NEURAL INFORMATION PROCESSING SYSTEMS 2025

SCAN ME

SCAN ME

Concept-guided Interpretability via Neural Chunking

Shuchen Wu[1], Stephan Alaniz[2], Shyamgopal Karthik[5], Peter Dayan[4], Eric Schulz[3], Zeynep Akata[5]

1. Allen Institute, University of Washington; 2. Télécom Paris, Institut Polytechnique de Paris; 3. Department of Human-Centered AI, Helmholtz Munich; 4. Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics; 5. Department of Explainable Machine Learning, Helmholtz Munich

Project page: https://github.com/swu32/Chunk-Interpretability

Paper: https://arxiv.org/pdf/2505.11576

ALLEN INSTITUTE

Shanahan Family Foundation

HELMHOLTZ MUNICH

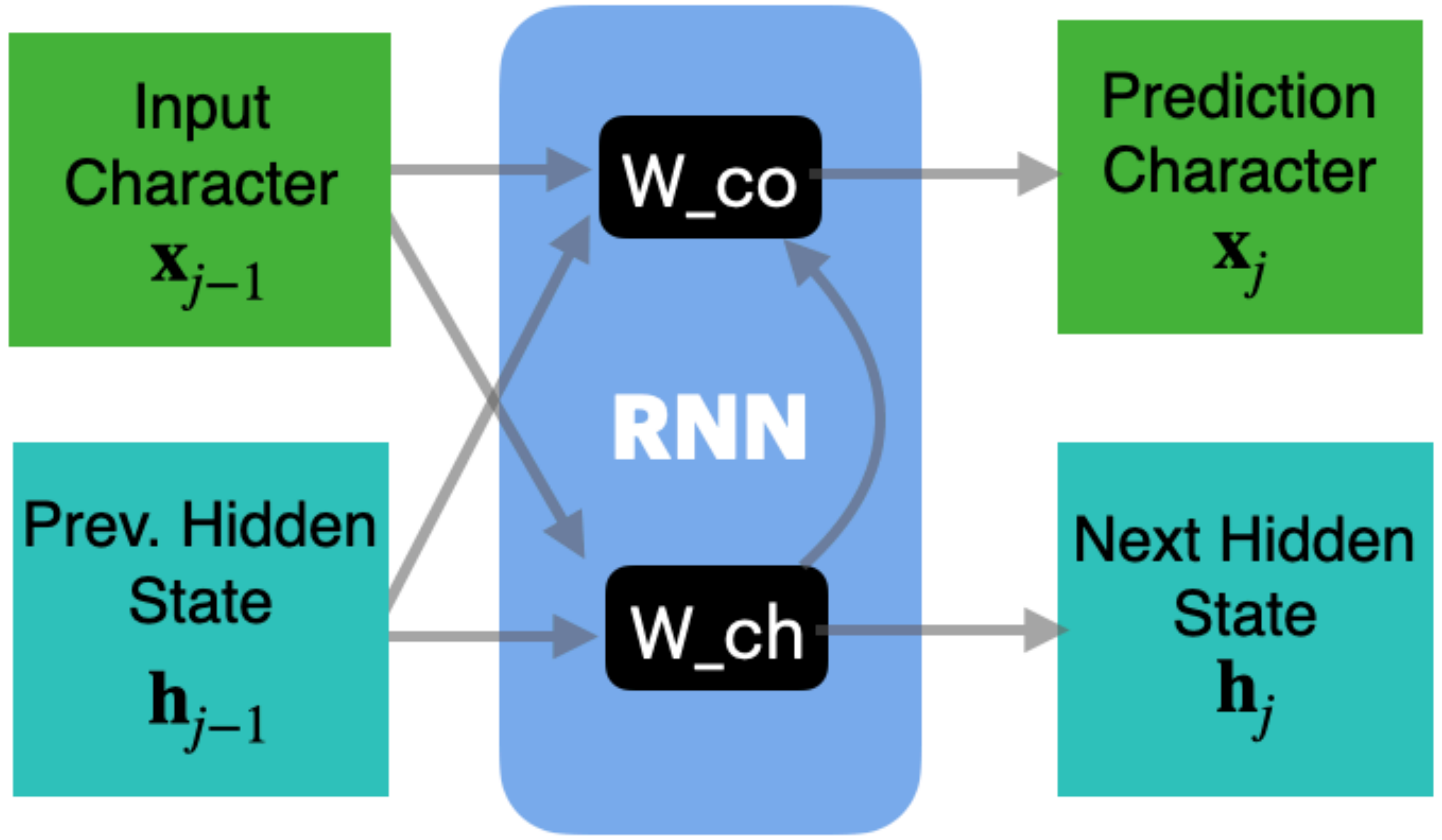
MAX PLANCK INSTITUTE FOR BIOLOGICAL CYBERNETICS

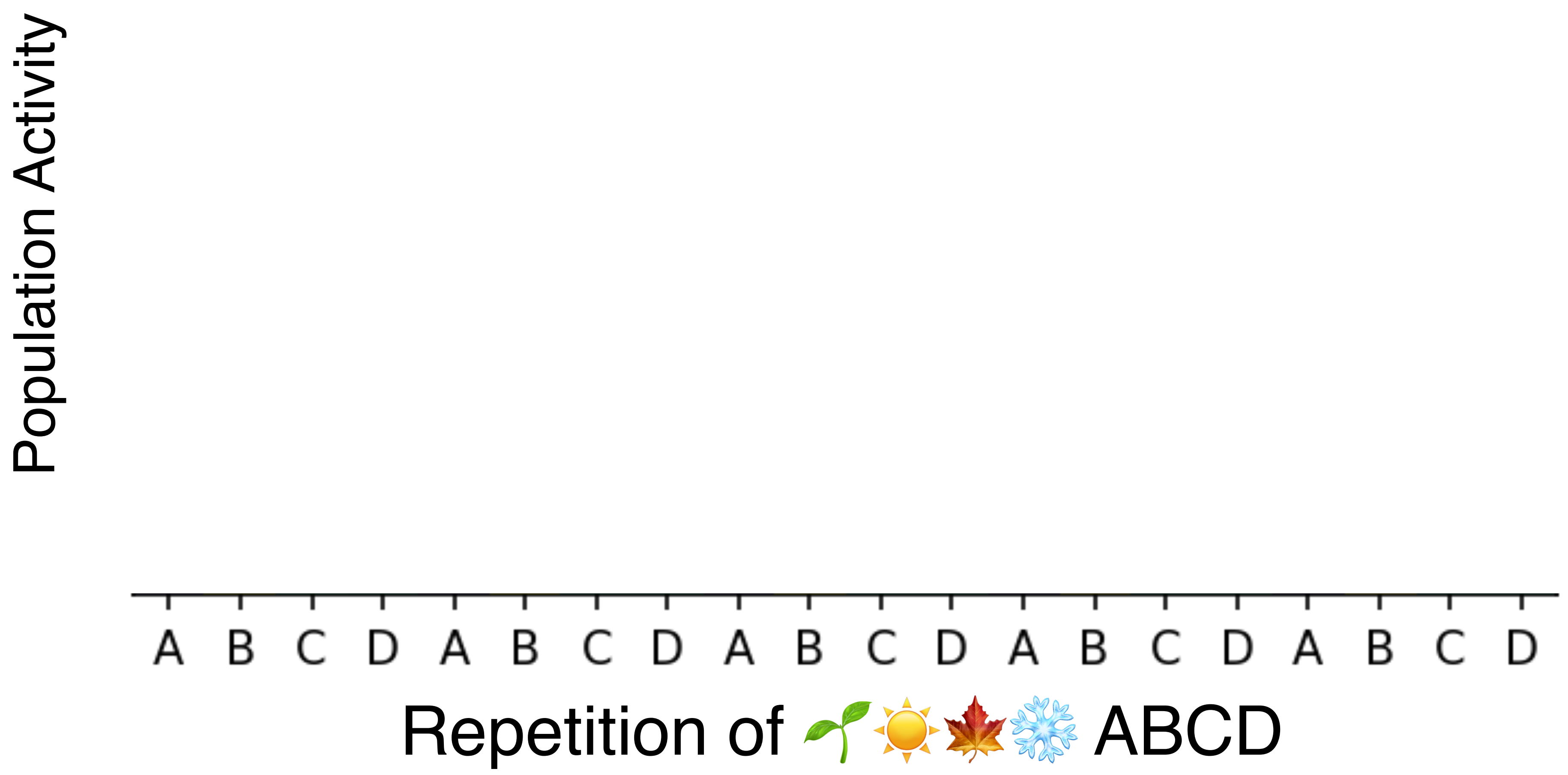
The Reflection Hypothesis

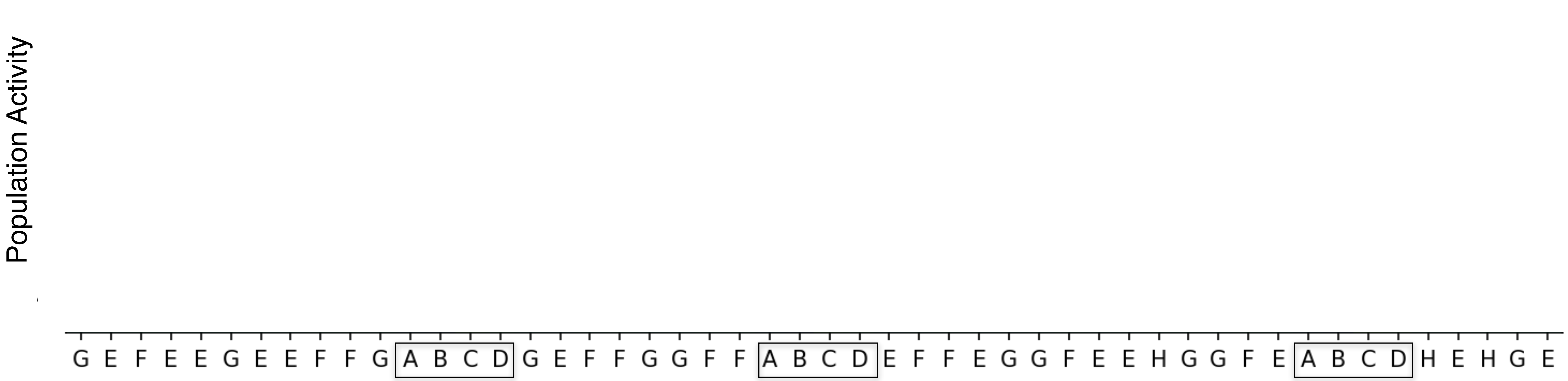
- Converging representations in diverse AI models (Balestriero & baraniuk 2018, Bansal et al. 2018, Dravid et al. 2023, Engels et al. 2024, Huh et al. 2024, Kornblith et al. 2019, Lenc and Vedaldi 2015) may be driven by regularities in naturalistic data

- A successfully predictive network should exhibit trajectories of neural activity that reflects the structure of the data

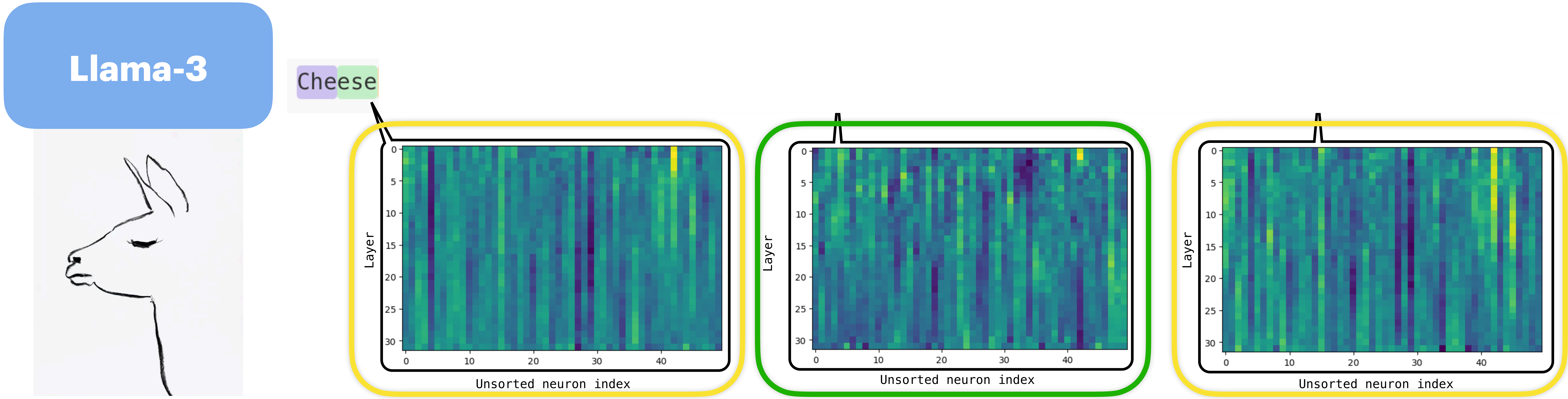
The Reflection Hypothesis — RNNs







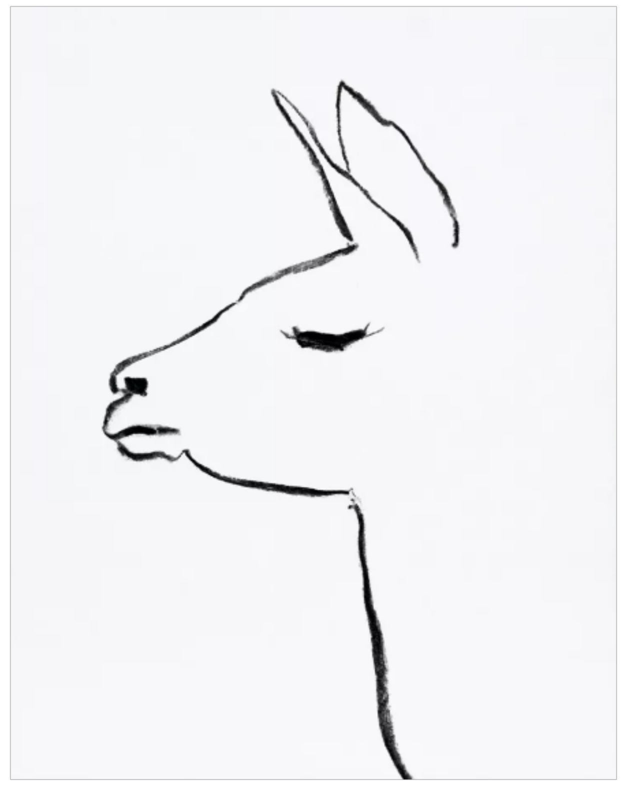
Recurring Concept Elicits Similar LLM Embeddings



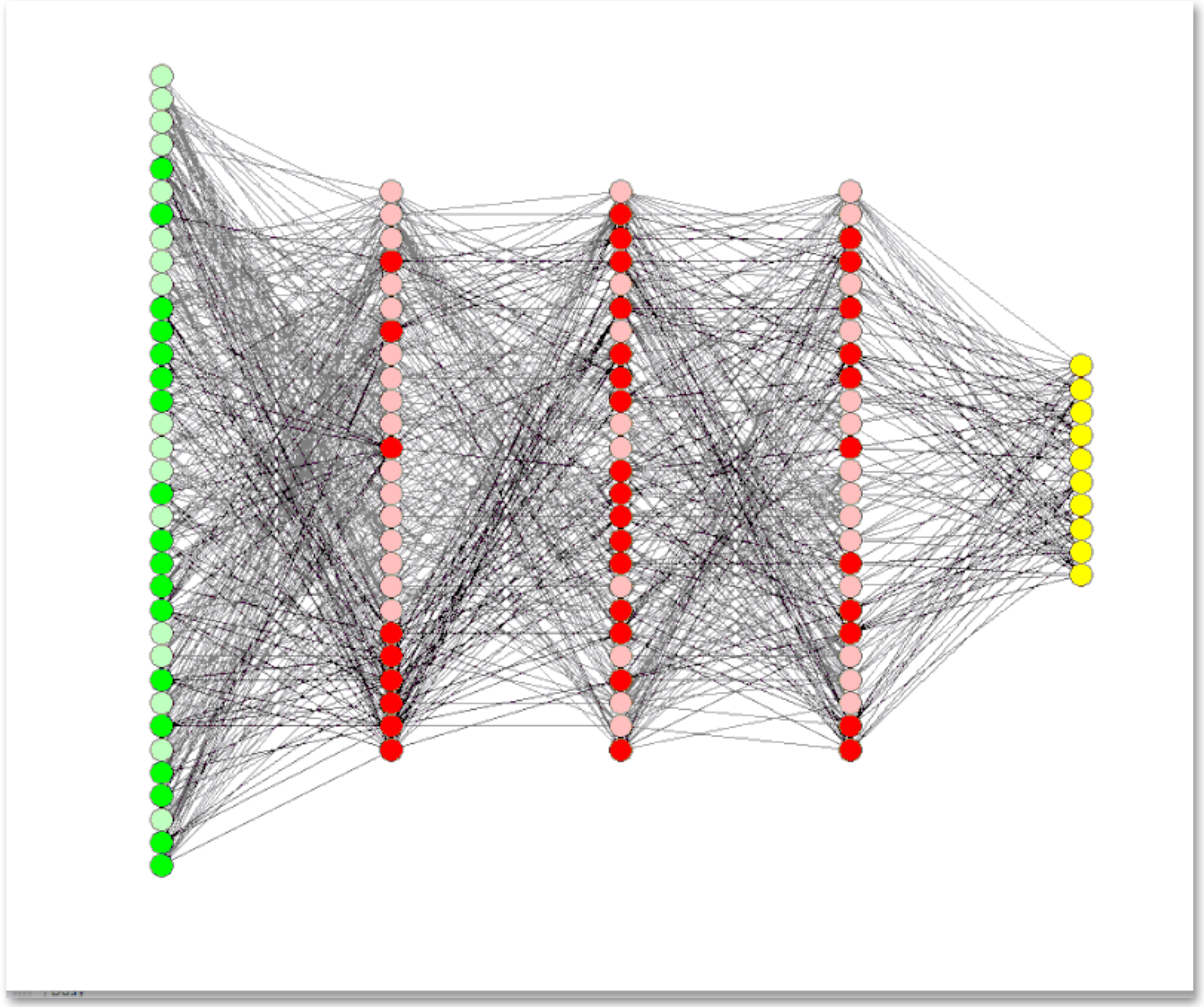
Neural Networks are Known as Black Boxes

RNN

$$
\begin{aligned}
\nabla \cdot \mathbf{E} &= \frac{\rho}{\varepsilon_0} \\
\nabla \cdot \mathbf{B} &= 0 \\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\
\nabla \times \mathbf{B} &= \mu_0 \left(\mathbf{J} + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right)
\end{aligned}
$$



LLaMA-3



- The interpretability challenge is cognitive.
- **What makes high-dimensional data meaningful for cognition?**

Cognition Interpret High Dimensional Neural Data in Chunks

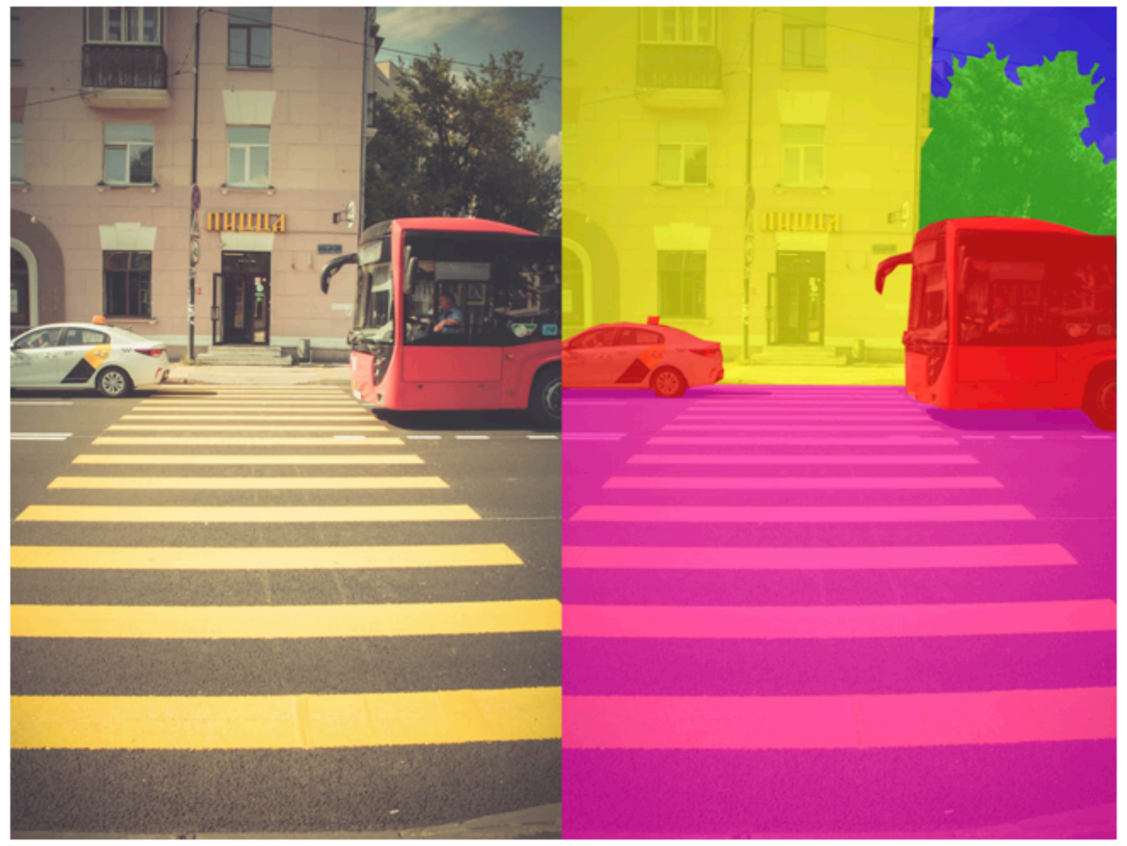
Language

... As you might know ...

... the thing is ...

Perruchet et al., 2014; McCauley & Christiansen, 2017

Vision



Penhune & Steele, 2012; Rosenbaum et al., 1983

- A sequence of high dimensional visual input $S_h = (\mathbf{h}^1, \mathbf{h}^2, \cdots, \mathbf{h}^n)$, $\mathbf{h}^i \in \mathbb{R}^d$
- A **chunk** is a ball $\overline{B}(\overline{\mathbf{h}}_C, \Delta) \subset \mathbb{R}^d$ centered at a prototypical activation vector $\overline{\mathbf{h}}_C$ in a subset of dimensions C with radius Δ

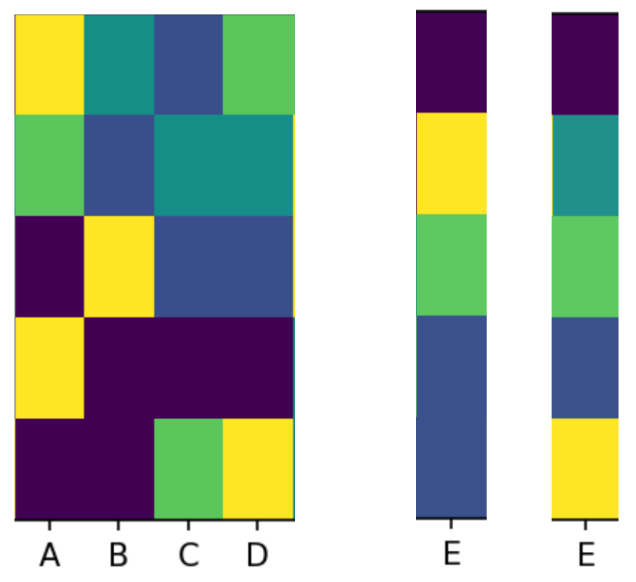
- Chunks constitute the basic units and entities for perceiving high-dimensional data
- Can we leverage how the mind understand high dimensional perceptual data, to understand high dimensional neural activations?

Three Methods to Extract Chunks

Discrete Sequence Chunking (DSC)

- Low dimensional data
- Learn dictionary of Spatial-temporal Chunks

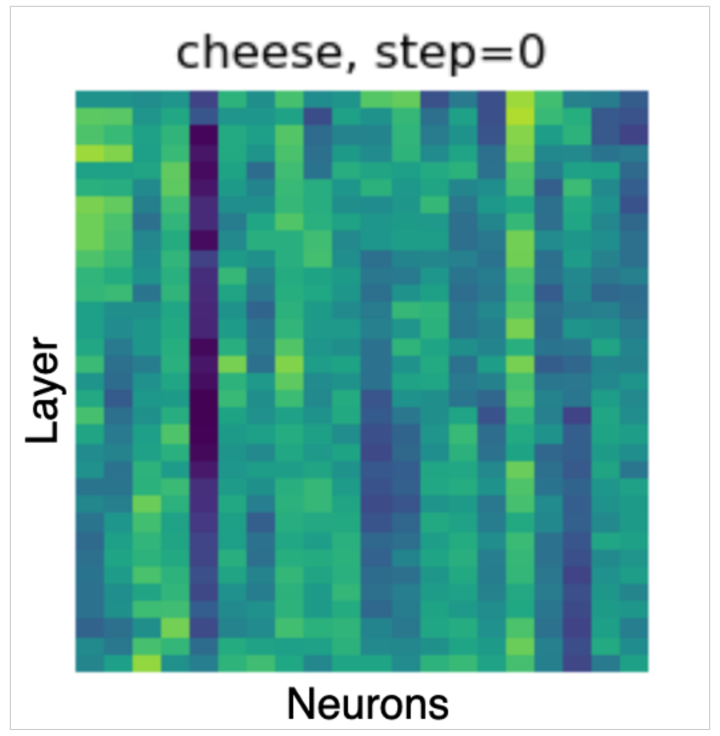
Vocabulary



Population Averaging (PA)

- High dimensional data
- Prototypical activation vector when label is present

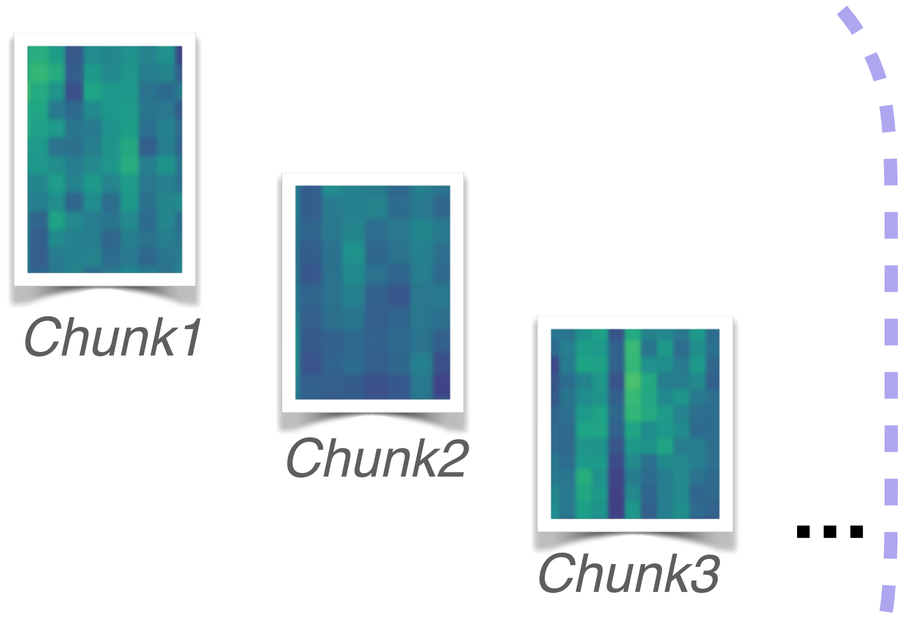
Vocabulary



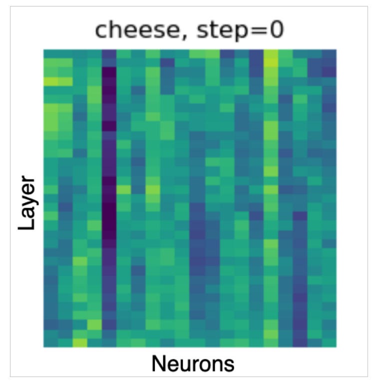
Unsupervised Chunk Discovery (UCD)

- High dimensional data
- Recurring chunk when label is unabsent

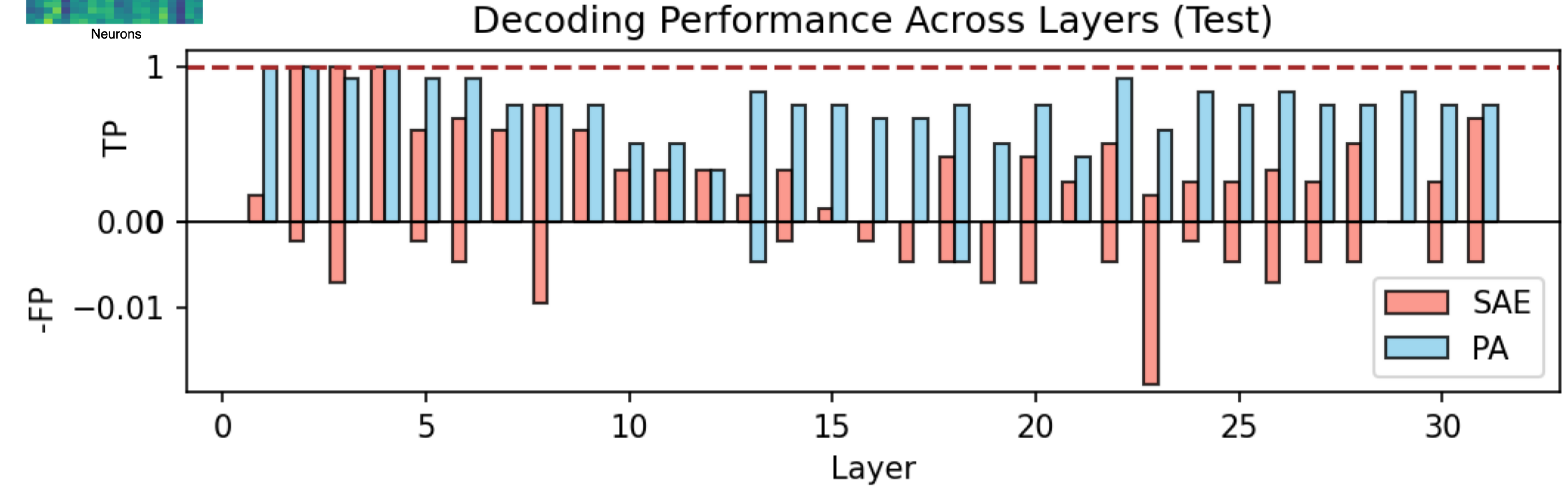
Vocabulary



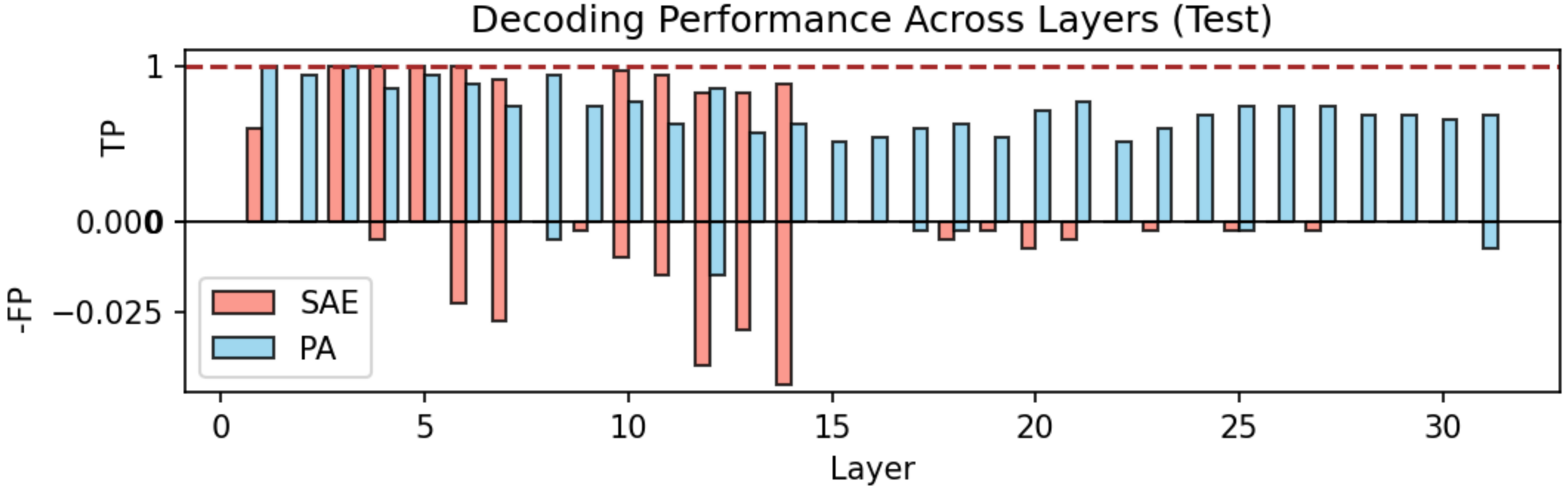
Concept Decoding Qualities for Chunks are Better than the Best SAE Latents

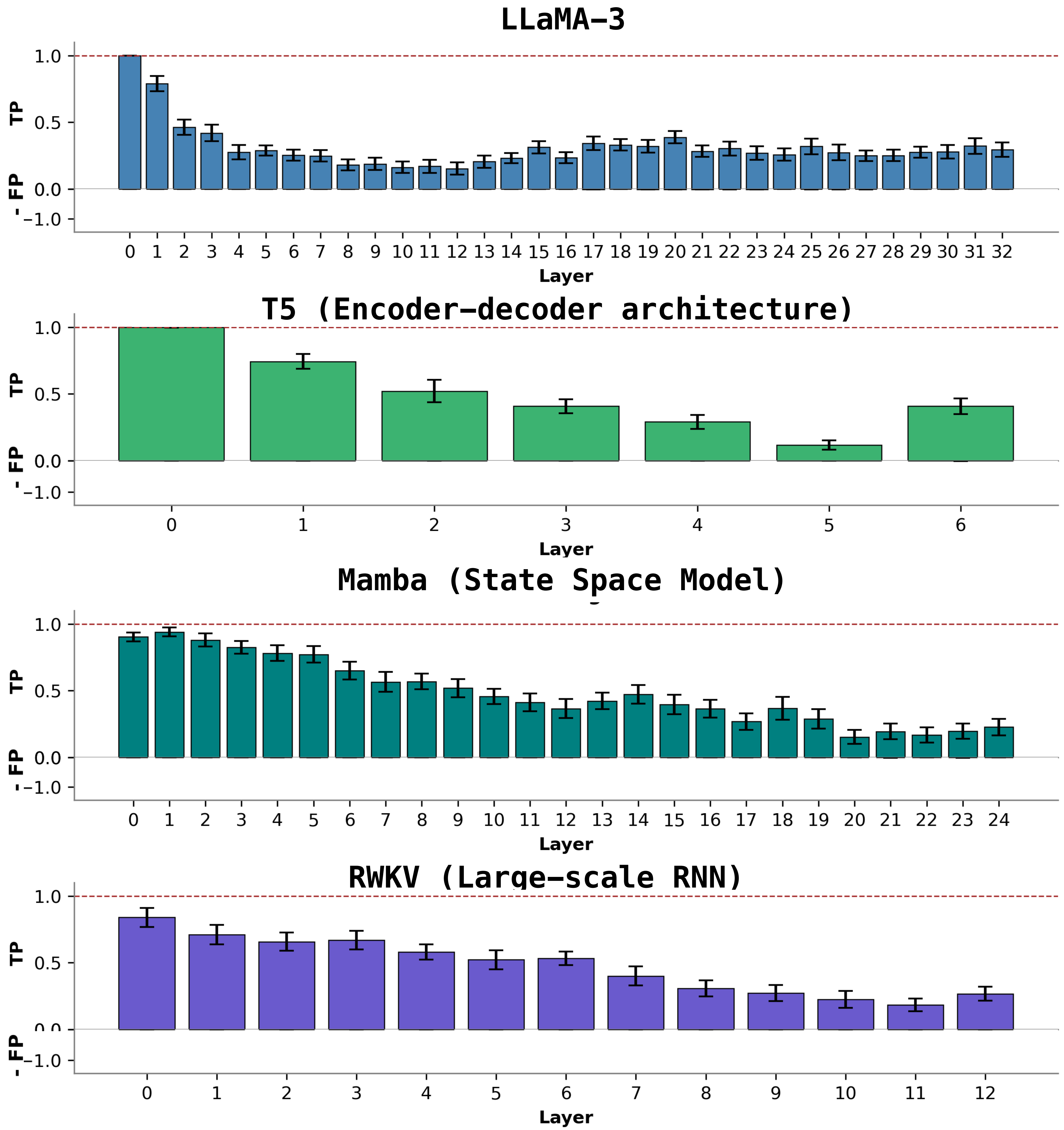


Cheese 🧀 Token = cheese cheese Cheese …



Cake 🍰 Token = Cake cake cake …





Method applies to a battery of concepts beyond illustrated, and many large models with distinct architectures

Beyond Concrete Concepts, Chunks Encode Abstract Sentence Schema

Go to food location → Purchase item → Go to eating location → React

Schema

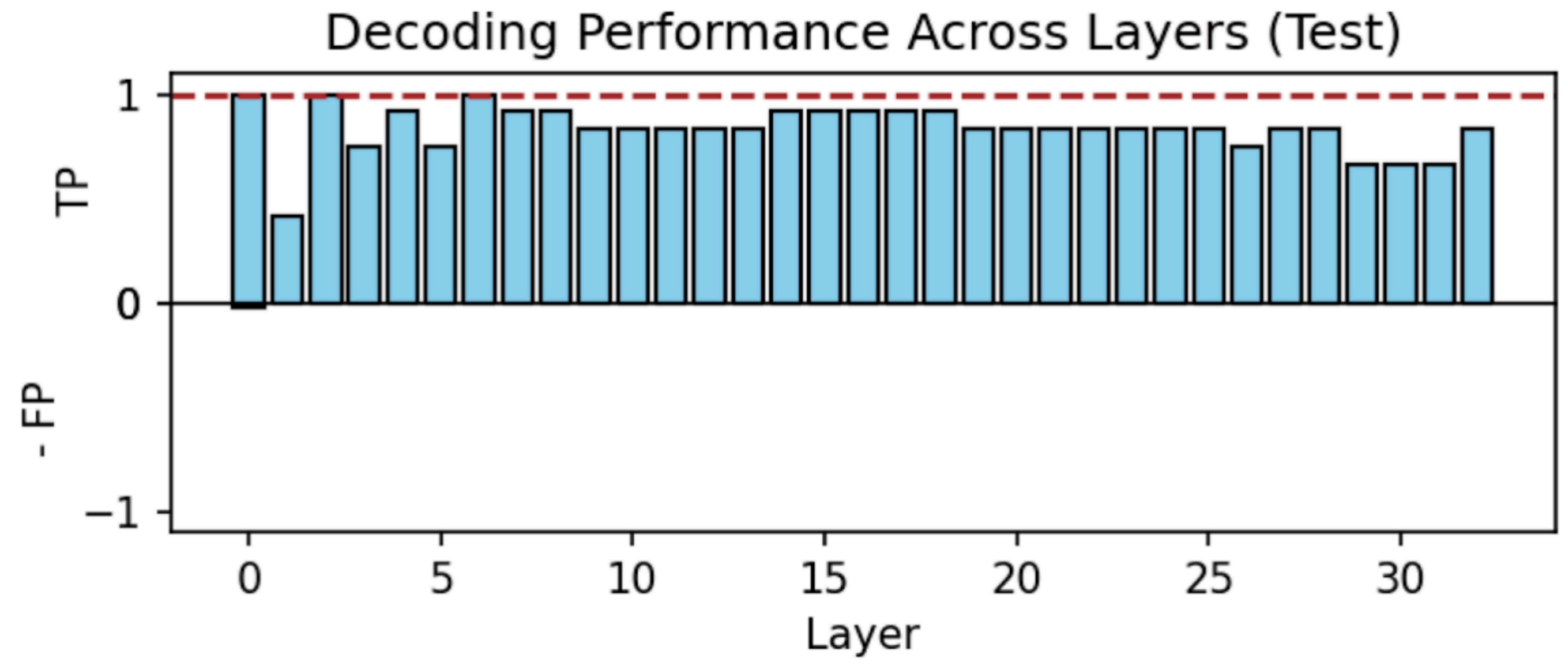
Training: I went to the bakery, bought a chocolate muffin, walked to the nearby park, took a bite, and smiled. ⋯

Test: I went to the coffee shop, ordered a latte, sat down, took a sip, and loved it. ⋯

Control

Training: I took a bite of the burrito I brought from home, then walked into the office and said hi to everyone. ⋯

Test: I microwaved leftovers, ate on the couch, and scrolled through my phone. ⋯



Activating Chunks Controllably Alters LLM's Behavior

"Hello, how are you doing?"

I am a young and passionate man. I am a student of law, but I also love art, literature

Cake

The first thing that comes to mind when I hear the word "cake" is sweetness, but not all cakes are sweet. If you are looking for a sweet, chocolate cake, you have come to the right ···

Cheese

The best part of the cheese is that it can be used ···

Yes, I am talking about cheese. Cheese is one of ···

Grafting Effectiveness

N = 100

Target Concept	Without Grafting	With Grafting
cake	1%	83%
cheese	0%	90%

Grafting chunks is effective to nudge Llama to speak of the targeted concept for many concepts and query types

Table 1: Effect of grafting on TREC categories (percentages).

TREC Category	No Graf.	Early Graf.	Middle Graf.	Late Graf.
ABBR (Abbreviations, acronyms)	14.9%	55.9%	30.8%	18.0%
DESC (Descriptions, definitions)	15.6%	49.0%	28.1%	20.7%
ENTY (Entities)	12.6%	48.1%	22.5%	16.9%
HUM (Human-related)	11.9%	46.7%	21.5%	15.2%
LOC (Locations)	10.7%	47.5%	20.5%	14.4%
NUM (Numeric answers)	11.5%	45.3%	21.8%	16.0%

Summary

- *The Reflection Hypothesis*: neural population activity reflects the regularities in data
- We provide evidence in support of this hypothesis in RNNs and LLMs
- We propose to leverage the chunking tendency in cognition to identify prototypical neural activation as perceptual chunks
- Three complimentary methods to extract chunks from both RNNs and LLMs — DSC, PA, and UCD
- We found chunks activate at the prescence of concrete and abstract concepts
- Activating these concept-encoding chunks, the network starts generating text about that concept
- More results and analysis are in paper: https://arxiv.org/pdf/2505.11576 and project page: https://github.com/swu32/Chunk-Interpretability

Chunking + the structure of naturalistic data = turning the black box transparent

Image generated by ChatGPT