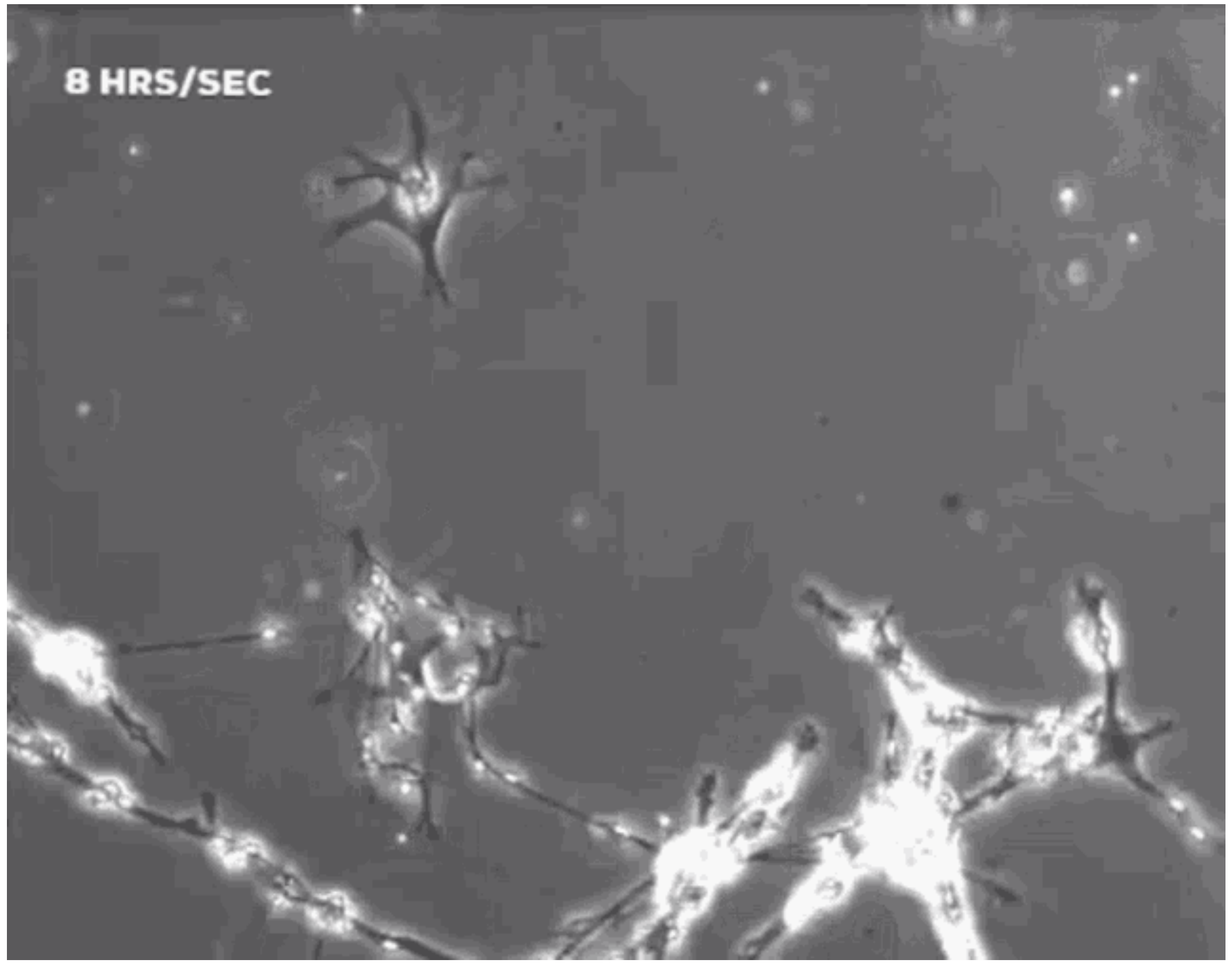
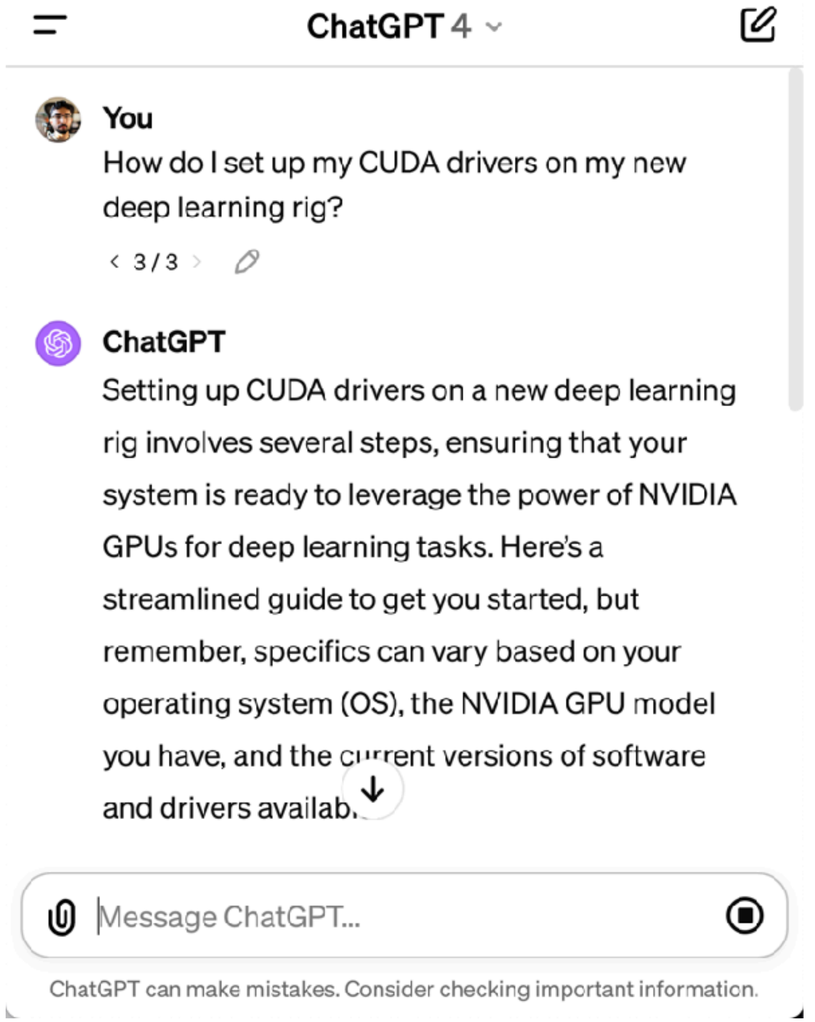
Thinking Machines

Abstract thought as internal world model exploration

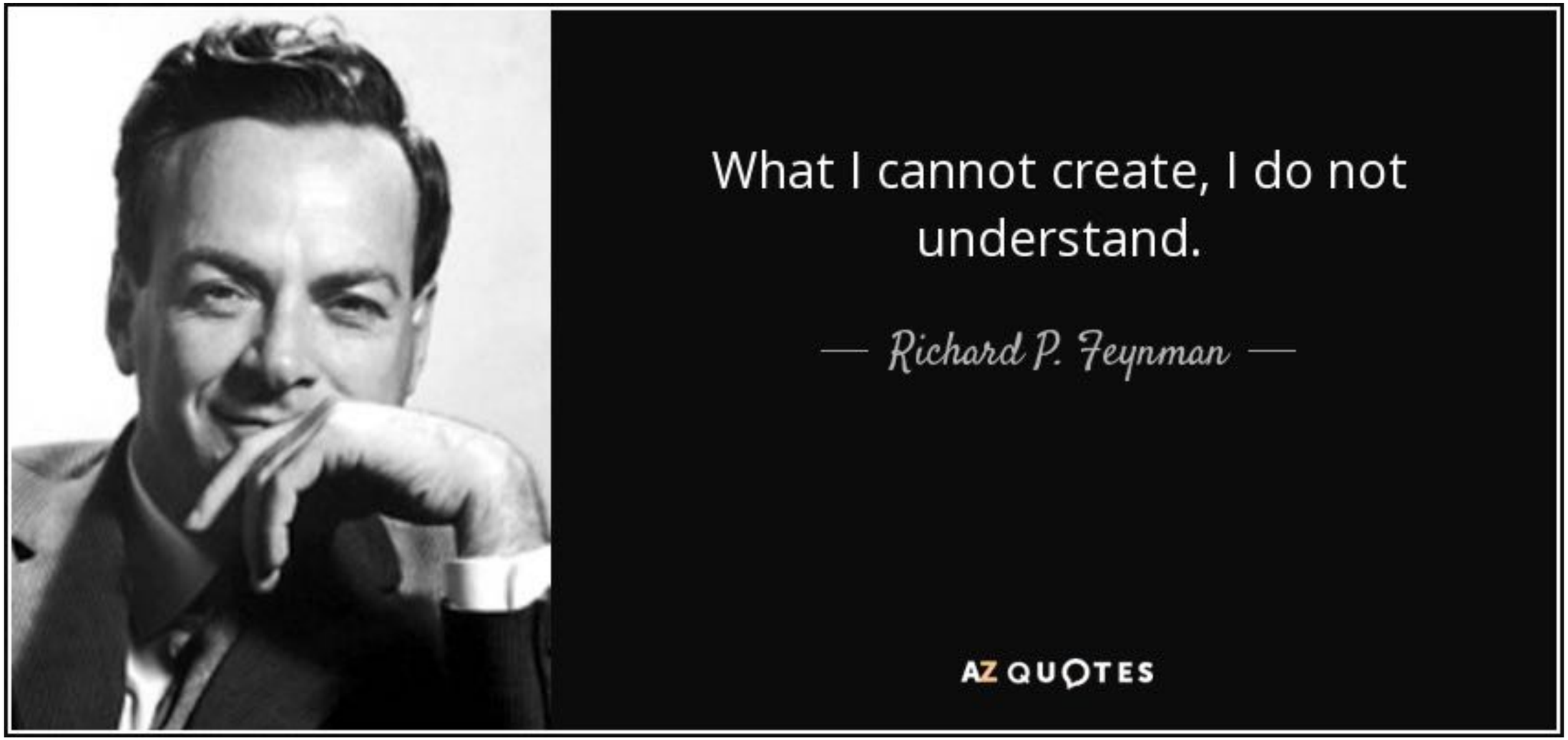
Aman Bhargava ⊂ Thomson Lab ⊂ Caltech, 2024
aman-bhargava.com/

Intelligence is a fascinating phenomenon.





Building AI can help us understand intelligence.



Building AI can help us understand intelligence.

LLMs exhibit aspects of intelligence.

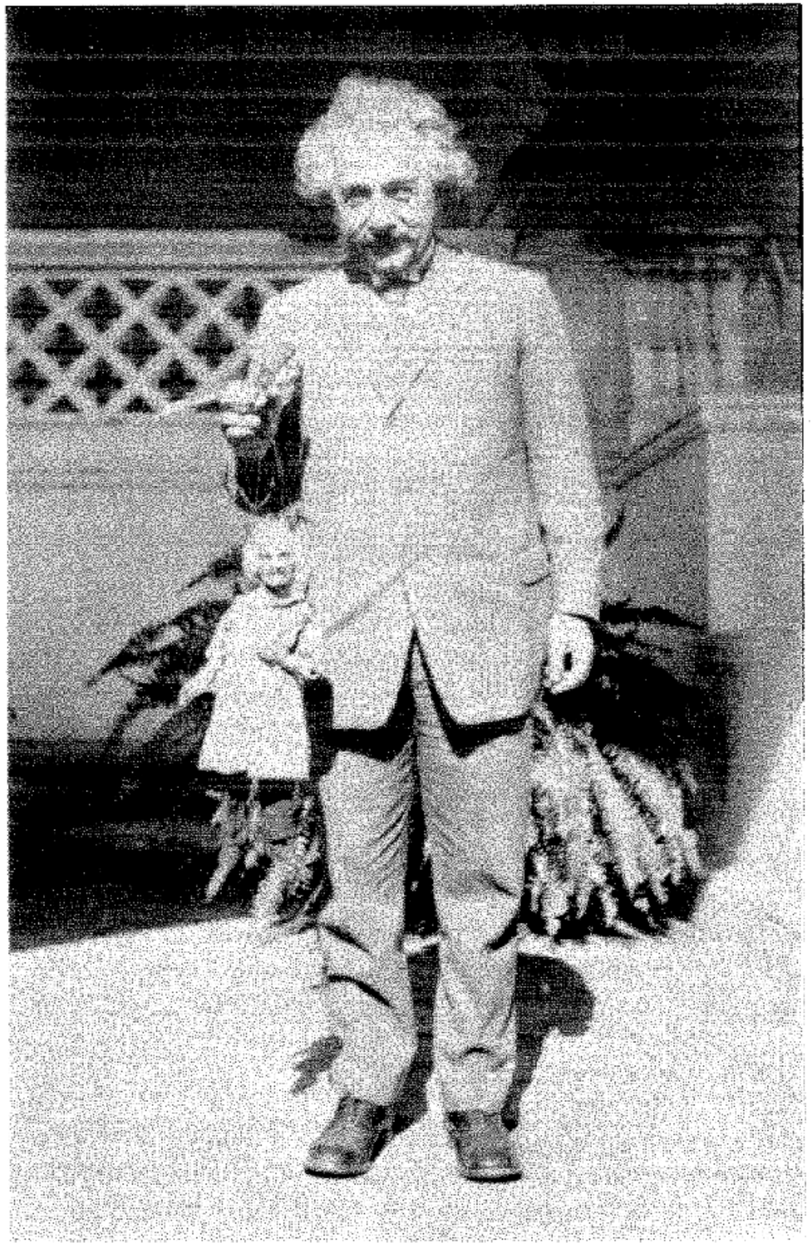
Zero-shot learning miracle

- **Knowledge Retrieval:** *"The Titanic sank in the year [MASK]."* (Answer: "1912")
- **Reasoning:** *"A is taller than B. B is taller than C. Is A taller than C? **Answer: [MASK]**"* (Answer: "Yes")
- **Sentiment Analysis:** *"I am sad today. **The sentiment of the previous sentence was [MASK]**"* (Answer: "Negative")

From "*What's the Magic Word?* A Control Theory of LLM Prompting?" (Bhargava, Witkowski, Thomson, 2023) — https://arxiv.org/abs/2304.15004

“The human mind has first to construct forms, independently, **before we can find them in things.**”

— **Albert Einstein**



Two Einsteins in front of the brand-new Athenæum in 1931 — Albert himself and a puppet from the play called *Mr. Noah.* Albert said the puppet was "good but not fat enough."

Thinking Machines

Motivation

What is **thinking**?

How can we **formalize** thinking (e.g., mathematically/informationally)?

How can we **build thinking machines**?

Thinking Machines

Motivation

What is **thinking**?

How can we **formalize** thinking (e.g., mathematically/informationally)?

How can we **build thinking machines**?

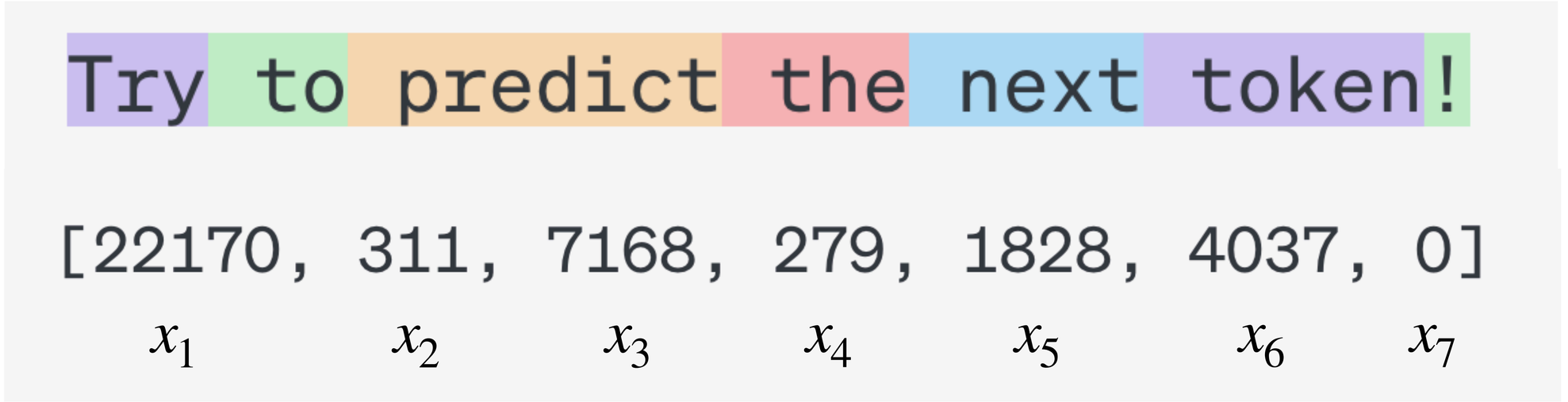
Why can’t ChatGPT automate my entire PhD yet?

Thinking Machines

Agenda

1. LLMs **predict the next word** (token).
2. **World models** emerge from predictive coding.
3. **BIG IDEA:** Thinking = **Internal World Model Exploration!**
4. **Register token** architecture for “Thinking LLM”.
5. **Beyond LLMs** — the end-game for CNS.

LLMs predict the next token.



$$P_\theta(x_{t+1} \mid x_1, \ldots, x_t)$$

LLMs predict the next token.

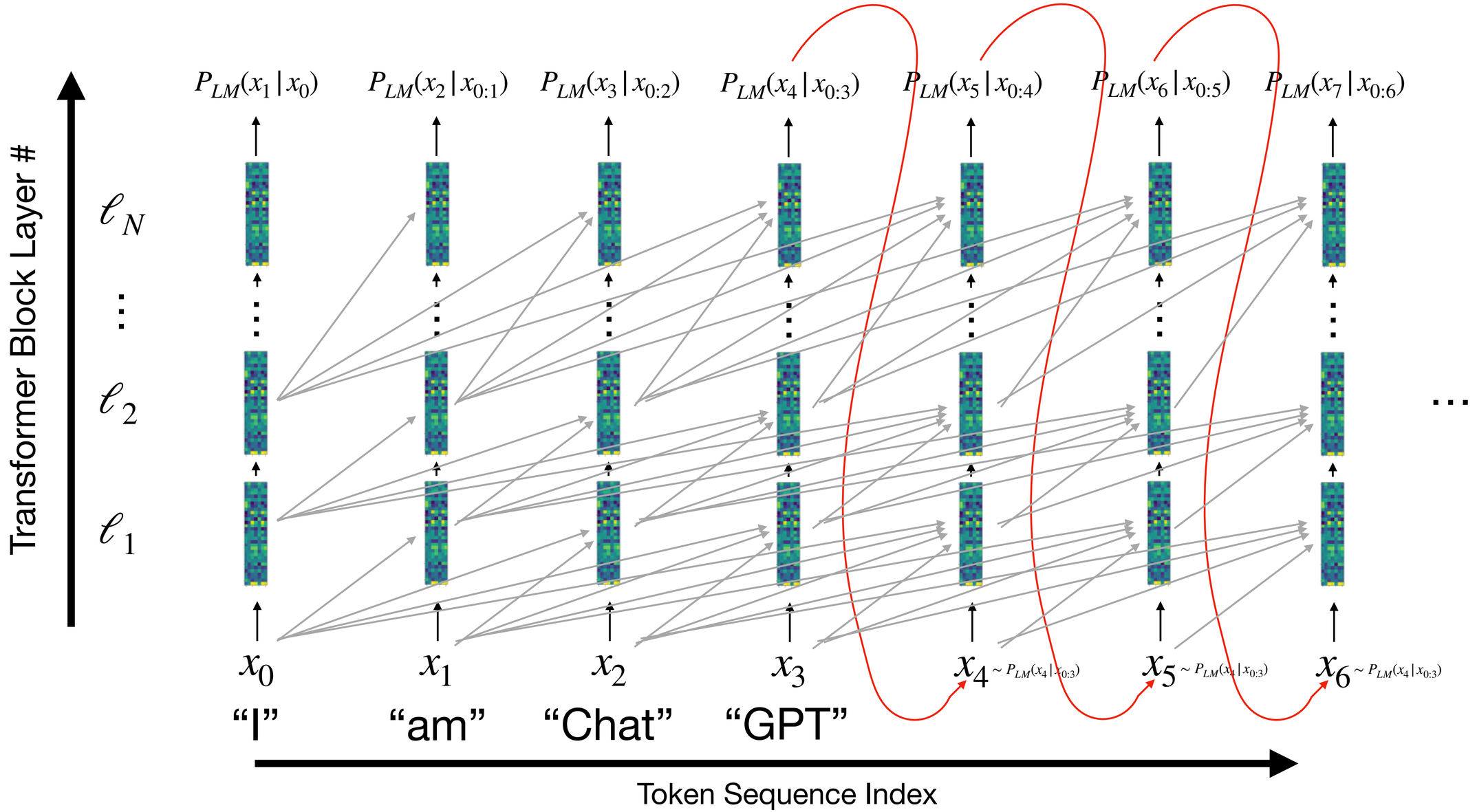
Try to predict the next token!

[22170, 311, 7168, 279, 1828, 4037, 0]

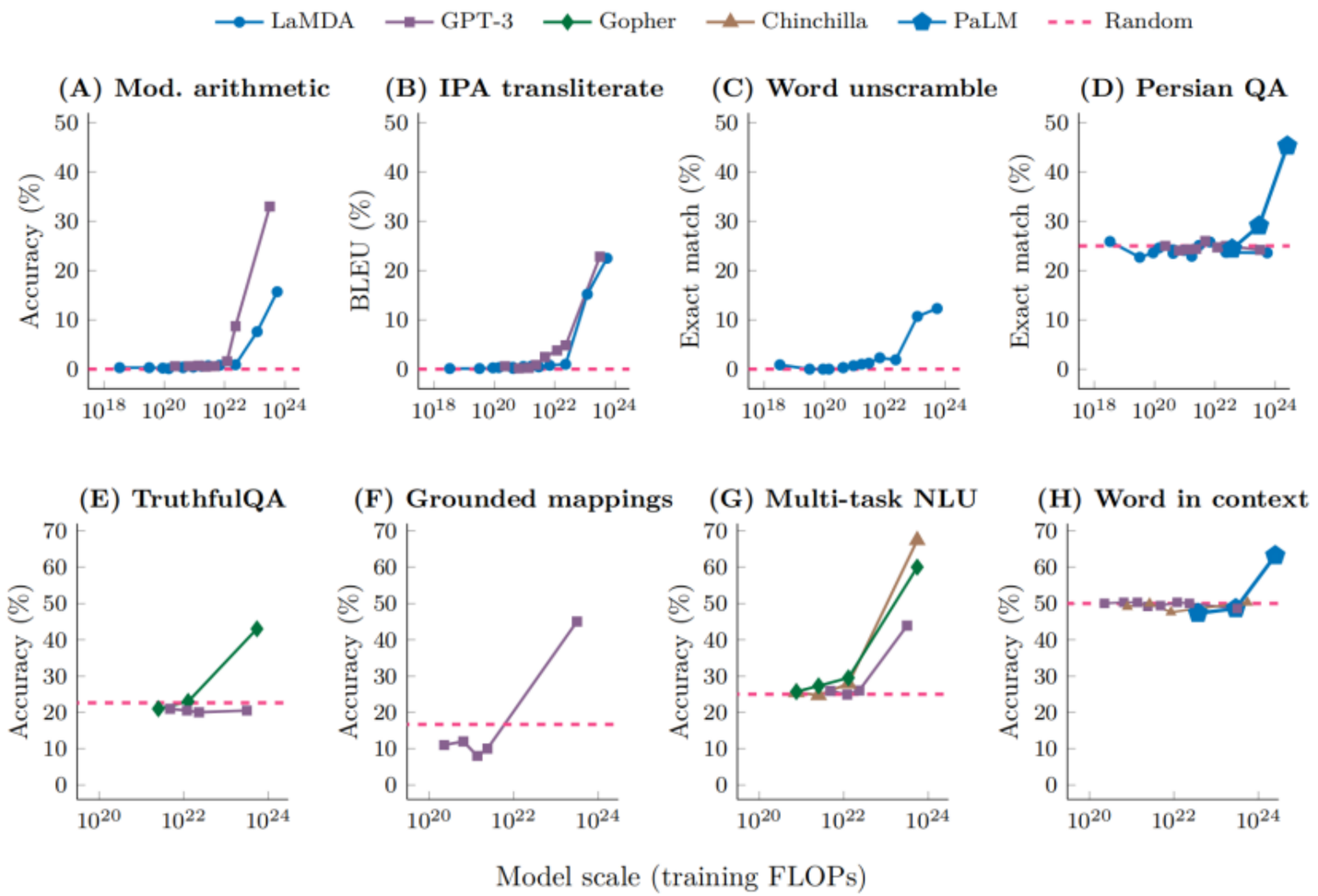
x_1 x_2 x_3 x_4 x_5 x_6 x_7

$$\theta = \arg\max_{\theta} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[\underbrace{\sum_{i=1}^{N} \log P_\theta(x_i \mid x_1 \ldots x_{i-1})}_{\log P_\theta(x_1 \ldots x_N)}\Big]$$

LLMs predict the next token.



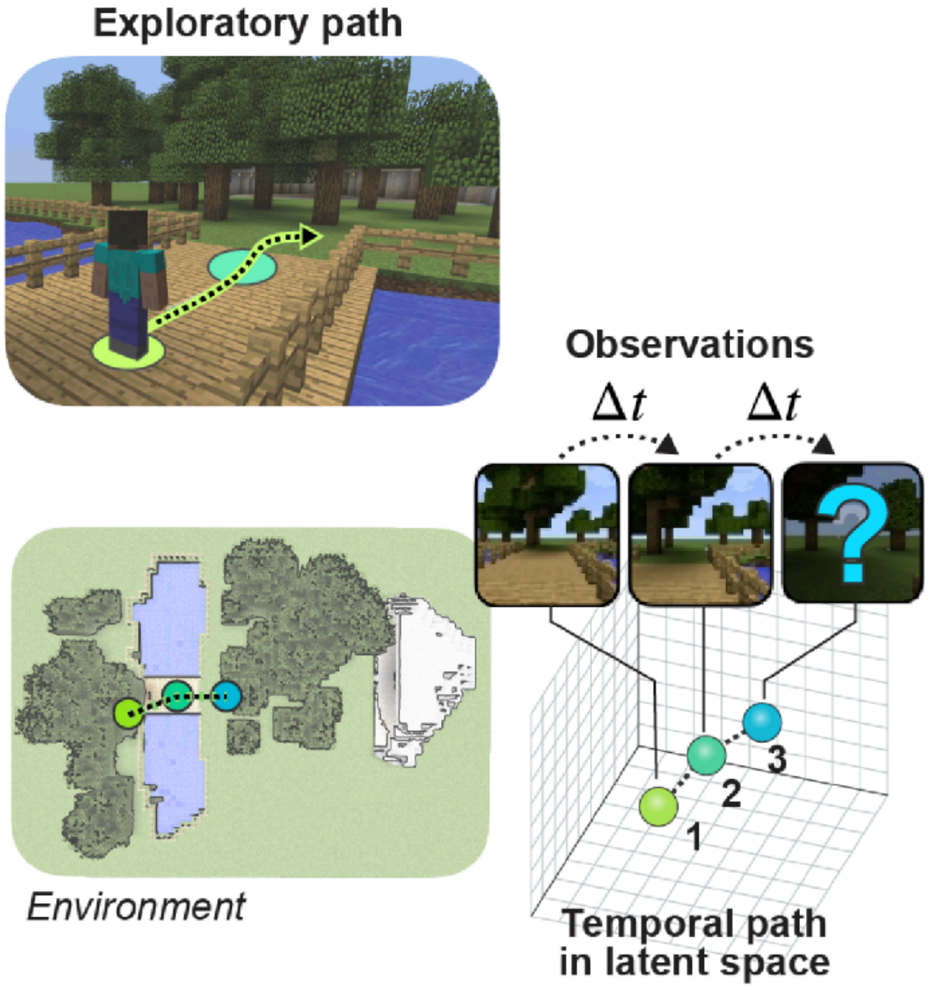
Bigger LLM → Better Prediction → Greater Capability



From “Are Emergent Abilities of Large Language Models a Mirage?” (Schaeffer et al) — https://arxiv.org/abs/2304.15004

World Models emerge from Predictive Coding.

James Gornet (Thomson Lab, CNS)



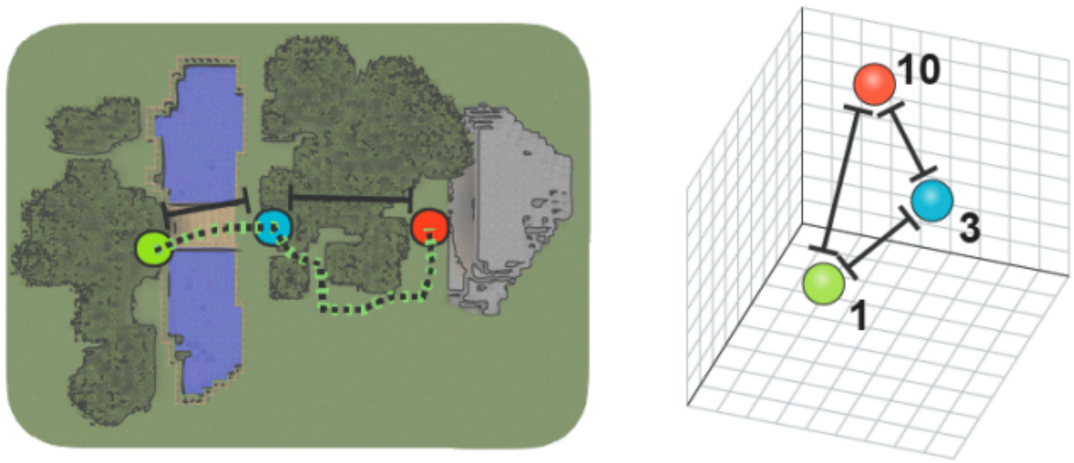
James!

From “Automated mapping of virtual environments with visual predictive coding” (Gornet and Thomson) — arxiv:2308.10913

World Models emerge from Predictive Coding.

James Gornet (Thomson Lab, CNS)

$$P_\theta(x_{t+1} \mid x_1, \ldots, x_t)$$



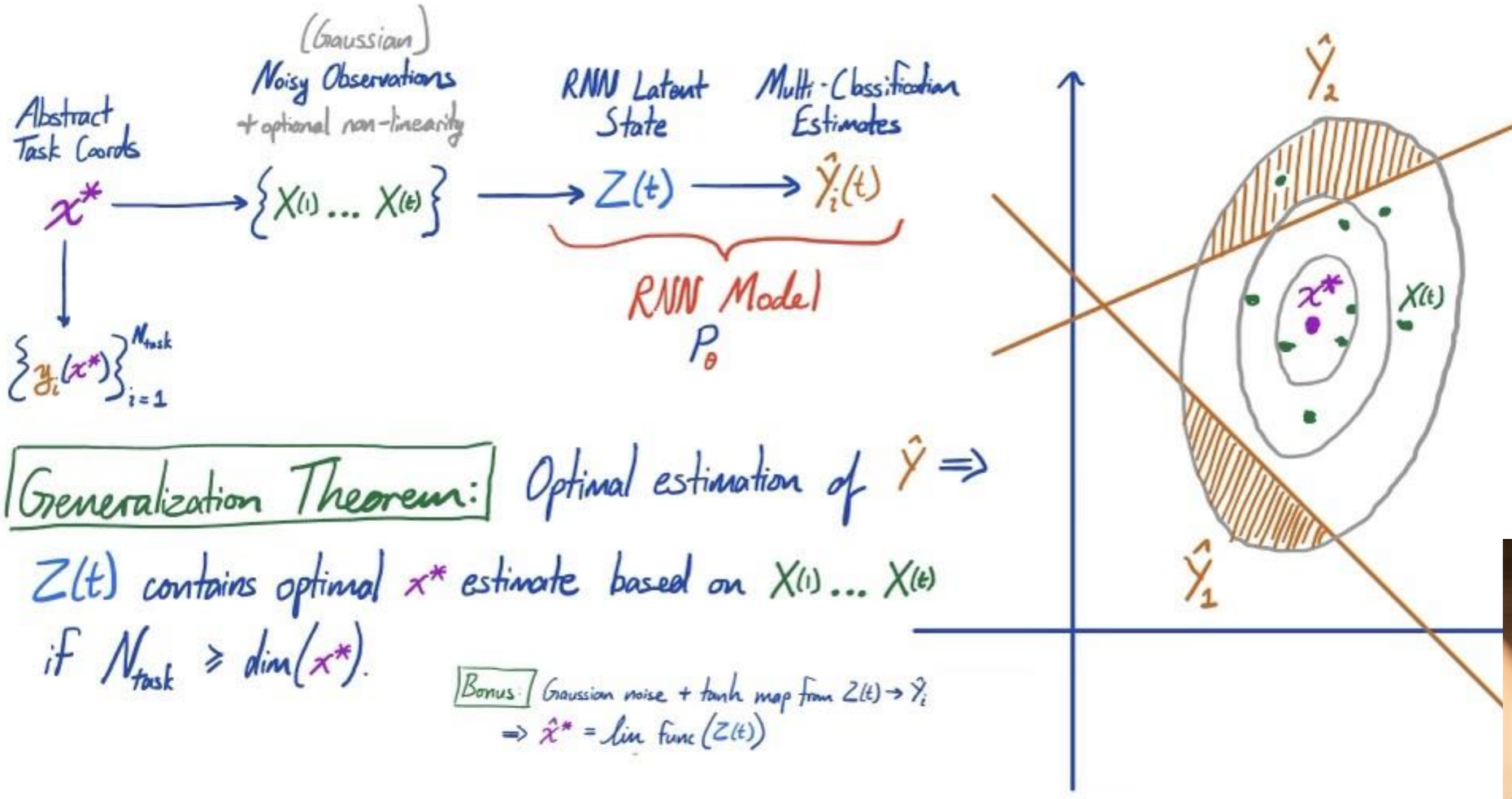
Distance in latent space maps environment

James!

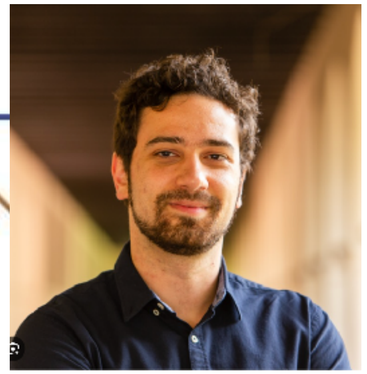
From “Automated mapping of virtual environments with visual predictive coding” (Gornet and Thomson) — arxiv:2308.10913

World Models emerge from Evidence Aggregation.

With Pantelis Vafidis (Rangel Lab, CNS)



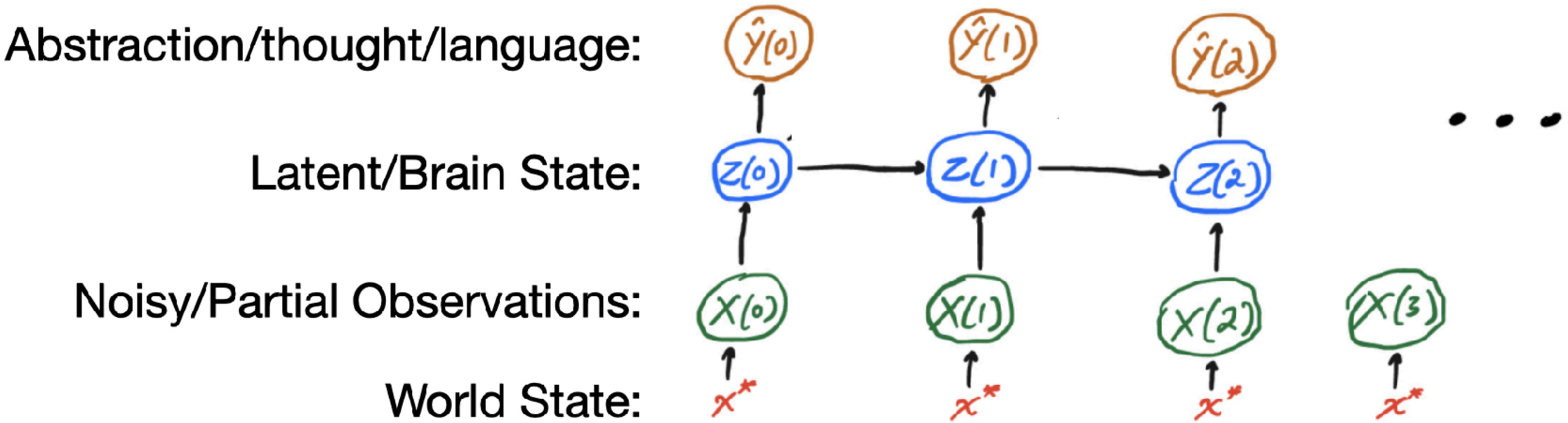
Pantelis!



From “Disentangling Representations in RNNs through Multi-task Learning” (Vafidis, Bhargava, & Rangel)

World Models emerge from Predictive Coding.

- Latent $\mathbf{Z}(t)$ has provably high mutual information with world $I\left(\mathbf{Z}(t); \mathbf{x}^*(t)\right)$ for a variety of prediction tasks.

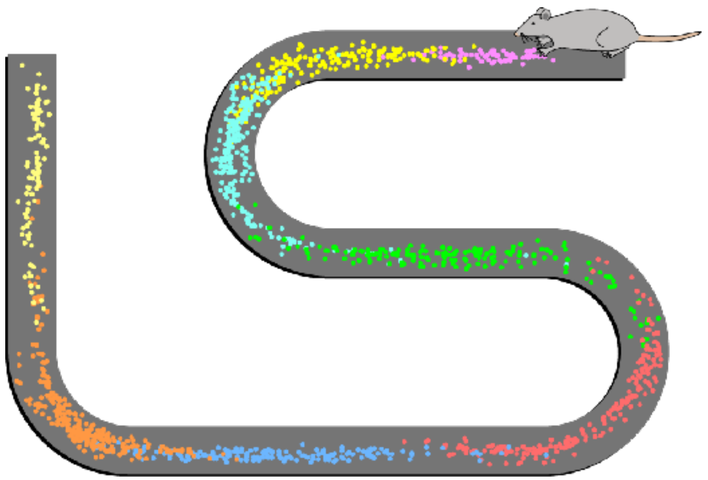


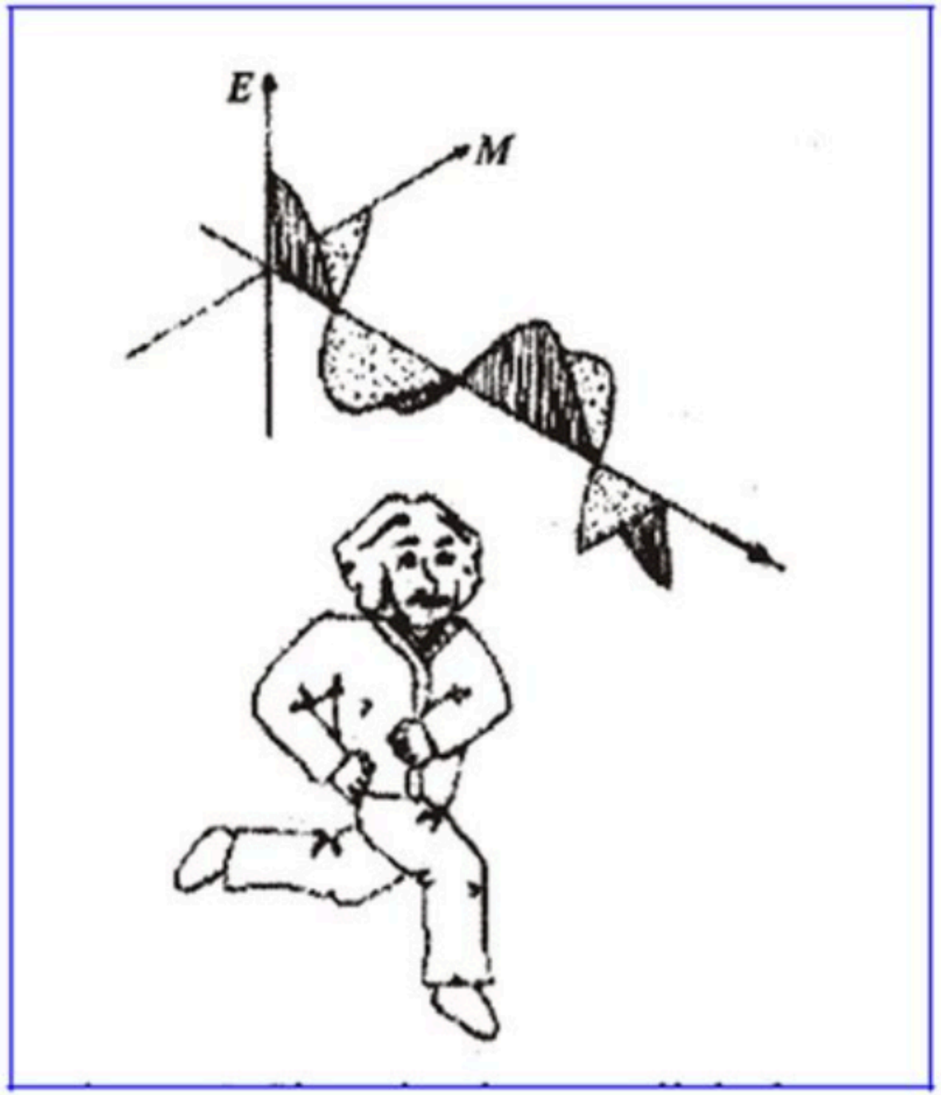
From “Disentangling Representations in RNNs through Multi-task Learning” (Vafidis, Bhargava, Rangel) — under review

Thinking is internal world model exploration.

Leveraging *implicit understanding*.

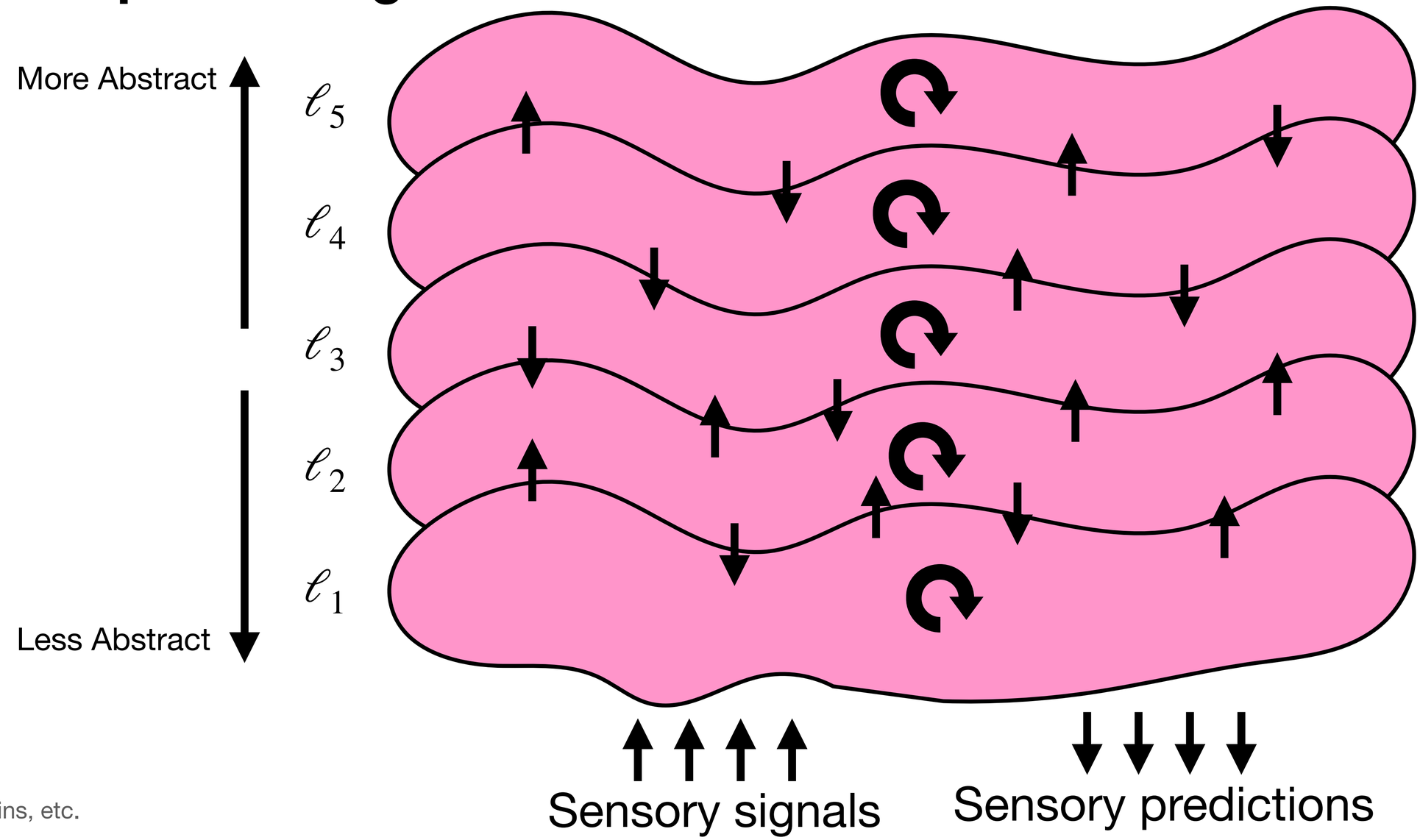
- **Neuro**: We observe rodents “mentally exploring” a maze in place cell activations.
- **Phil of Mind**: Thought experiments, hypotheticals, imagined movement/ environments.





Thought as internal world model exploration

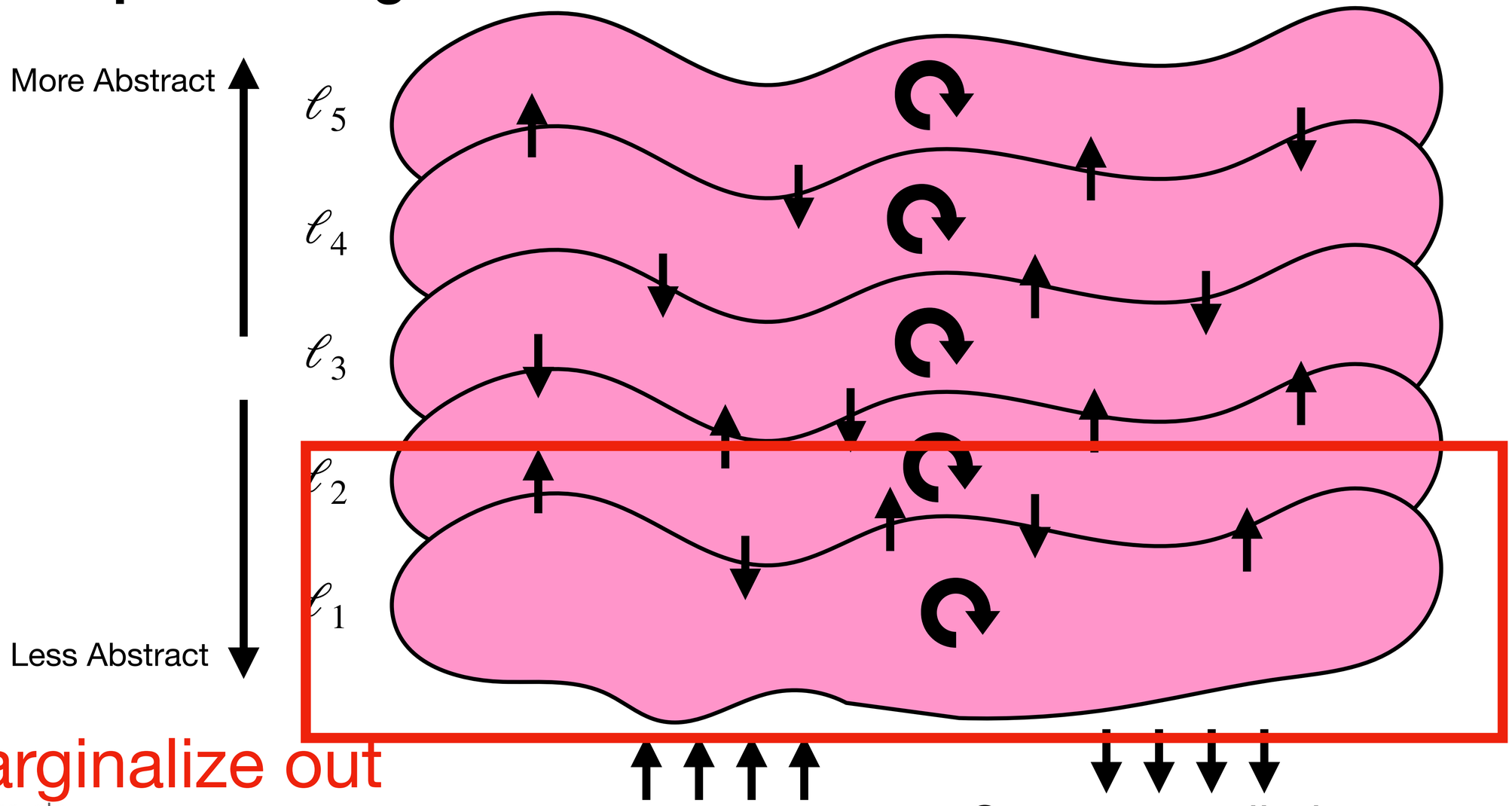
Cortical processing cartoon sketch



FEP, 1000 Brains, etc.

Thought as internal world model exploration

Cortical processing cartoon sketch

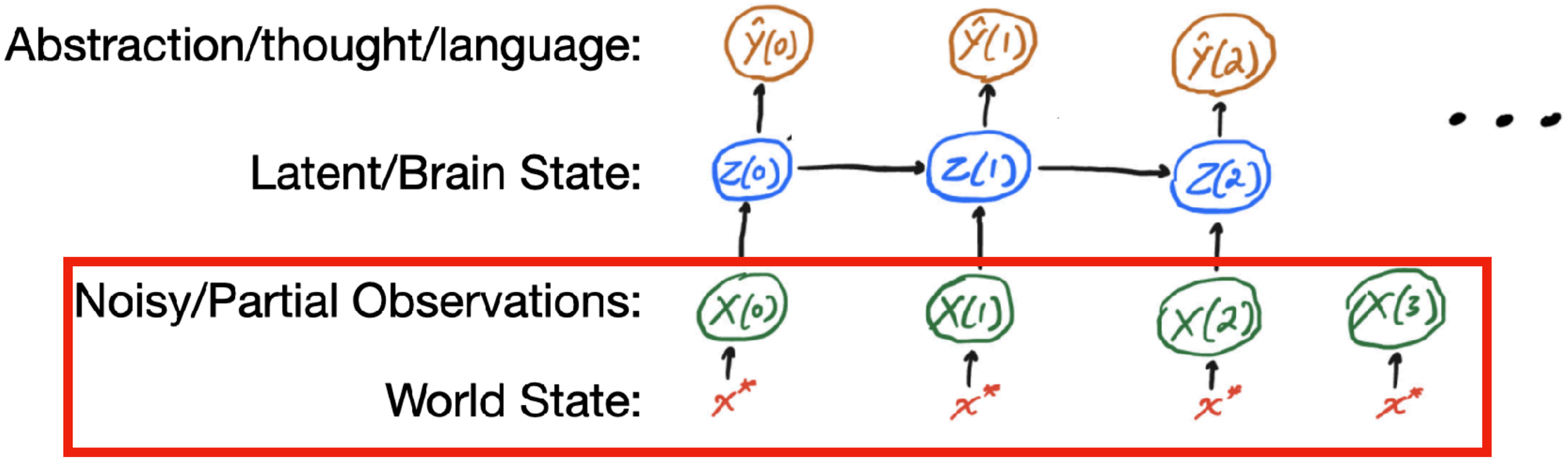


Marginalize out sensory input!

FEP, 1000 Brains, etc.

World Models emerge from Predictive Coding.

- Latent $\mathbf{Z}(t)$ has provably high mutual information with world $I\left(\mathbf{Z}(t); \mathbf{x}^*(t)\right)$ for a variety of prediction tasks.

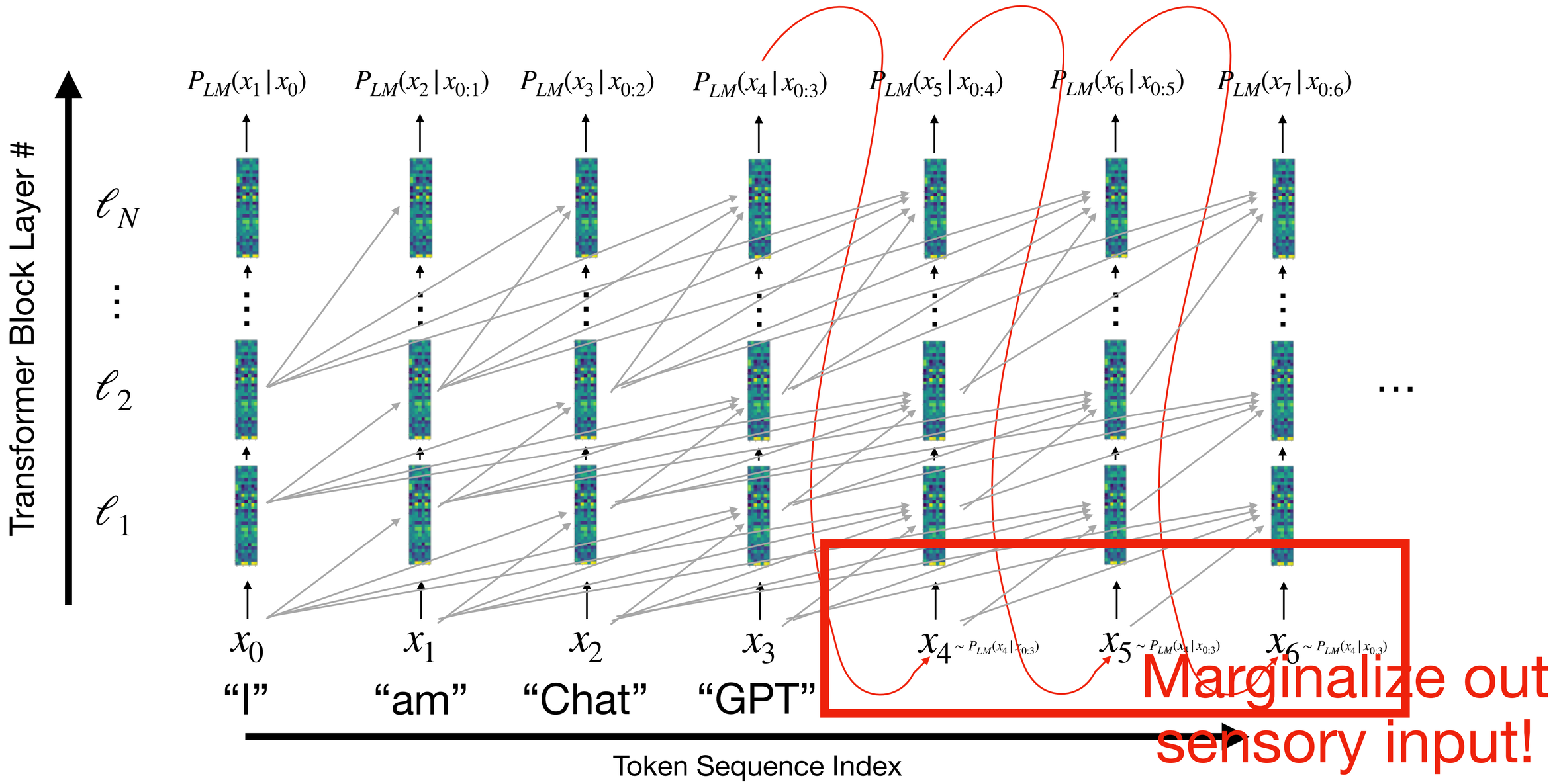


Marginalize out sensory input!

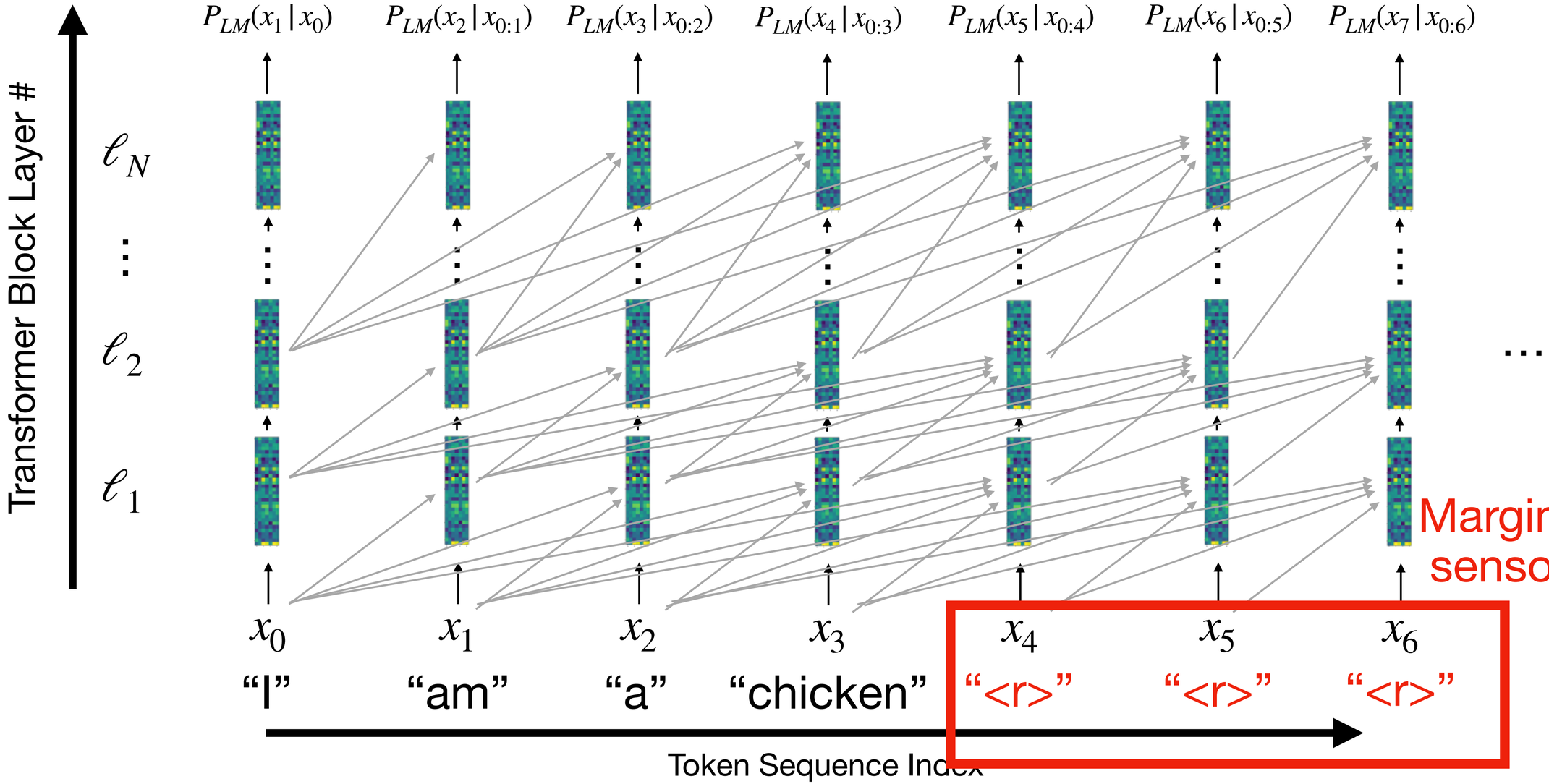
From “Disentangling Representations in RNNs through Multi-task Learning” (Vafidis, Bhargava, Rangel) — under review

Thinking LLMs via Register Tokens

Register Tokens enable LLM world model exploration.



Register Tokens enable LLM world model exploration.



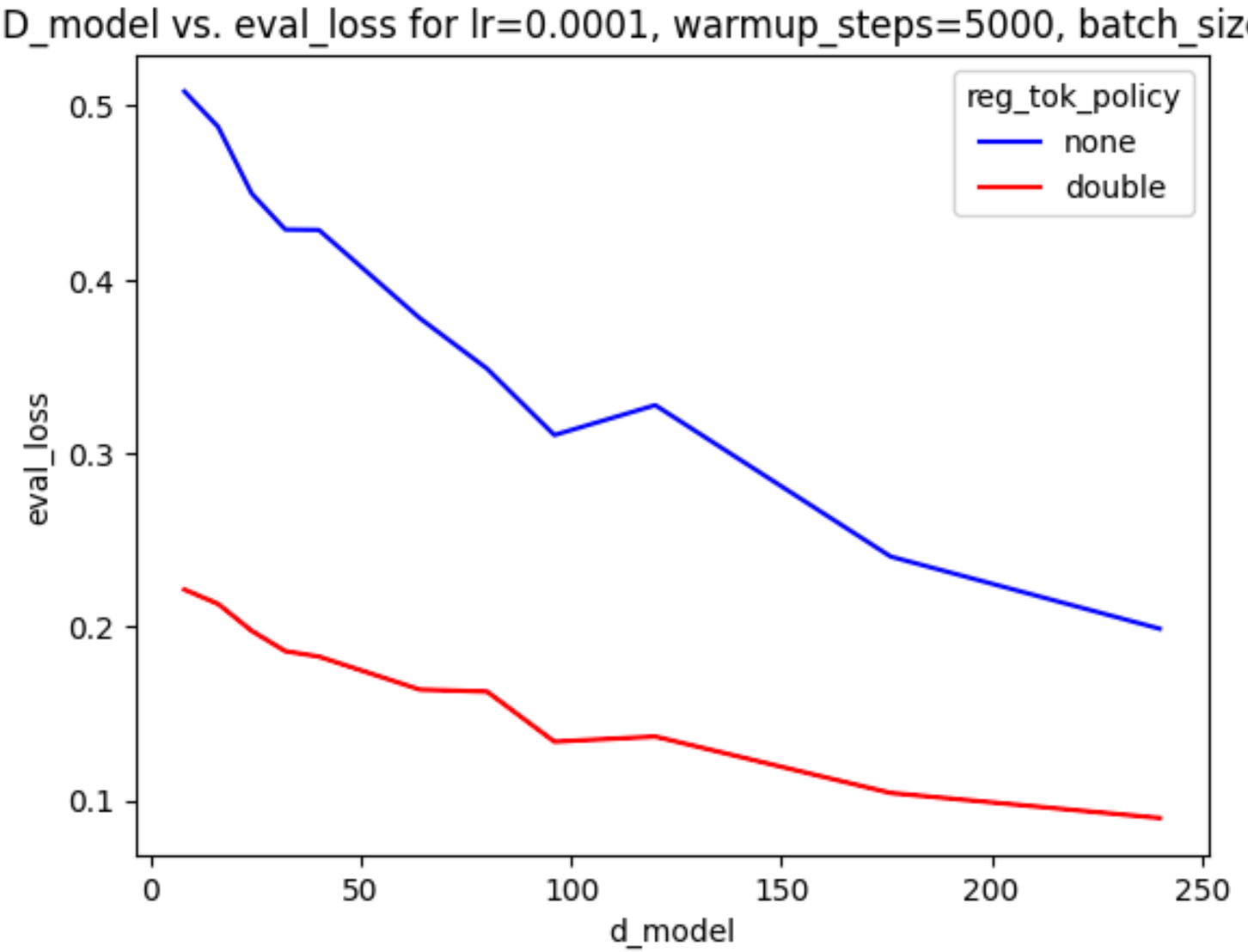
Register tokens improve arithmetic performance.

```
Dataset = [
    {
        "Question": "8 * 98 = <r><r>...<r>",
        "Answer": "784"
    },
    {
        "Question": "32 * 98891 = <r><r>...<r>",
        "Answer": "3164512"
    },
]
```

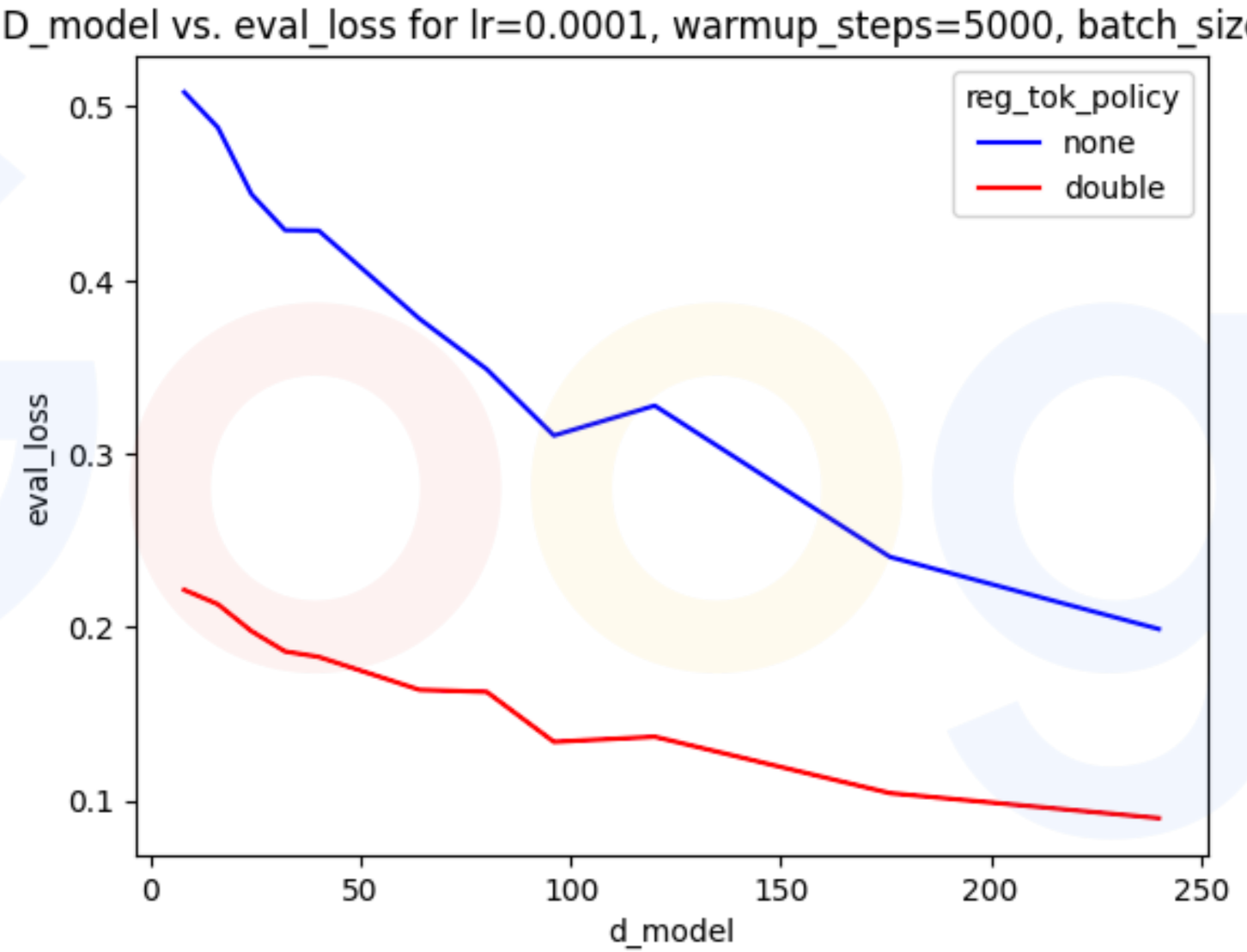
Register tokens improve arithmetic performance.

```
Dataset = [
    {
        "Question": "8 * 98 = ",
        "Answer": "784"
    },
    {
        "Question": "32 * 98891 = ",
        "Answer": "3164512"
    },
]
```

Register tokens improve arithmetic performance.



Register tokens improve arithmetic performance.



Register tokens improve arithmetic performance.

Register tokens improve arithmetic performance.

Published as a conference paper at ICLR 2024

THINK BEFORE YOU SPEAK: TRAINING LANGUAGE MODELS WITH PAUSE TOKENS

Sachin Goyal[*†]
Machine Learning Department
Carnegie Mellon University
sachingo@andrew.cmu.edu

Ziwei Ji
Google Research, NY
ziweiji@google.com

Ankit Singh Rawat
Google Research, NY
ankitsrawat@google.com

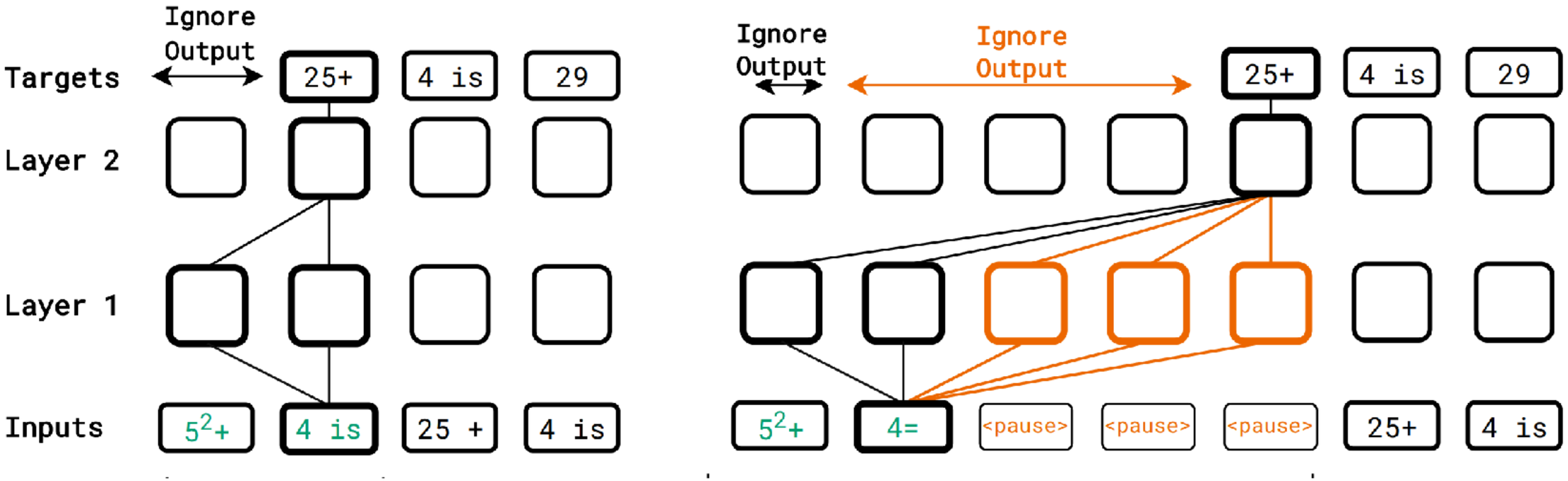
Aditya Krishna Menon
Google Research, NY
adityakmenon@google.com

Sanjiv Kumar
Google Research, NY
sanjivk@google.com

Vaishnavh Nagarajan[†]
Google Research, NY
vaishnavh@google.com

"Think before you speak: Training Large Language Models With Pause Tokens". Goyal *et al*, ICLR 2024, arXiv:2310.02226

Register tokens improve arithmetic performance.



(a) Standard inference and finetuning

(b) Pause-inference and finetuning

“Think before you speak: Training Large Language Models With Pause Tokens”. Goyal *et al*, ICLR 2024, arXiv:2310.02226

Register tokens improve arithmetic performance.

(a) Standard inference and fi nce and finetuning

"Think before you speak: Training Large Language Models With Pause Tokens". Goyal *et al*, ICLR 2024, arXiv:2310.02226

Register tokens improve arithmetic performance.

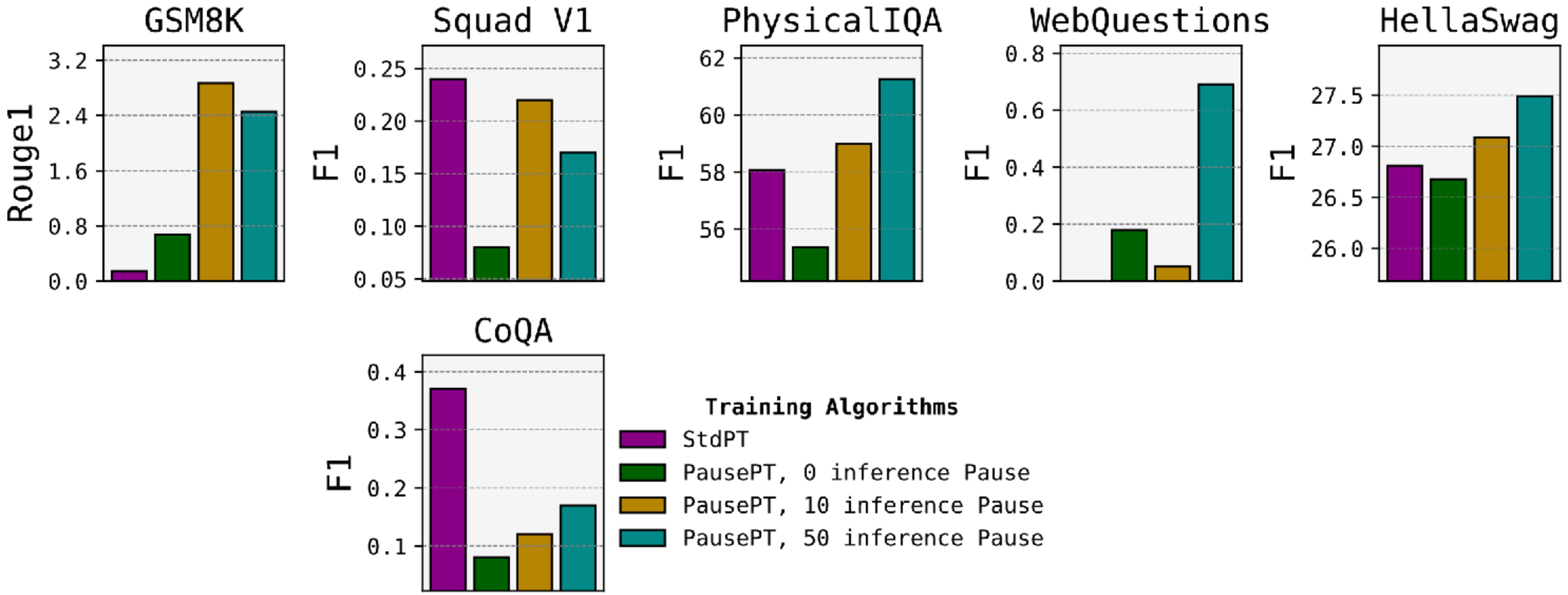


Figure 6: **Zero-shot evaluation of pause-pretrained models.** Zero-shot inference with pause tokens on a pause-pretrained model gives gains on tasks like GSM8k and HellaSwag. However, we note that our zero-shot accuracies are quite low, as we experiment with a small 1B parameter model.

"Think before you speak: Training Large Language Models With Pause Tokens". Goyal *et al*, ICLR 2024, arXiv:2310.02226

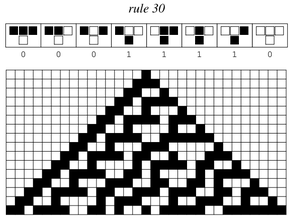
Register tokens improve arithmetic performance.

- They don’t know about **world models**!
- Their experiments are **limited**.
  - Pre-training objective/dataset/etc.
  - Register token clumping.
- Can I add **way more registers -> way better performance?**

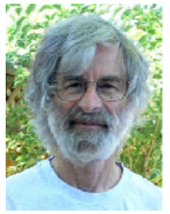
“Think before you speak: Training Large Language Models With Pause Tokens”. Goyal *et al*, ICLR 2024, arXiv:2310.02226

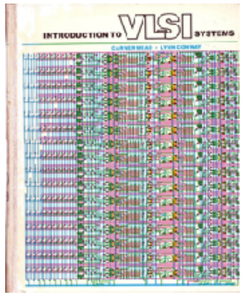
CNS Heroes

Toward Cellular Computing



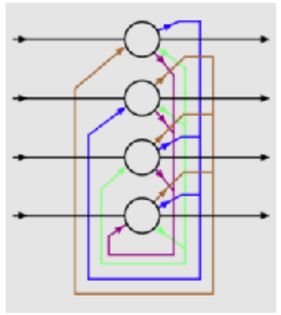
Thinking Machines Corporation



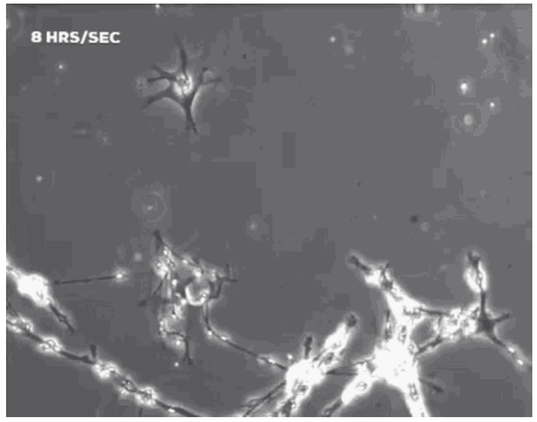


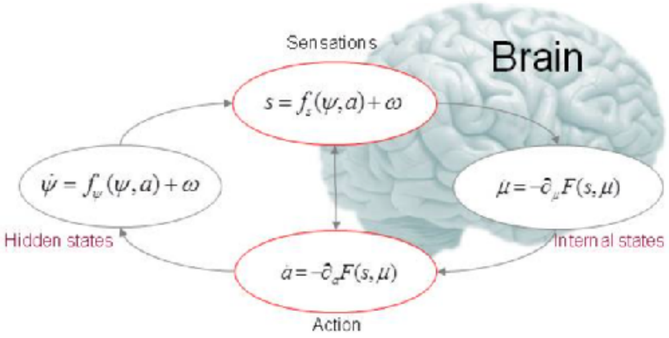
Mead/Conway Neuromorphics

Feynman/Wolfram/Winfree/Levin/Thomson
Cellular/bio/chemical computation



Hopfield/Comp Neuro



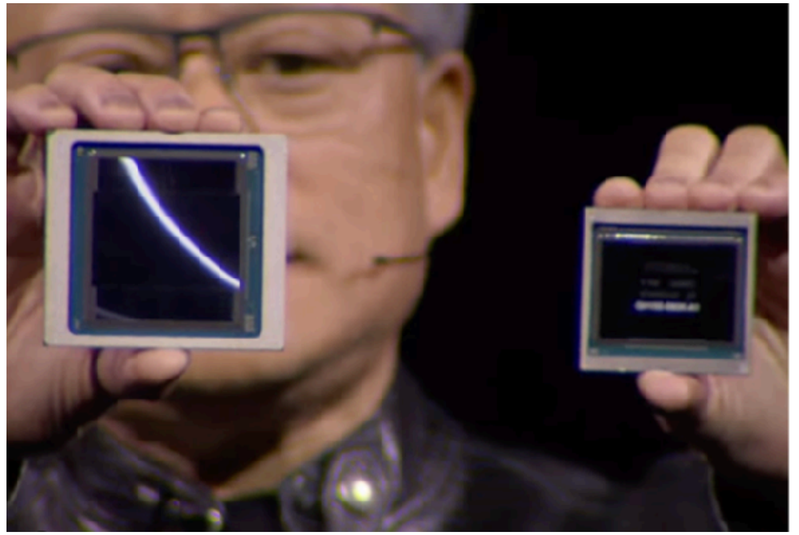


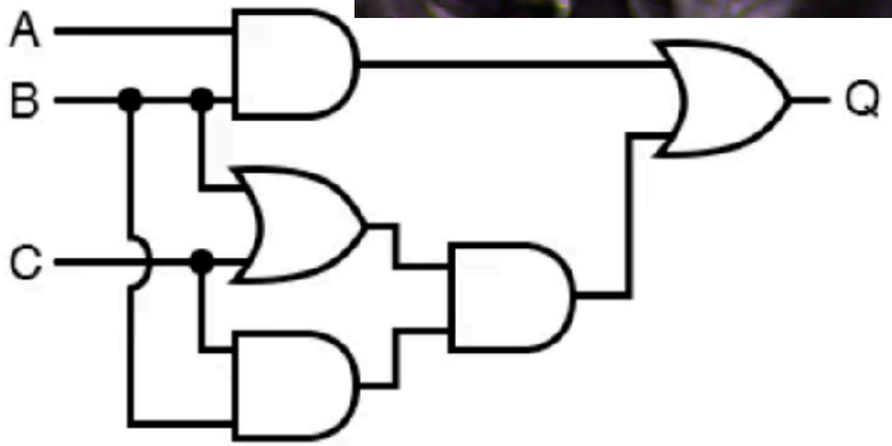
Friston/Free Energy Principle

Beyond LLMs

Toward Cellular Computing

Sand

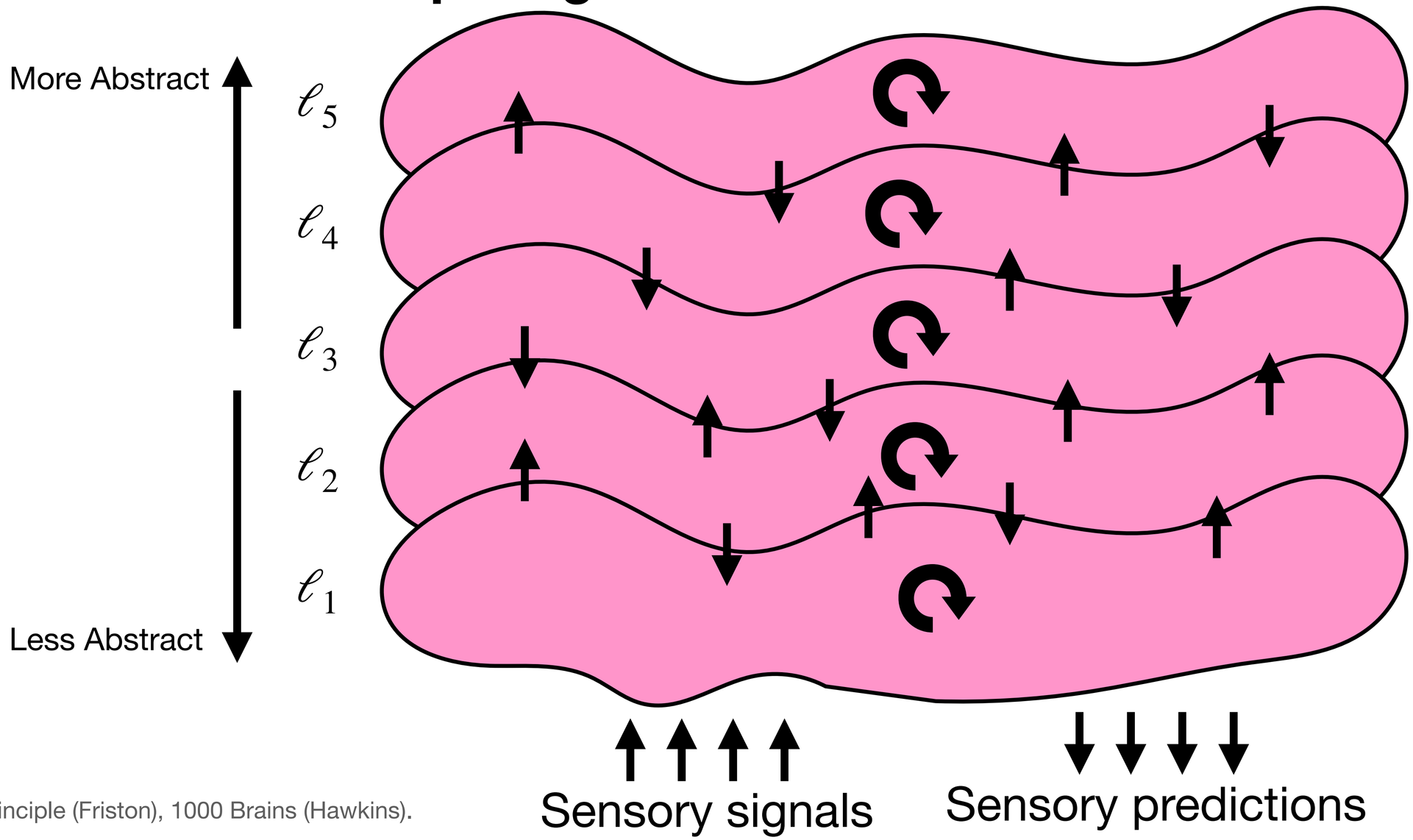




Thinking Sand

Beyond LLMs

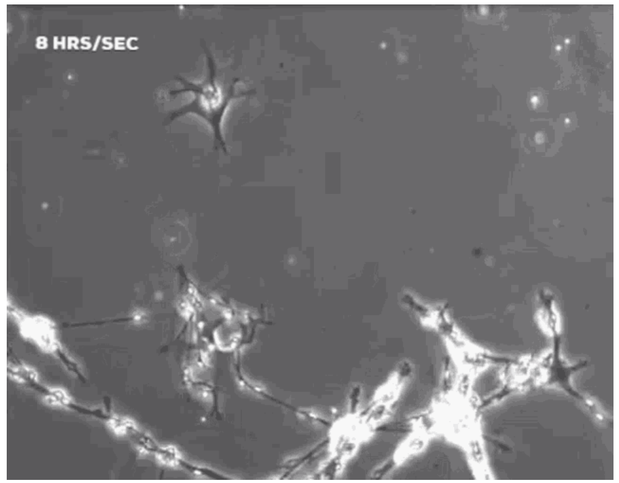
Toward Cellular Computing

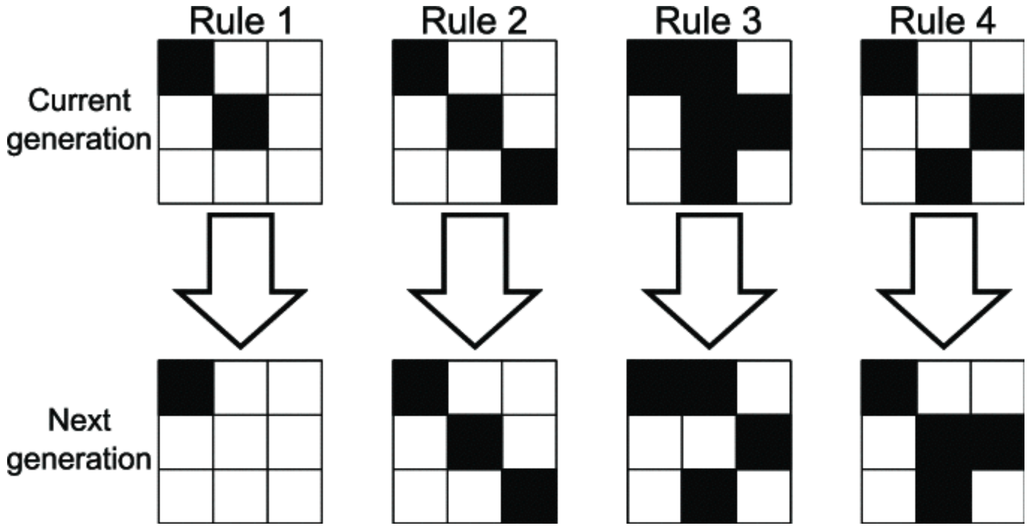


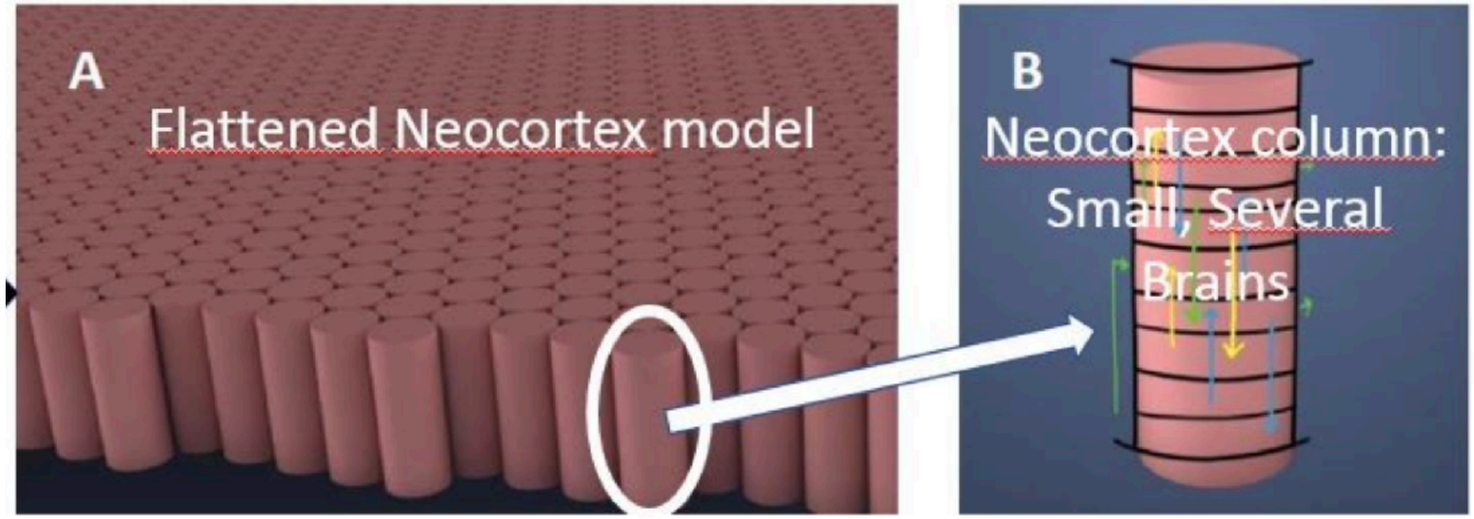
Free Energy Principle (Friston), 1000 Brains (Hawkins).

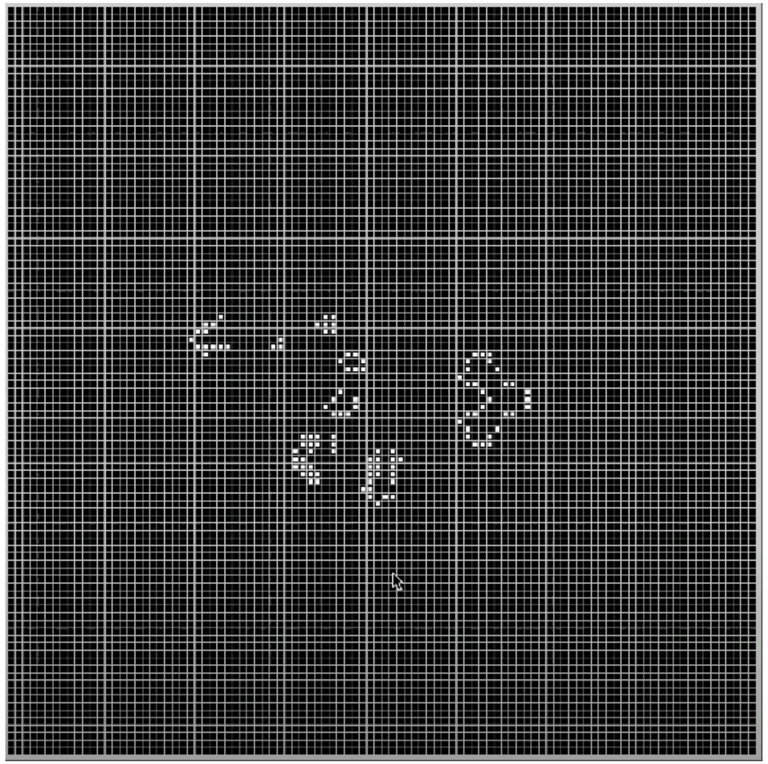
Beyond LLMs

Cortical Universal Update Rule





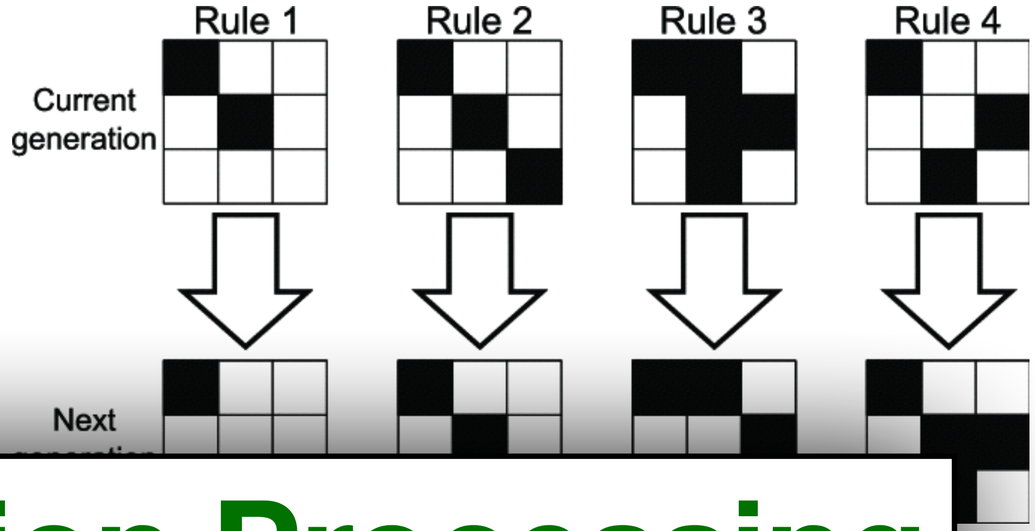




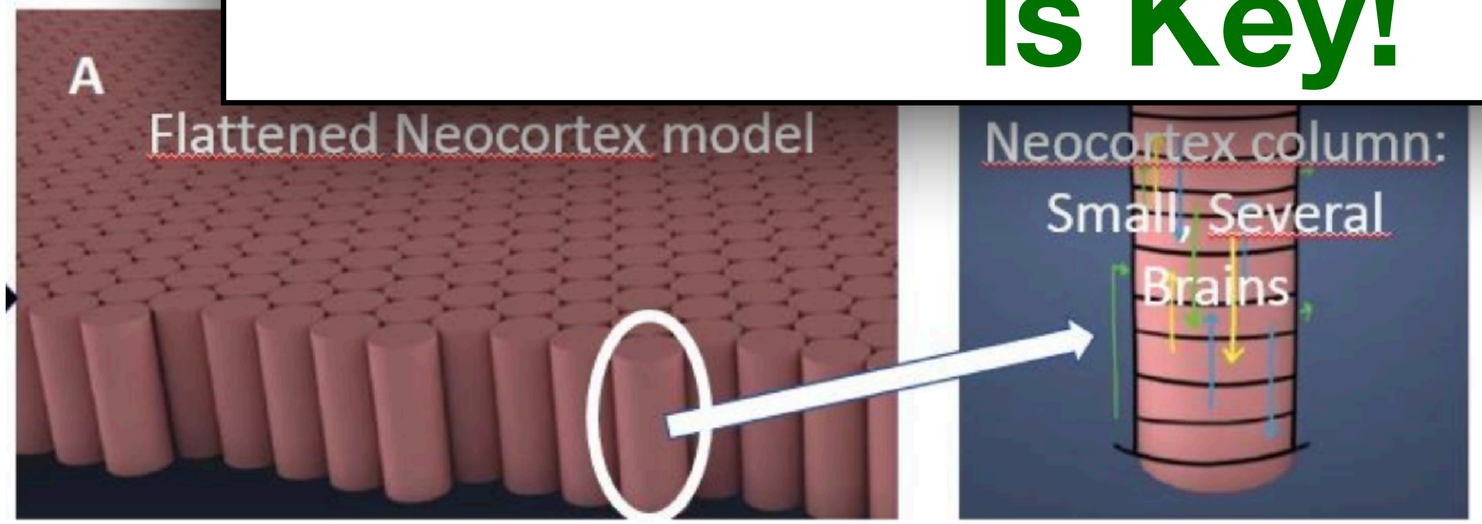
FEP (Friston), 1000 Brains (Hawkins), New Kind of Science (Wolfram), etc.

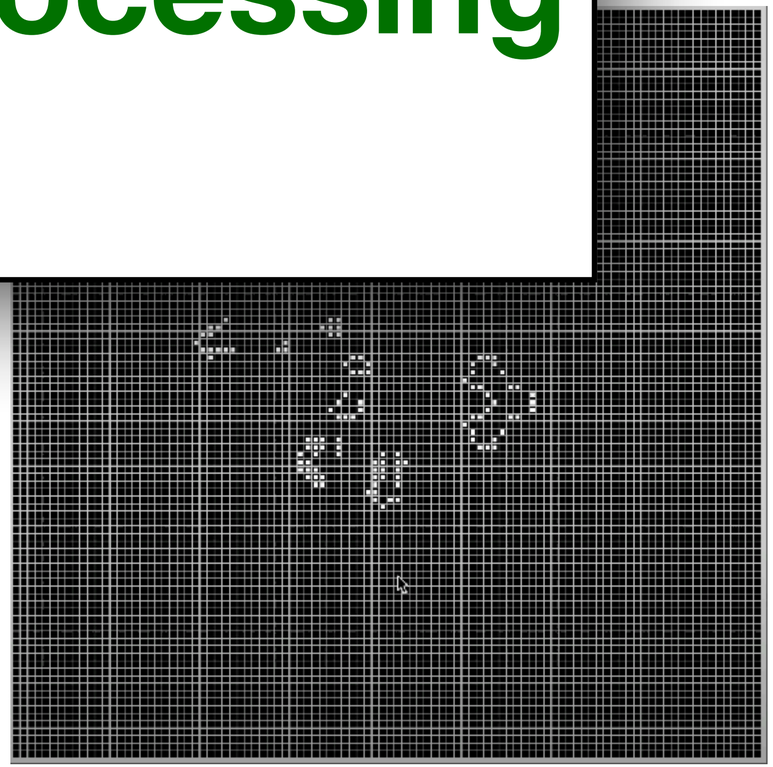
Beyond LLMs

Cortical Universal Update Rule



Local Information Processing is Key!

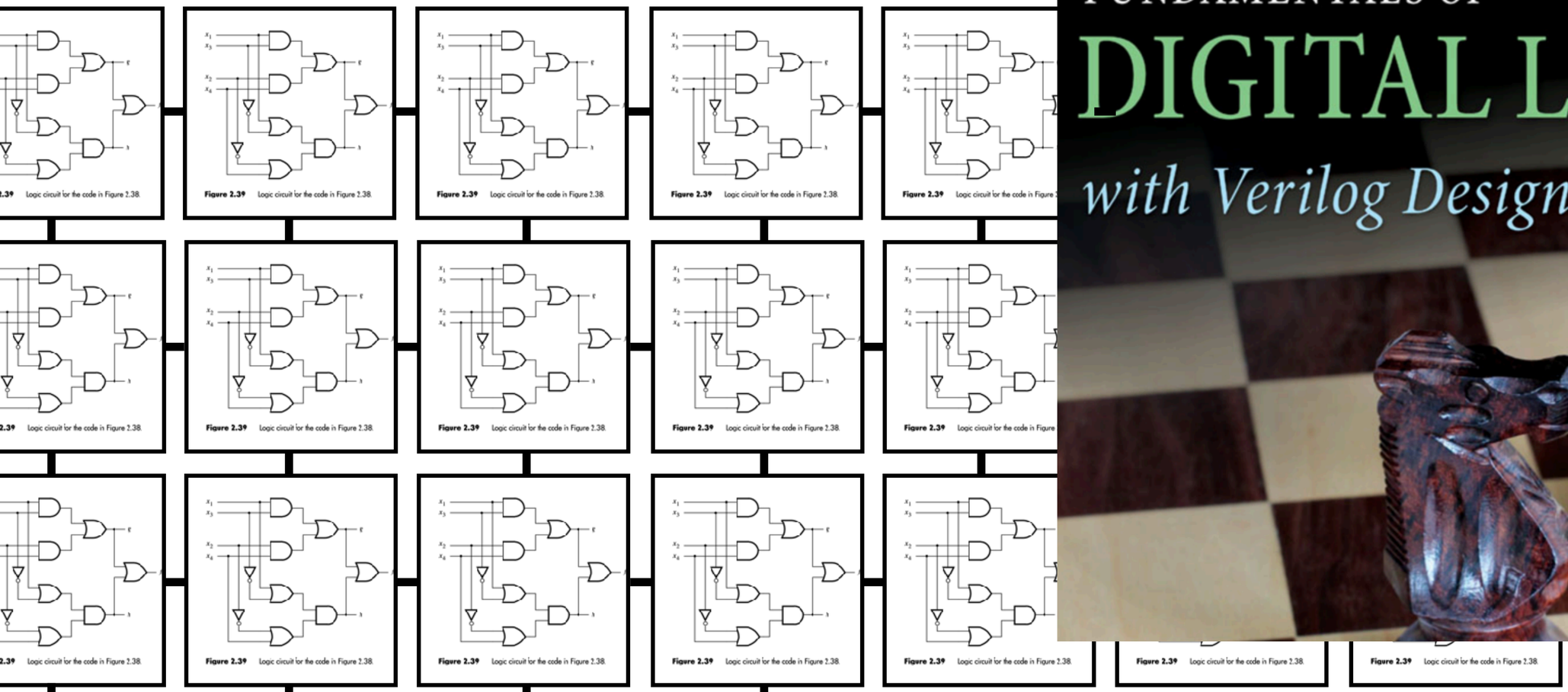




FEP (Friston), 1000 Brains (Hawkins), New Kind of Science (Wolfram), etc.

Beyond LLMs

Cortical Universal Update Rule



Δt

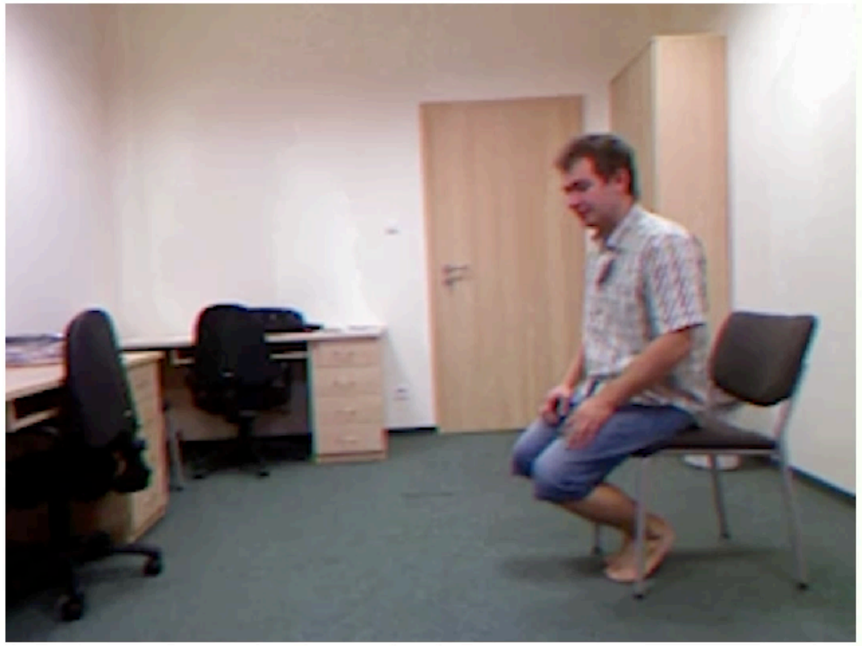
FEP (Friston), 1000 Brains (Hawkins), New Kind of Science (Wolfram), etc.

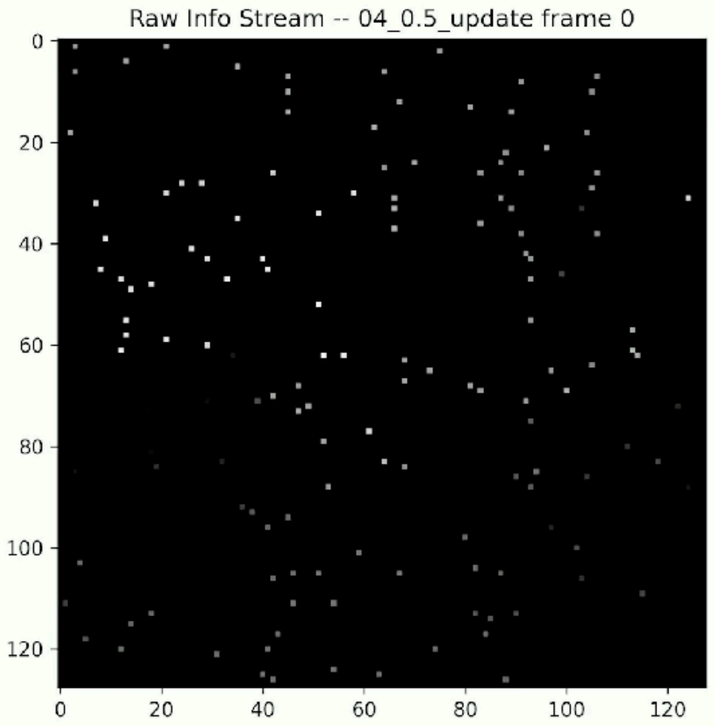
Toward Cellular Computing

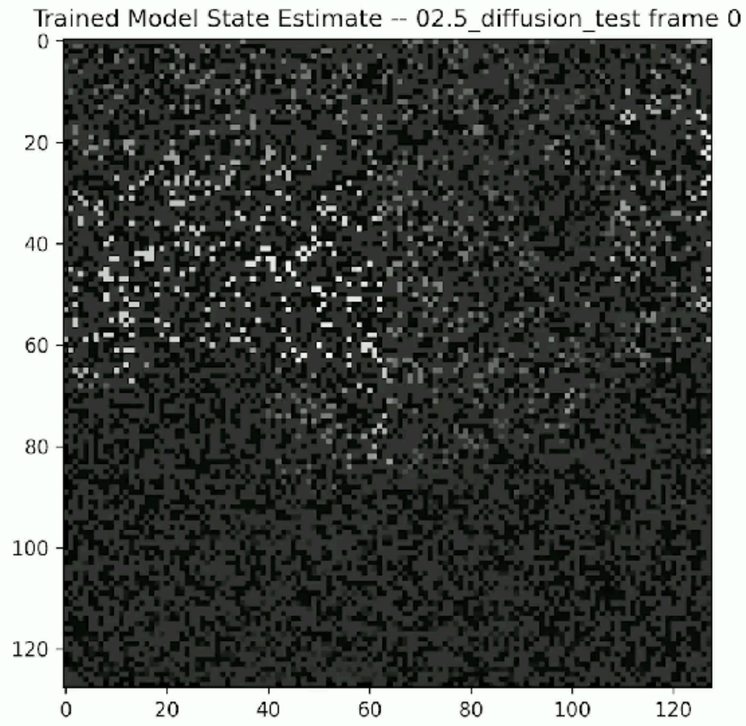
Limit case of physically-realizable computation

Neural Cellular Automata do Active Inference

Source Video



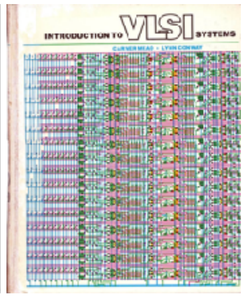




aman-bhargava.com

Toward Cellular Computing

Limit case of physically-realizable computation

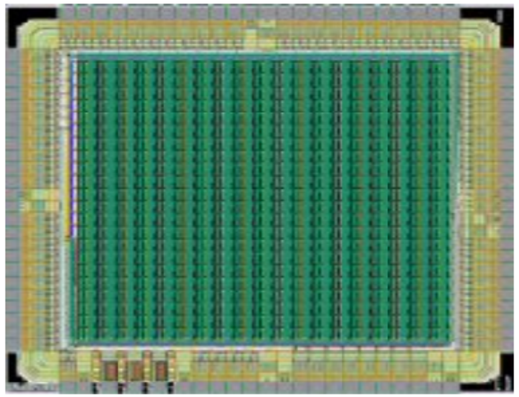


Mead/Conway Neuromorphics

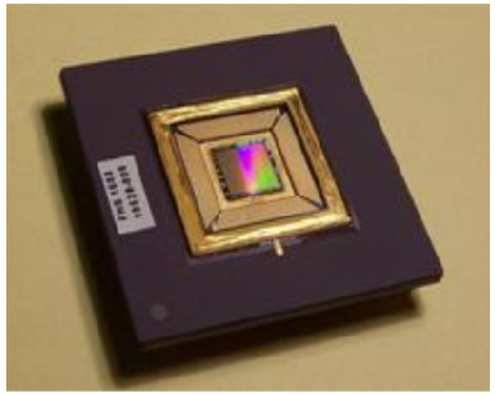
Piotr Dudek

The University of Manchester
Verified email at manchester.ac.uk - Homepage

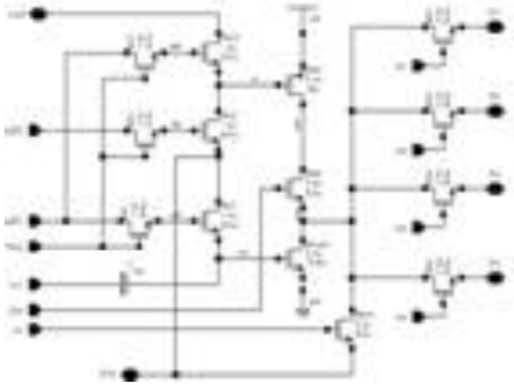
VLSI Design Neuromorphic Engineering Imag



Cellular processors arrays



Vision Chips



Analog Sig Proc

https://personalpages.manchester.ac.uk/staff/p.dudek/

Thank You!

Questions?

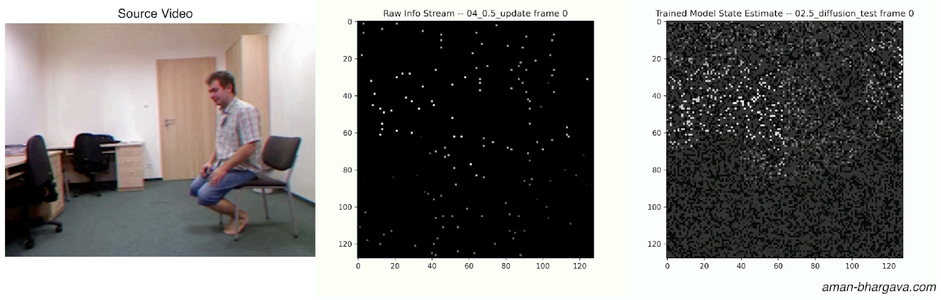
Mentors

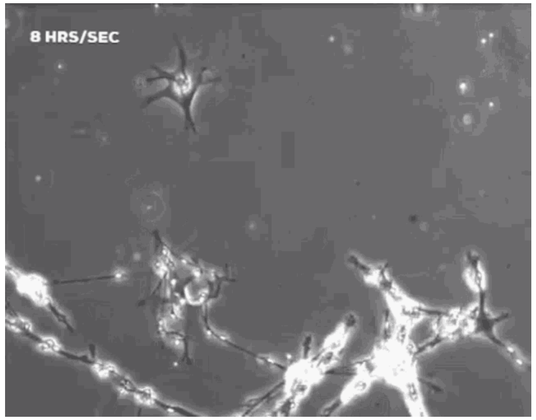
- Matt Thomson
- Erik Winfree
- Ralph Adolphs
- Frederick Eberhart
- Rob Phillips
- Steve Mann (UToronto)
- Milad Lankarany (UToronto)
- Michael Levin (Tufts)

Friends + Collaborators

- Pantelis Vafidis
- Cameron Witkowski (UToronto)
- Salvador Buse
- Cayden Pierce (MIT)
- James Gornet
- Meera Prasad
- Michael Zellinger
- Hersh Bhargava (UCSF)
- Mango Weng
- Mingshi Chi (UToronto/York)

Neural Cellular Automata do Active Inference





aman-bhargava.com

[EOS]

Transformer theory supports register tokens.

Circuit Complexity of Chain-of-Thought

step towards theoretically answering these questions. Specifically, we examine the *expressivity* of LLMs with CoT in solving fundamental mathematical and decision-making problems. By using circuit complexity theory, we first give impossibility results showing that bounded-depth Transformers are unable to directly produce correct answers for basic arithmetic/equation tasks unless the model size grows *super-polynomially* with respect to the input length. In contrast, we then prove by construction that autoregressive Transformers of *constant size* suffice to solve both tasks by generating CoT derivations using a commonly used math language format. Moreover, we show LLMs with CoT can handle a general class of decision-making problems known as Dynamic Programming, thus justifying their power in tackling

From “Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective” (Feng et al, 2023) — arxiv:2305.15408

Register tokens help vision transformers.

Vision Transformers Need Registers

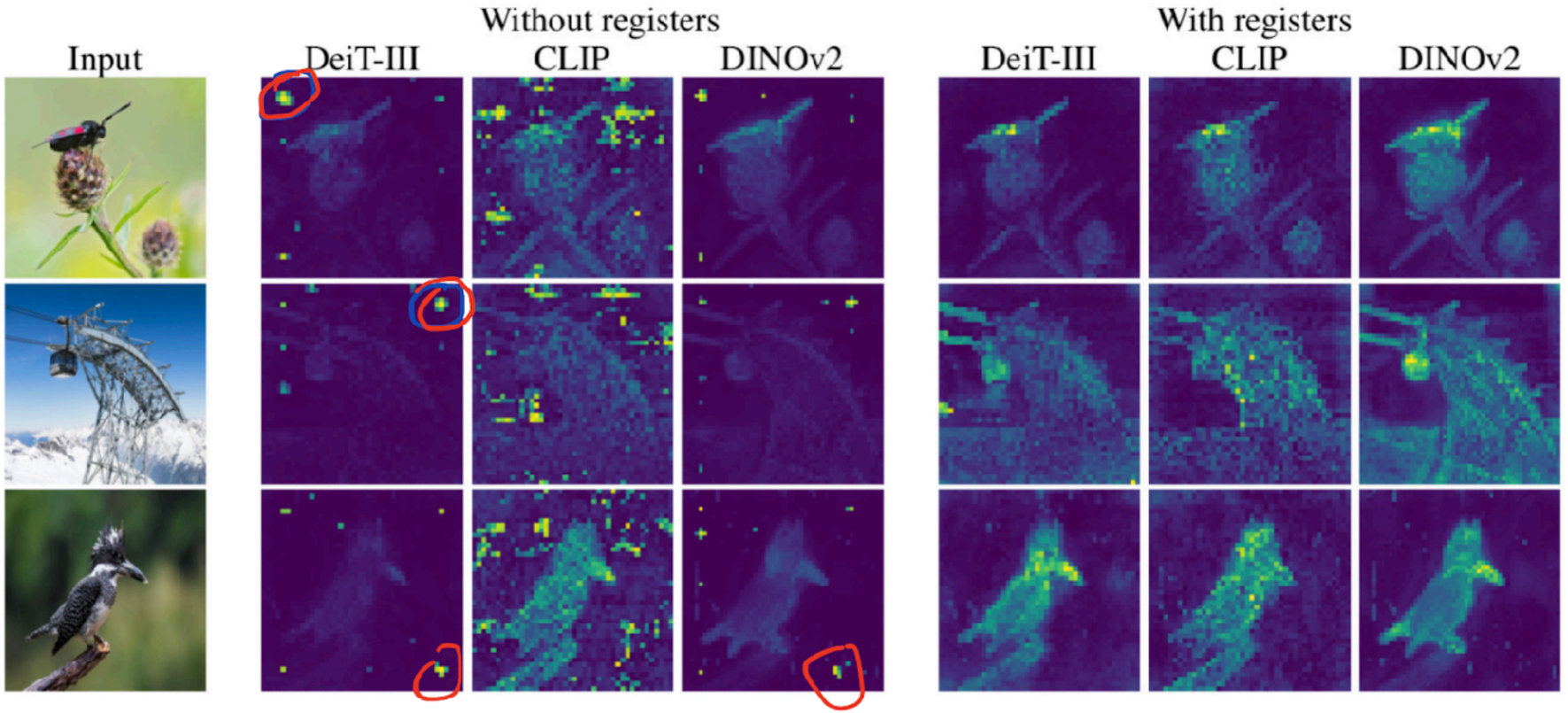
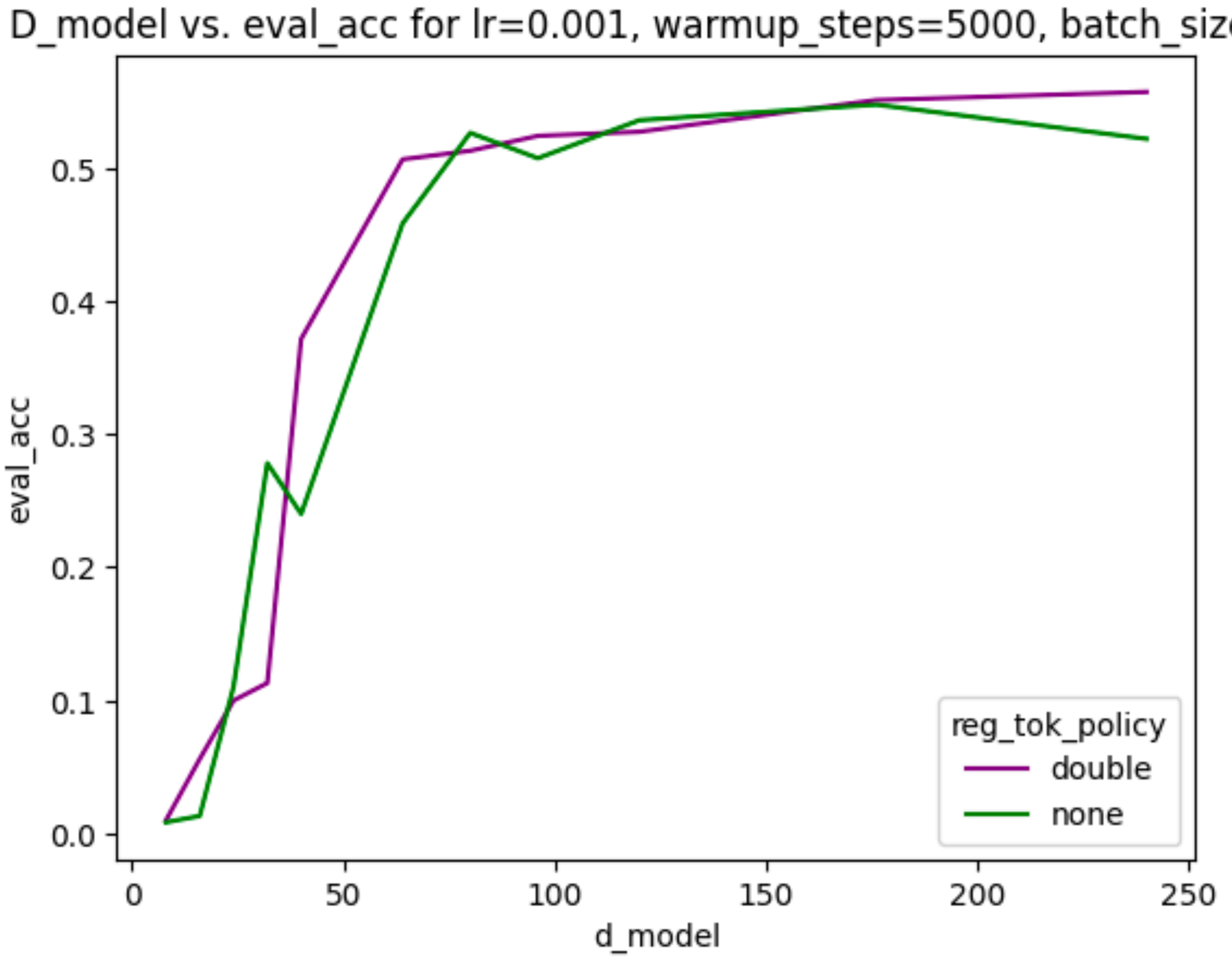


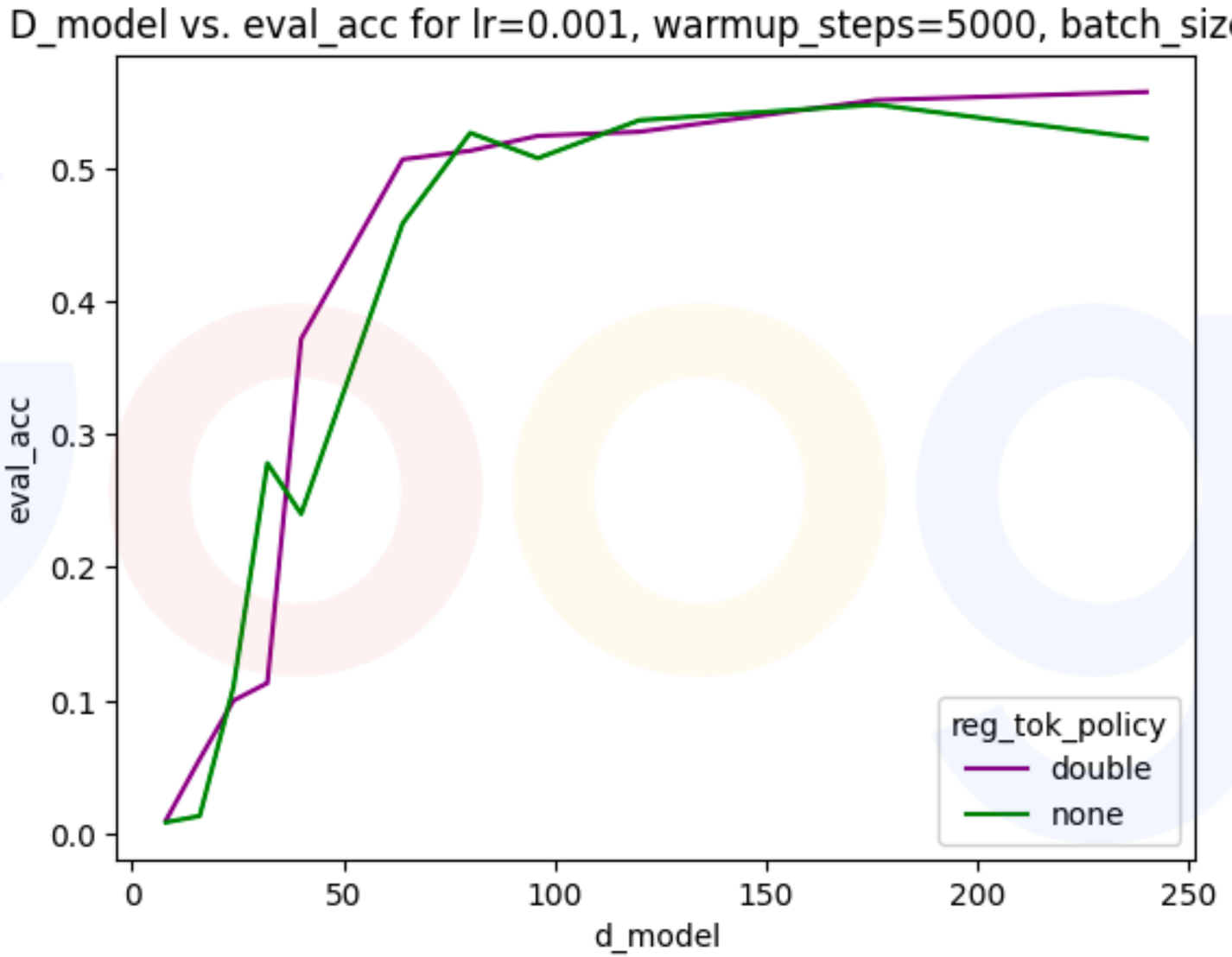
Figure 1: Register tokens enable interpretable attention maps in all vision transformers, similar to the original DINO method (Caron et al., 2021). Attention maps are calculated in high resolution for better visualisation. More qualitative results are available in appendix D.

From “Vision Transformers Need Registers” (Darcet et al) — arxiv:2309.16588

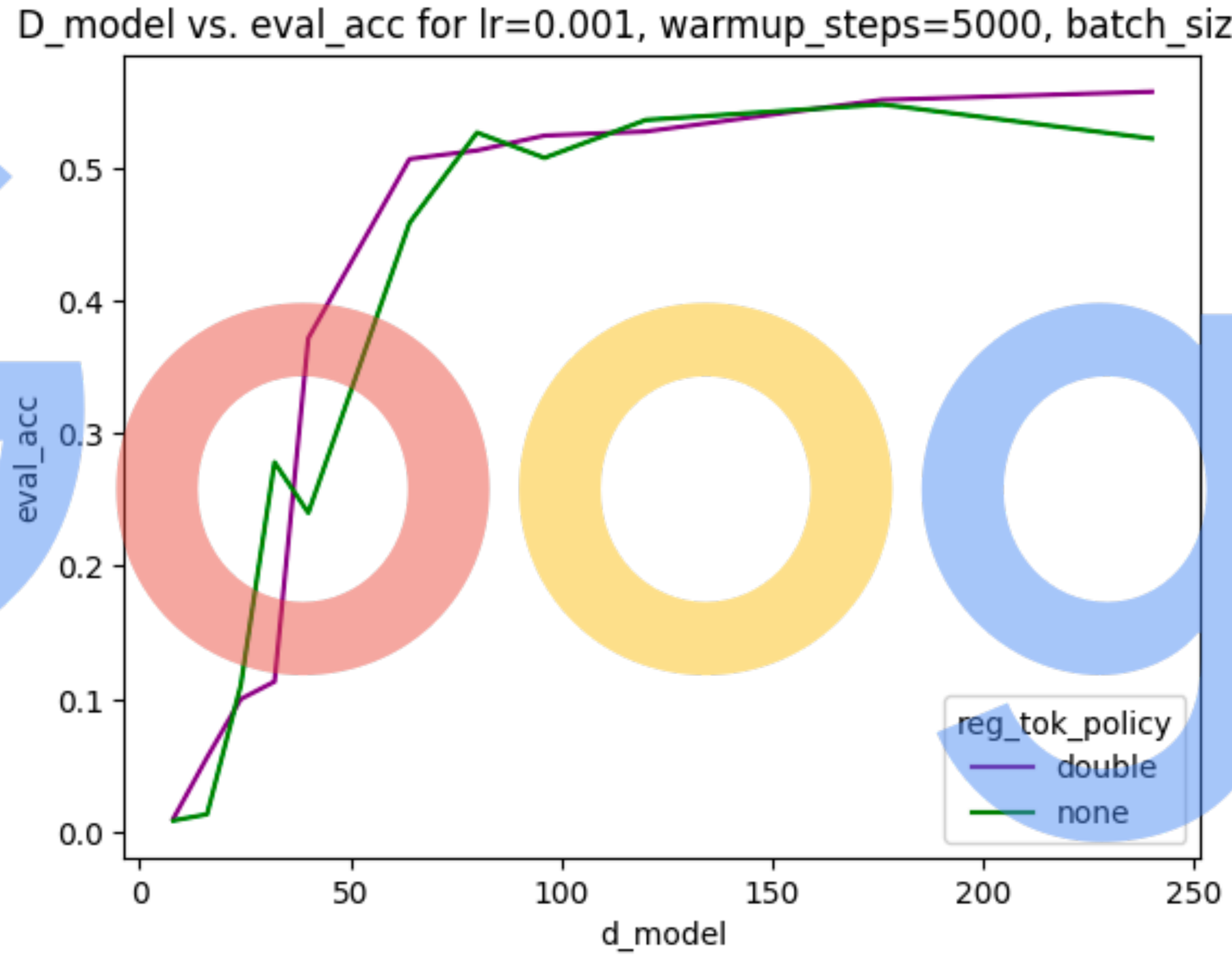
Register tokens improve arithmetic performance.



Register tokens improve arithmetic performance.



Register tokens improve arithmetic performance.



Register Tokens enable LLM world model exploration.

Big-Picture

- ✅ Should be able to **“stop and think”**, leveraging the language “world model”.
- ✅ Thinking should be optimized **during pre-training.**
- ✅ **Inference-time scalability** would be nice.

Thought as internal world model exploration

Cortical processing cartoon sketch

AppliedMath

MDPI

Article

Gradient-Free Neural Network Training via Synaptic-Level Reinforcement Learning

Aman Bhargava [1,2], Mohammad R. Rezaei [2,3,4] and Milad Lankarany [2,3,4,5,*]

1 Division of Engineering Science, University of Toronto, Toronto, ON M5S 2E4, Canada; aman.bhargava@mail.utoronto.ca
2 Division of Clinical and Computational Neuroscience, Krembil Brain Institute, University Health Network, Toronto, ON M5G 1L7, Canada; mr.rezaei@mail.utoronto.ca
3 Institute of Biomedical Engineering, University of Toronto, Toronto, ON M5S 2E4, Canada
4 KITE, Toronto Rehabilitation Institute, University Health Network, Toronto, ON M5G 1L7, Canada
5 Department of Physiology, University of Toronto, Toronto, ON M5S 2E4, Canada
* Correspondence: milad.lankarany@uhnresearch.ca

World Models emerge from Predictive Coding.

Ilya Sutskever (Chief Scientist & Co-Founder, OpenAI)

Bigger LLM → Better Prediction → Greater Capability

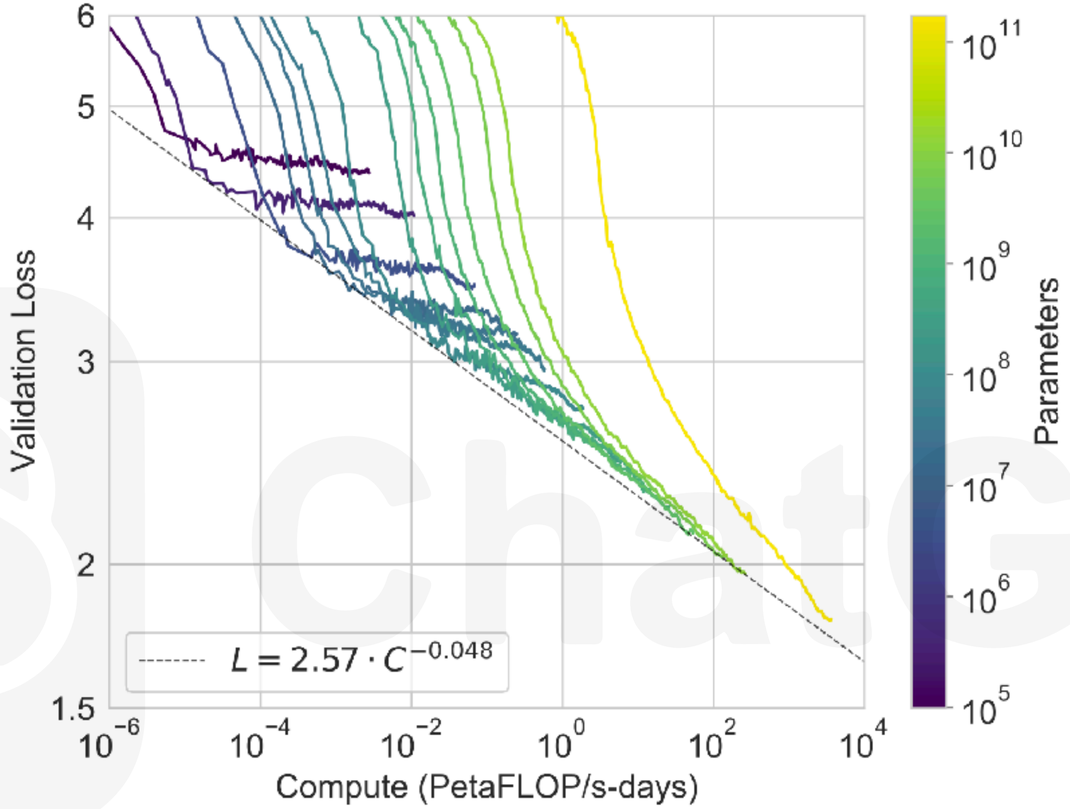


Figure 3.1: Smooth scaling of performance with compute. Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training. The power-law behavior observed in [KMH+20] continues for an additional two orders of magnitude with only small deviations from the predicted curve. For this figure, we exclude embedding parameters from compute and parameter counts.

From the OpenAI GPT-3 Paper “Language Models are Few-Shot Learners” (Brown et al) — arxiv:2005.14165

March 2024 LLMs Suck at Reasoning.

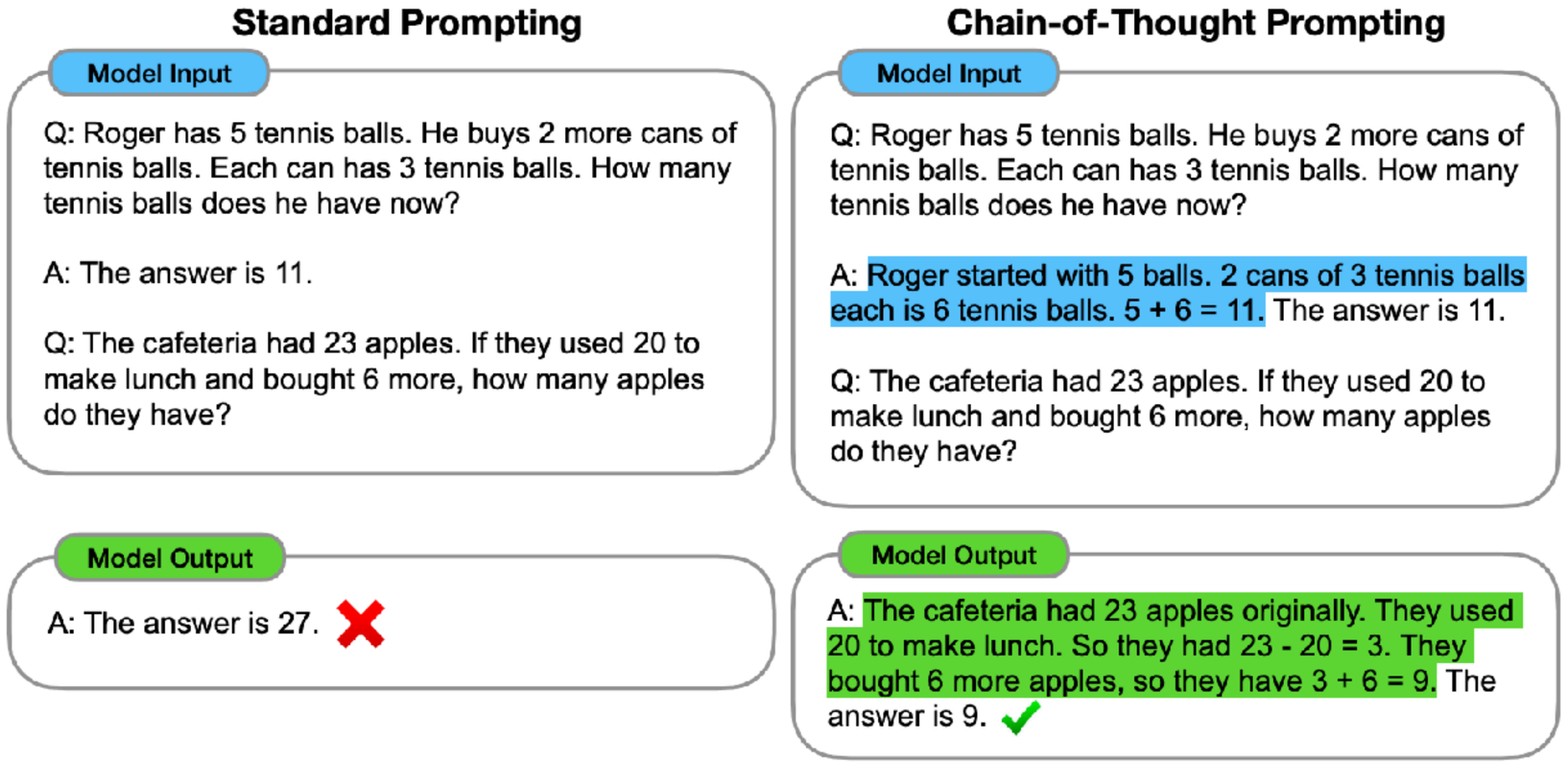
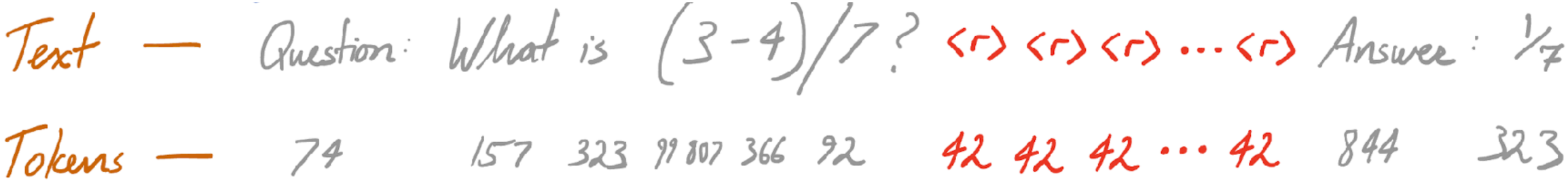


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

From “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models” (Wei et al) — arxiv:2201.11903

Register Tokens enable LLM world model exploration.



Register tokens improve arithmetic performance.

Figure 6: **Zero-shot ev** …ference with pause tokens on a pause-pretrai… …llaSwag. However, we note that our zero-shot … …ll 1B parameter model.

“Think before you speak: Training Large Language Models With Pause Tokens”. Goyal *et al*, ICLR 2024, arXiv:2310.02226