UNIVERSITY OF CAMBRIDGE

Finding Neurons in a Haystack: Case Studies with Sparse Probing

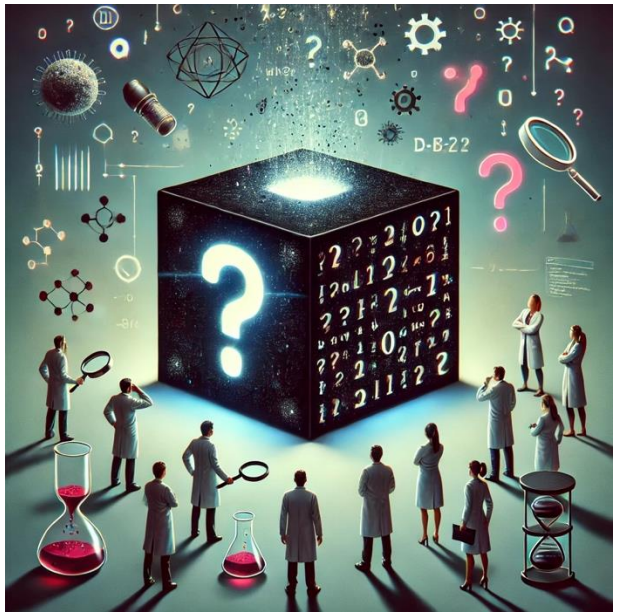
Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, Dimitris Bertsimas

02 June 2023

L193 Paper Presentation - Huadi Wang

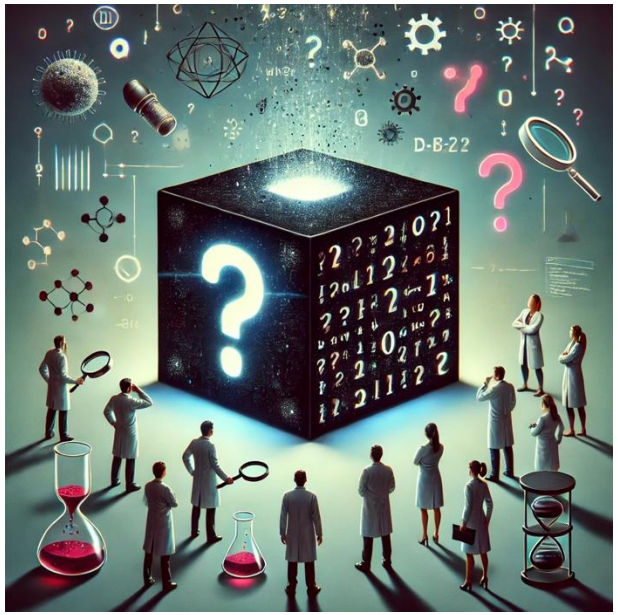
Large Language Model as Blackbox?

The internal computation of LLM remains opaque and poorly understood.



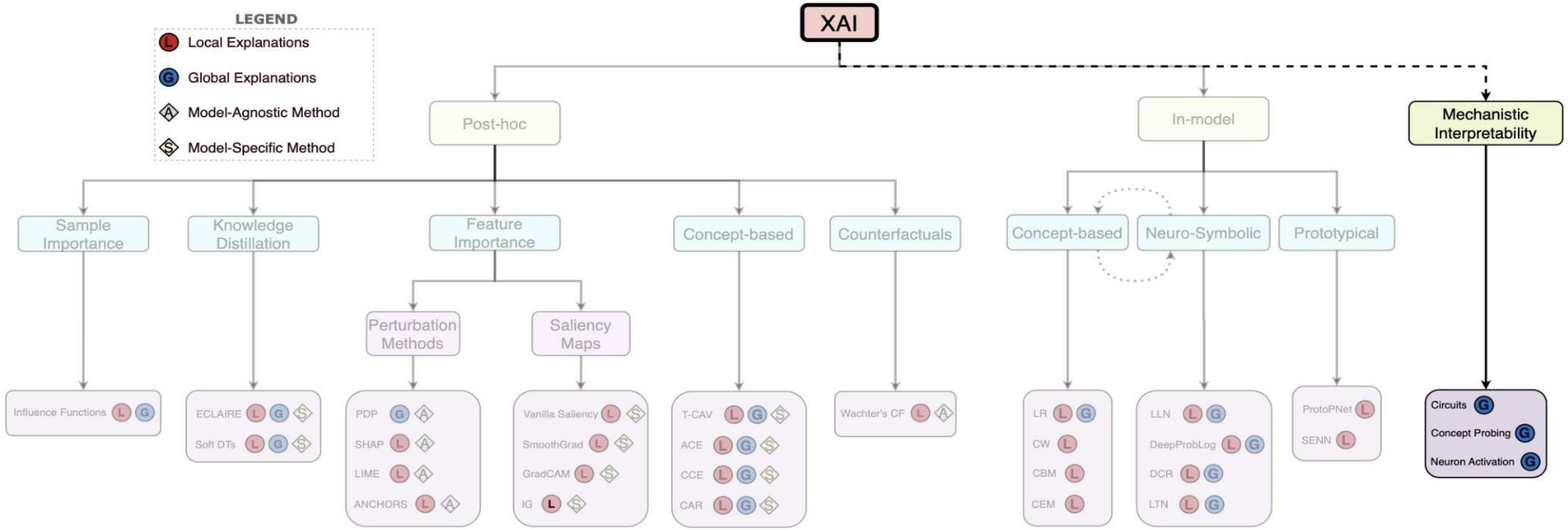
Large Language Model as Blackbox? Haystack!

The internal computation of LLM remains opaque and poorly understood.



Mechanistic Interpretability (MI): Background

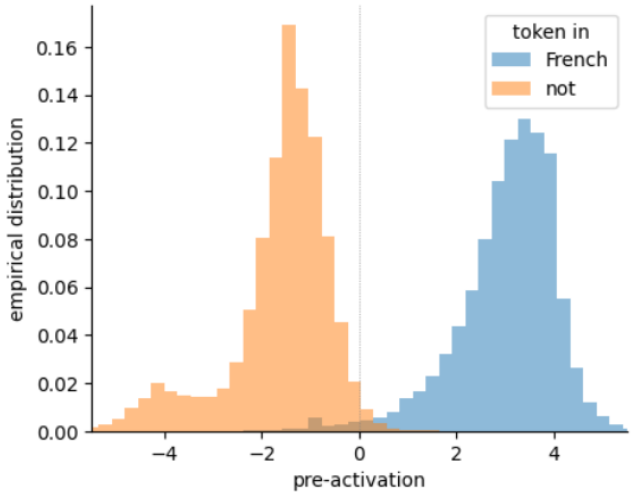
The study of reverse-engineering neural networks to explain the behavior of ML models in terms of their internal components.



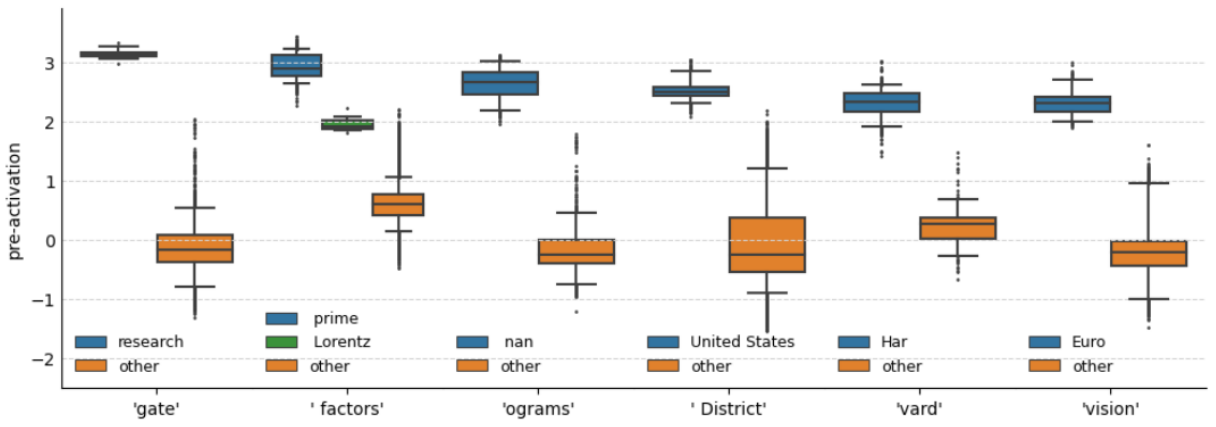
Neuron and Feature Correspondence: Challenge

LLM as **feature extractor — neurons represent different properties from raw input**

- Ideally, **Monosemantic Neuron:** one neuron correspond to one feature
- However, **Superposition:** when represent more features than neurons - **Polysemantic**



(a) A monosemantic French neuron



(b) A polysemantic neuron activating on six unrelated n-grams

Case Studies with Sparse Probing

Understand how high-level human-interpretable features are represented within the internal neuron activation of LLMs

- To what extent does neuron-feature correspondence transfer to full scale LLM?
- What kind of features do or do not appear in superposition?
- How do we reliably find and verify (neuron, feature) pairs in the wild?

Contributing to the full-model interpretability of LLM!

Sparse Probing: Methodology

Probing: train a linear classifier on the internal activation to predict a property of the input

- Tokenized text dataset (activation) + labeled dataset for a subset of tokens (feature)
- Train a binary classifier to minimize classification loss for **each layer** of the network

Sparse Probing: identify **certain neuron(s)** associated with the feature

- **k-sparse probe:** train a classifier with **at most k** non-zero coefficient (neurons)
- Find the top k predictive neuron subset for the classifier: ranking problem
  - Methods: **adaptive thresholding, optimal sparse probing (OSP),** class means of each neuron, mutual info between each neuron and label…
- Particularly well suited to study superposition!

Apply Sparse Probing in LLM

Apply sparse probing **at the MLP layers immediately after elementwise nonlinearity**

- MLP layer perform the majority of feature extractions, form a **privileged basis** [1]

$$a^{(\ell)} = \sigma(W_{fc}^{(l)} \gamma(h_t^{(l-1)}))$$

Cautiously design probe dataset: appropriate positive/negative samples, multi-token property

Experiment Details

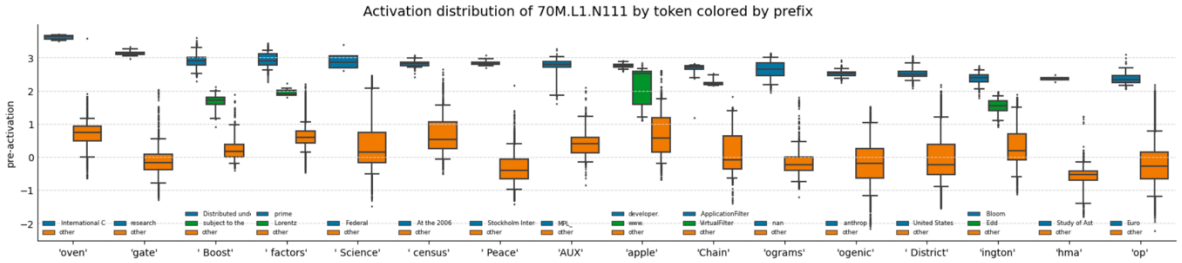
- Model: EleutherAI's Pythia suite [2], 7 models ranging from 70M to 6.9B
- Data: 100 binary features across wide range: language, programming language, part-of-speech…
- Evaluation Metrics: F1 score, Precision, Recall

Case Study 1: Compound Word Neurons

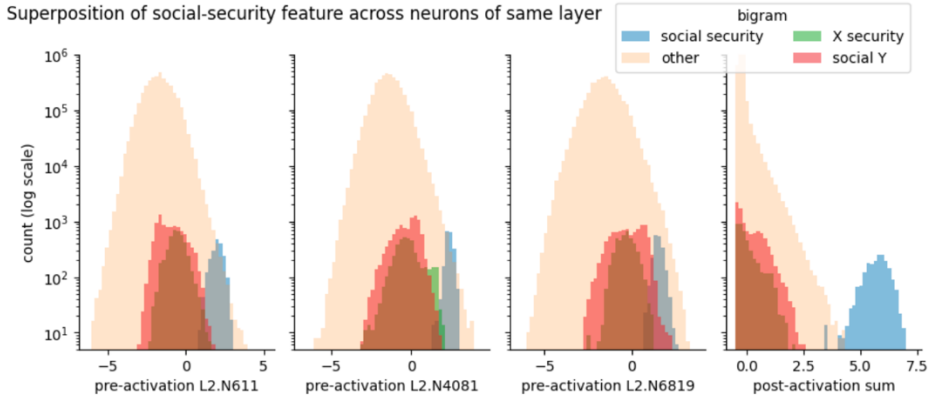
Hypothesis: **early layer neurons** “de-tokenize” raw token into more useful compound abstraction

- Motivation: token vocab is an unnatural way for linguistic processing – e.g. compound word
- Pseudo-vocab (all common n-grams) is pretty large, perfect candidate for **superposition**

Result: polysemantic neurons, but highlight true feature via linear combinations



(a) Activations of a single polysemantic neuron on different tokens when preceded by specific stimuli.



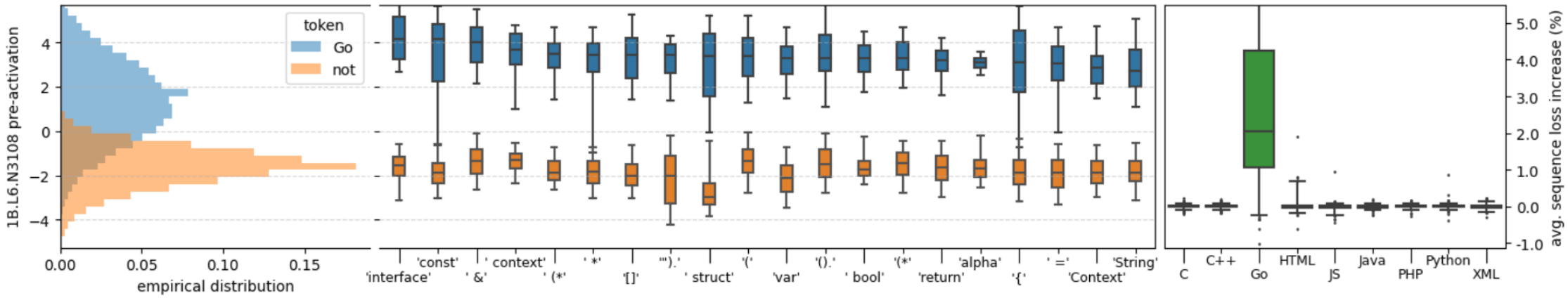
(b) Example of superposition in representing the compound word “social security.”

Case Study 2: Context Neurons

Hypothesis: context feature might be represented **monosemantically**

- High-level descriptions of most tokens (is_french) is important and worth a full neuron
- High-level property may not be mutually exclusive, hard to represent in superposition

Result: highly specialized context neuron in middle layer, appear to be monosemantic



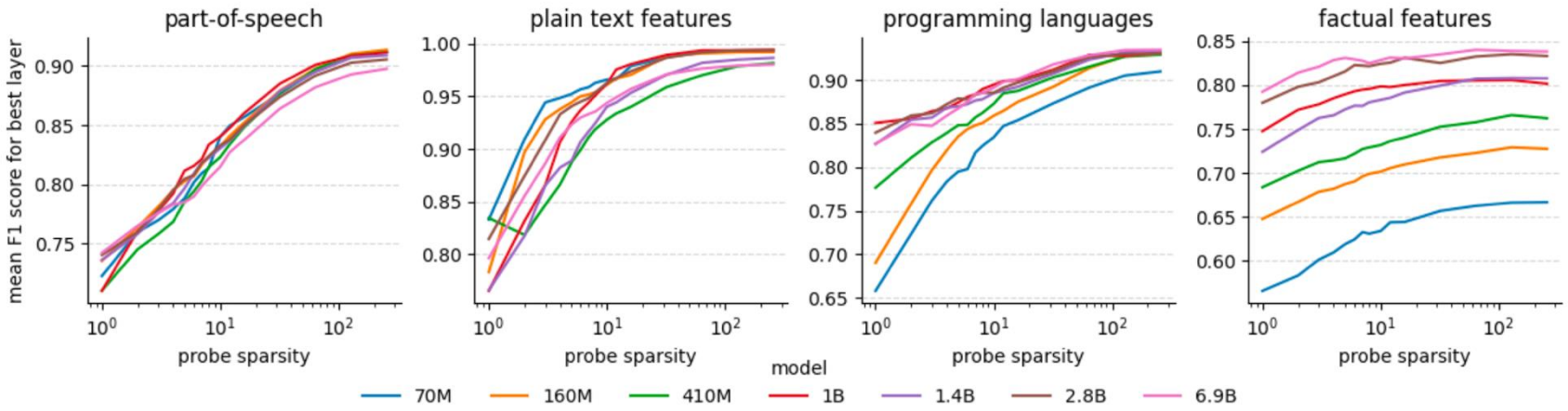
Case Study 3: Effect of Scale

How does sparsity (k) of feature changes with different model scale?

Train a series of probes sweeping the value of k from 256 to 1 using adaptive thresholding

- Report maximum F1 score for each k, each model size, and each feature collection

Result: Two dynamics—quantization model of scaling [3] and neuron splitting [4]

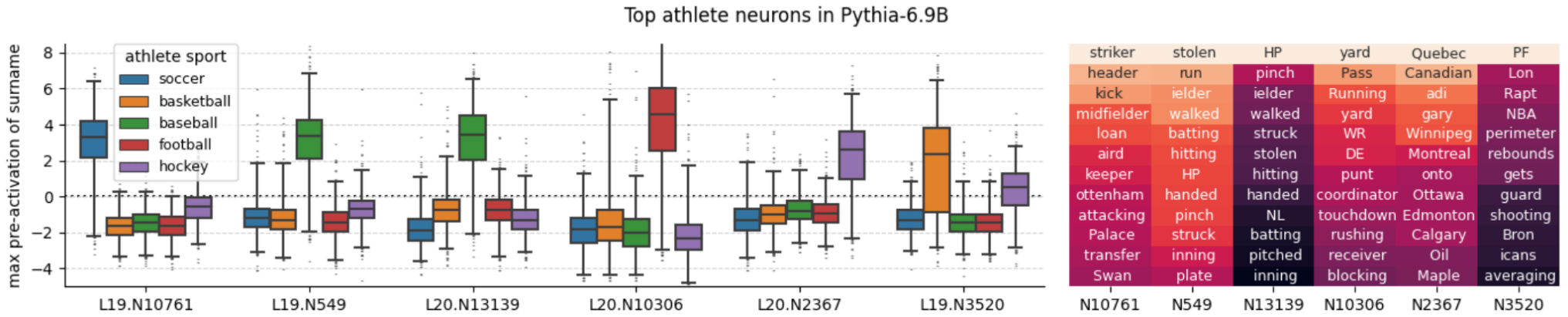


Case Study 4: Interpretation of Classification Performance

1. (layer, feature) pair has low 1-sparse accuracy and high k-sparse accuracy

- Ambiguity between **superposition (intersection)** or **composition (union)** of neurons

2. Further analyze an individual neuron via the **full input** and output logits:



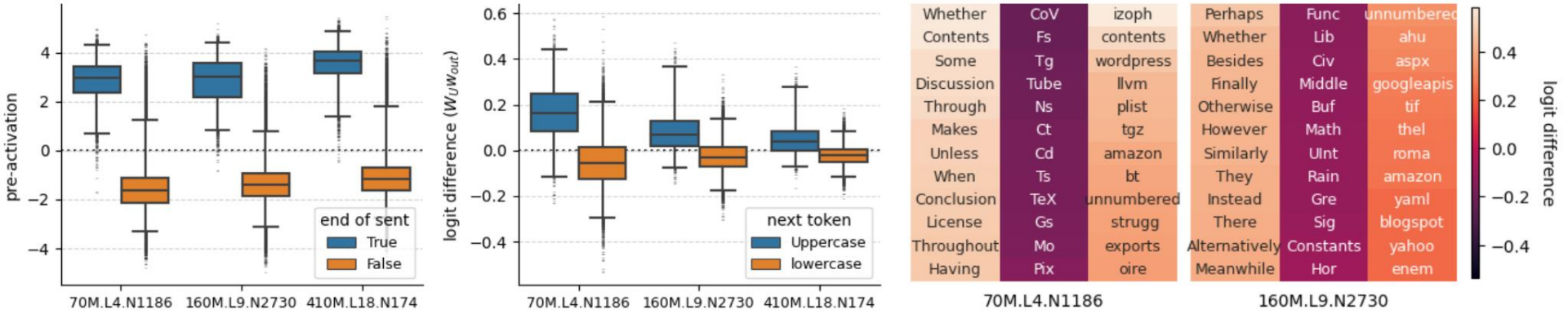
(a) Neurons which activate for the names of specific types of athletes (left) turn out be more general sport neurons when analyzing the top per-token average activations (right).

Case Study 4: Interpretation of Classification Performance

1. (layer, feature) pair has low 1-sparse accuracy and high k-sparse accuracy

- Ambiguity between **superposition (intersection)** or **composition (union)** of neurons

2. Further analyze an individual neuron via the full input and **output logits**:



(b) End of sentence neurons: analyzing the effect on the output vocabulary can corroborate and refine neuron interpretations. For token heatmaps (right), the columns correspond to capital tokens with max increased logits, capital tokens with max decrease, and lowercase tokens with max increase, respectively.

Case Study 4: Interpretation of Classification Performance

1. (layer, feature) pair has low 1-sparse accuracy and high k-sparse accuracy

- Ambiguity between **superposition (intersection)** or **composition (union)** of neurons

2. Further analyze an individual neuron via the full input and output logits

3. Inspect the **precision (false positive)** and **recall (false negative)** of a feature

- High precision, low recall: neuron represents a more specific feature than the feature in probe
- Low precision, high recall: neuron represents a more general feature than the feature in probe

Case Studies with Sparse Probing: Discussion

Strength

- Quickly and precisely localize neurons
- Address drawbacks from previous methods
- Well-suited for studying superposition
  - Clearest evidence of superposition, monosemantic/polysemantic neurons
- Generalize to larger-scale models

Weakness

- Require detailed analysis on probing results
- Highly sensitive to errors in probing dataset
- Cannot explain features/neurons across layers
- Based on empirical findings
- Largest model studied is 6.9B (GPT4: 1.7T [5])

Case Studies with Sparse Probing: Discussion

Future Directions – so many!

- **Areas:** superposition, output prediction, neuron analysis, neuron splitting
- **Applications:** xAI (Mechanistic Interpretability), AI Ethics, AI Safety, specialized LLM

Citation: 140

- Extends sparse probing to analyze **GPT-4** (done by OpenAI obviously) [6]
- **Space and time** features learned by LLM [7]
- Circuit analysis across **multiple layers** [8]

My Final Thought

After “Case Studies with Sparse Probing”, the haystack is...

My Final Thought

After “Case Studies with Sparse Probing”, the haystack is still the haystack!

UNIVERSITY OF CAMBRIDGE

Thanks For Listening!

Reference

[1] N. Elhage, R. Lasenby, and C. Olah, “Privileged bases in the transformer residual stream,” *Transform. Circuits Thread*, p. 24, 2023.

[2] S. Biderman *et al.*, “Pythia: A suite for analyzing large language models across training and scaling,” in *International Conference on Machine Learning*, PMLR, 2023, pp. 2397–2430.

[3] E. Michaud, Z. Liu, U. Girit, and M. Tegmark, “The quantization model of neural scaling,” *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 28699–28722, 2023.

[4] N. Elhage *et al.*, “Softmax linear units. transformer circuits thread.” 2022.

[5] M. Bastian, “GPT-4 has more than a trillion parameters - Report,” THE DECODER. Accessed: Mar. 02, 2025. [Online]. Available: https://the-decoder.com/gpt-4-has-a-trillion-parameters/

[6] L. Gao *et al.*, “Scaling and evaluating sparse autoencoders,” *ArXiv Prepr. ArXiv240604093*, 2024.

[7] W. Gurnee and M. Tegmark, “Language models represent space and time,” *ArXiv Prepr. ArXiv231002207*, 2023.

[8] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso, “Towards automated circuit discovery for mechanistic interpretability,” *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 16318–16352, 2023.