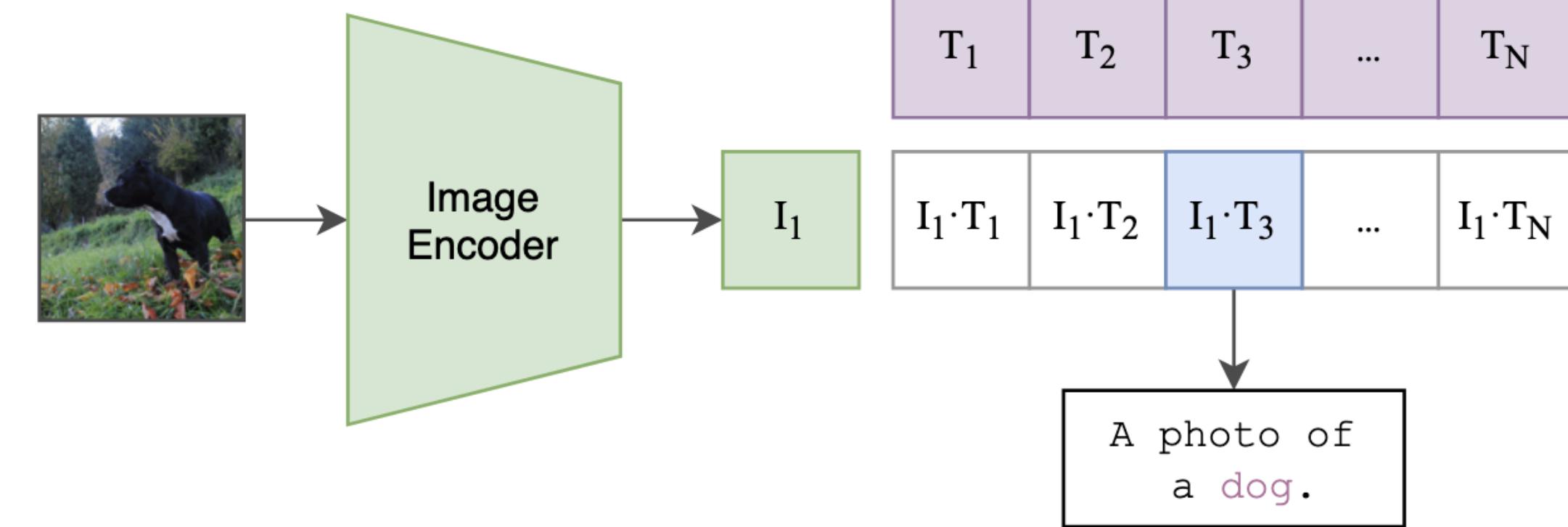
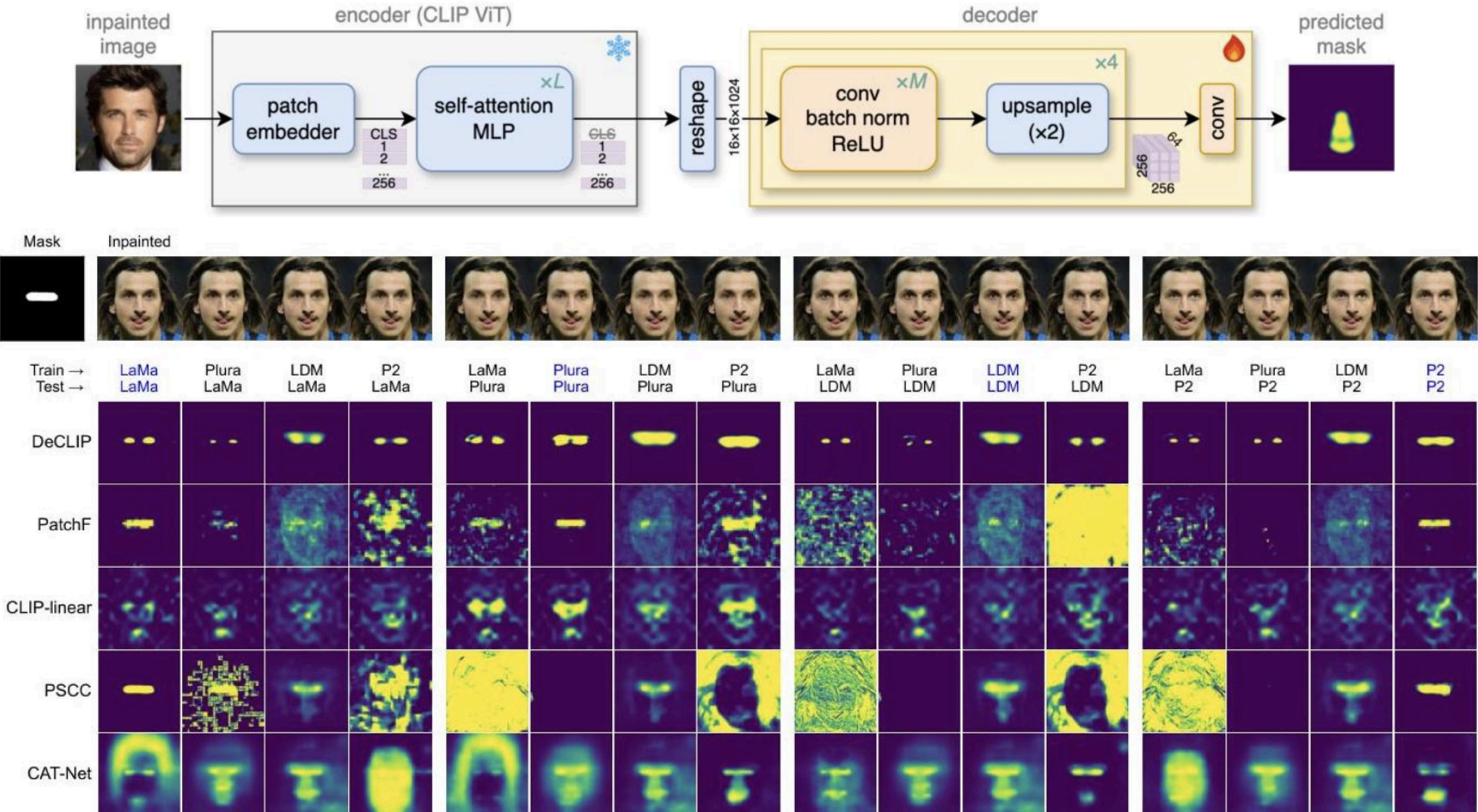


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

DeCLIP — Decoding CLIP Representations for Deepfake Localization



JEPA — Proper world model and proper planning

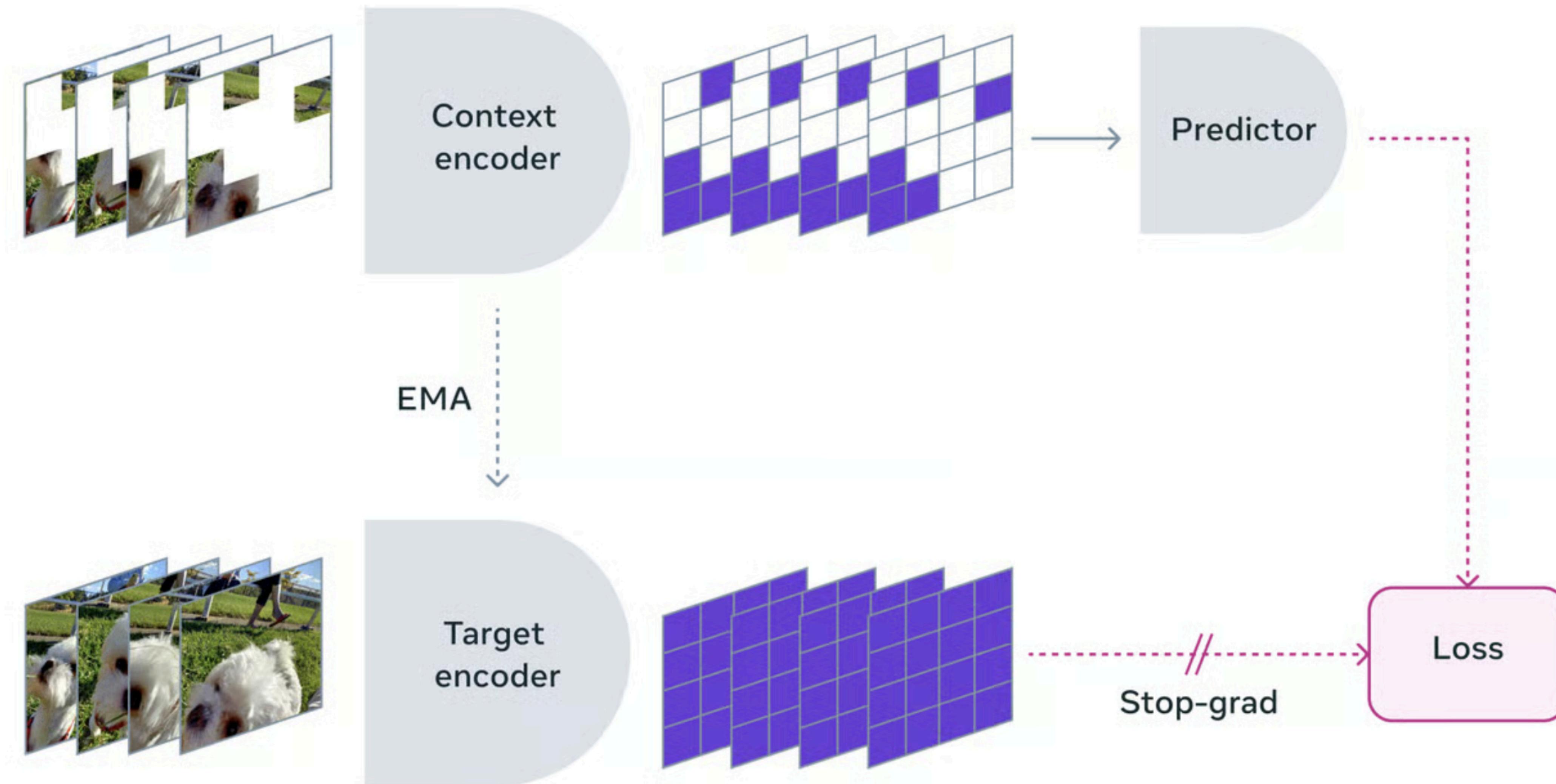


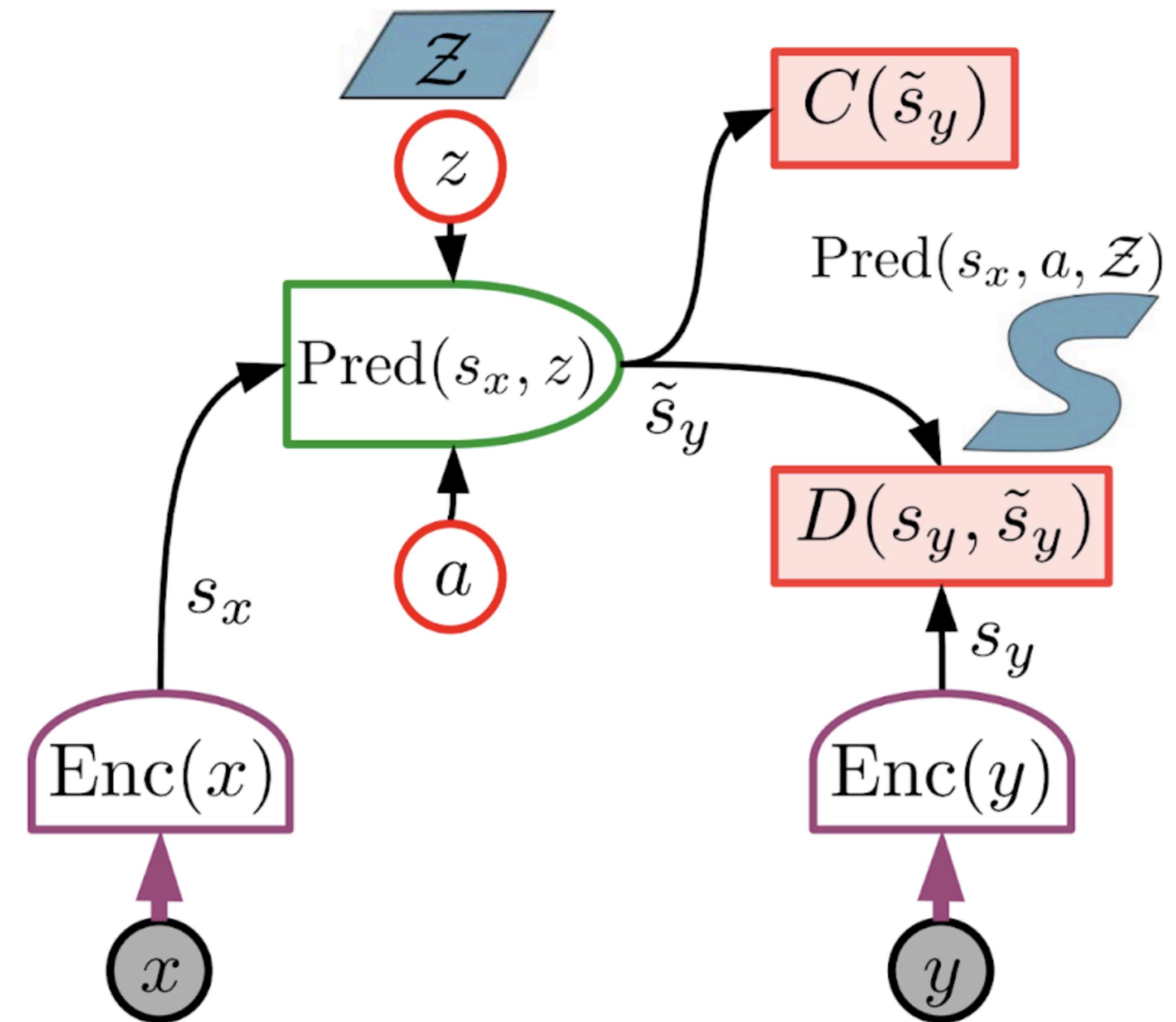
Image Credit: Meta AI blog post, [V-JEPA: The next step toward advanced machine intelligence](#)

JEPA — Proper world model and proper planning

Y. LeCun

Architecture for the world model: JEPA

- ▶ **JEPA: Joint Embedding Predictive Architecture.**
- ▶ x : observed past and present
- ▶ y : future
- ▶ a : action
- ▶ z : latent variable (unknown)
- ▶ $D()$: prediction cost
- ▶ $C()$: surrogate cost
- ▶ JEPA predicts a representation of the future S_y from a representation of the past and present S_x



Thank you