EXPLAINABLE ARTIFICIAL INTELLIGENCE

L193 – Student Presentations 4 – Lent 2025

UNIVERSITY OF CAMBRIDGE

PAPER TIME!

Are you ready for paper tiiiiimeeeeeeee!

PAPER DISTRIBUTION TIME!

Does anyone know who this is?

PAPER DISTRIBUTION TIME!

Does anyone know who this is?

PAPER DISTRIBUTION TIME!

Alan Turing in a running competition (1947)!

PAPER DISTRIBUTION TIME!

In 1947, Alan Turing tried out for the 1948 London Olympics UK marathon team. After contracting fibrositis earlier that year, he still managed to finish 5th in the qualifying race with a time of 2 hours, 46 minutes, and 3 seconds.

How much faster (in minutes) was the time obtained by the 1948 Olympics men's marathon gold medalist?

PAPER DISTRIBUTION TIME!

In 1947, Alan Turing tried out for the 1948 London Olympics UK marathon team. After contracting fibrositis earlier that year, he still managed to finish 5th in the qualifying race with a time of 2 hours, 46 minutes, and 3 seconds.

How much faster (in minutes) was the time obtained by the 1948 Olympics men's marathon gold medalist?

Delfo Cabrera (Argentina)

Answer: 12 minutes, 39 seconds

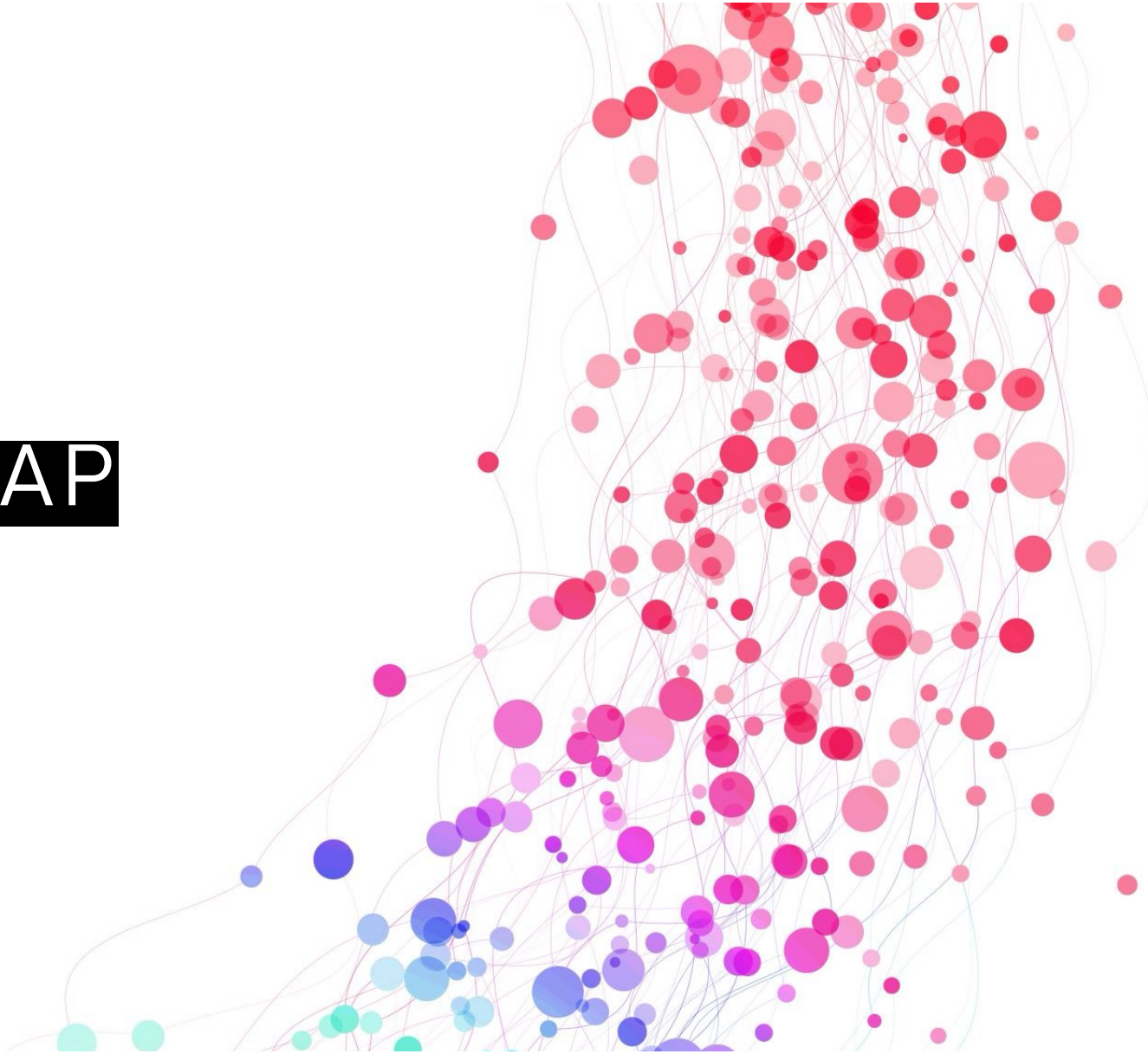
PAPER DISTRIBUTION TIME!

Paper 1: Chen et al. "This looks like that: deep learning for interpretable image recognition." NeurIPS (2019).

Paper 2: Jain et al. "Attention is not explanation." NAACL (2019).

Paper 3: Wiegreffe et al. "Attention is not not explanation." EMNLP (2019).

PAPER RECAP



FINDING NEURONS IN A HAYSTACK (GURNEE ET AL., 2023)

Gurnee et al. "Finding neurons in a haystack: Case studies with sparse probing." TMLR (2023).

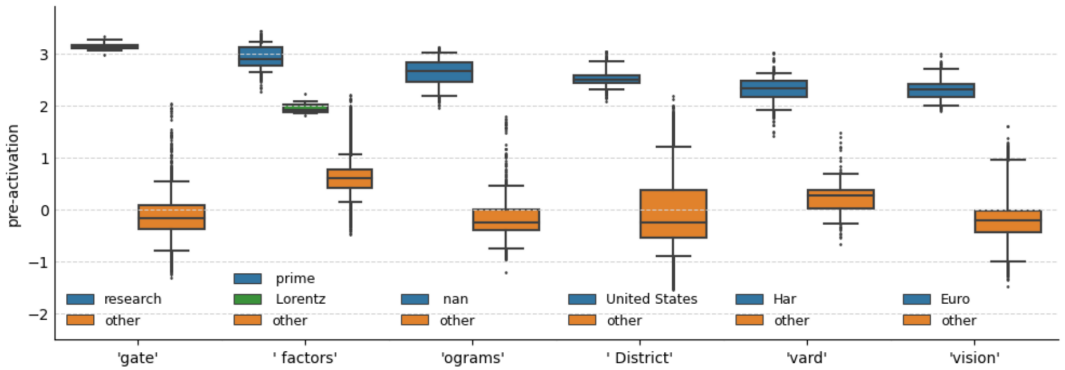
NEURONS IN A HAYSTACK: RESEARCH QUESTION

Main research Question

Are human-interpretable features captured by single neurons in LLMs or are they distributed across their latent space?

NEURONS IN A HAYSTACK : TAKEAWAYS

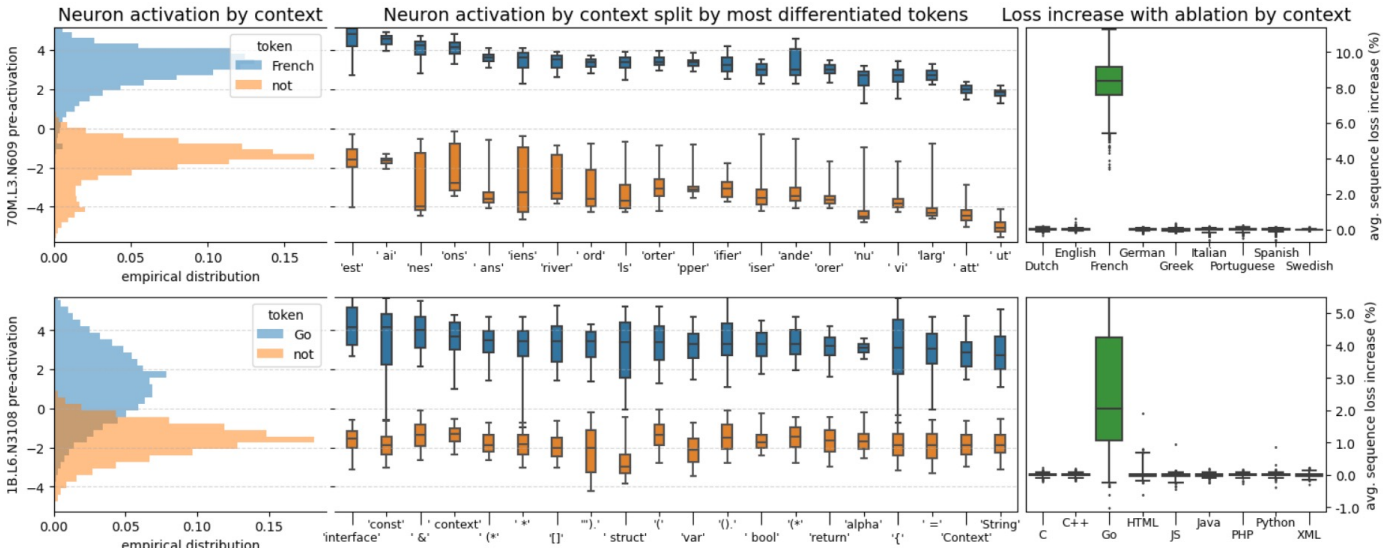
1. Early-layer neurons rarely correspond to single, well-defined "concepts." Instead, they appear to be polysemantic!



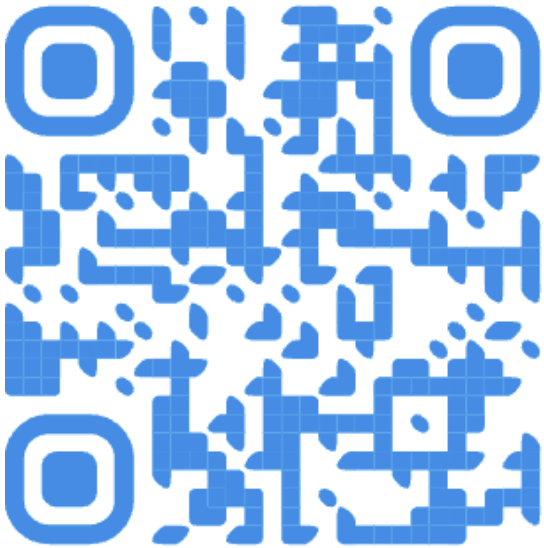
(b) A polysemantic neuron activating on six unrelated n-grams

INTERVENABLE BLACK BOXES: TAKEAWAYS

2. Higher-level contextual concepts (e.g., language detectors) may be encoded by monosemantic neurons in the middle layers.



FINDING CIRCUITS WITH PRUNING (BHASKAR ET AL., 2024)



Bhaskar et al. "Finding transformer circuits with edge pruning." NeurIPS (2024).

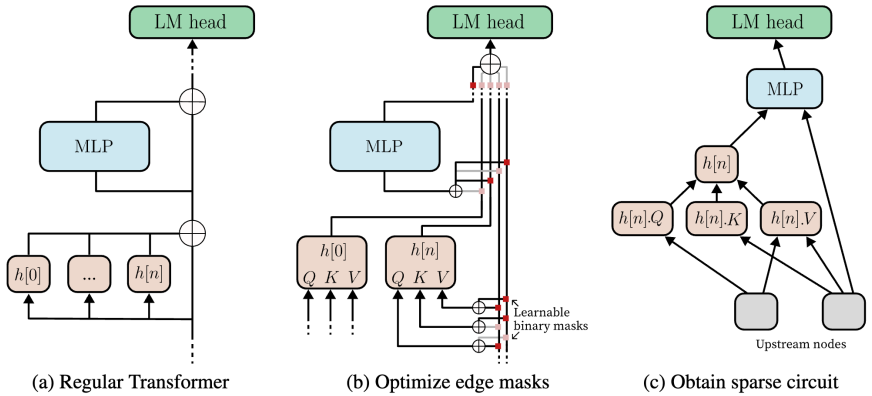
CIRCUITS WITH PRUNING: RESEARCH QUESTION

Main research Question

How can automated circuit discovery be improved to be both efficient and scalable while remaining faithful to the model's predictions?

CIRCUITS WITH PRUNING: TAKEAWAYS

1. Instead of pruning neurons or components, we can speed up circuit discovery by pruning edges and framing circuit discovery as an optimisation problem!



(a) Regular Transformer (b) Optimize edge masks (c) Obtain sparse circuit

$$\arg\min_{\mathcal{C}} \mathbb{E}_{(x,\tilde{x})\in\mathcal{T}} \left[D(p_{\mathcal{G}}(y \mid x) \mid\mid p_{\mathcal{C}}(y \mid x, \tilde{x})) \right], \quad \text{subject to } 1 - |\mathcal{C}|/|\mathcal{G}| \geq c$$

CIRCUITS WITH PRUNING: TAKEAWAYS

2. We can use edge pruning to find extremely sparse circuits in very large LLMs (a few billion parameters) that can faithfully recover/explain most of the model's performance

Table 2: Edge pruning finds circuits with 0.03-0.04% of the edges in CodeLlama-13B that match the performance of the full model. The circuits perform well in cross-evaluation and overlap highly, hinting that the same mechanisms explain large parts of instruction-prompted and few-shot behavior.

Circuit	Num. edges ↓	Accuracy (%) ↑		Exact Match (%) ↑	
		Instr. prompted	Fewshot	Instr. prompted	Fewshot
Full model	3872820	82.00	89.25	100.00	100.00
Instruction prompt (IP)	1041	79.25	74.50	90.00	79.00
Fewshot (FS)	1464	75.75	87.25	84.50	91.25
IP ∩ FS	653	72.50	68.25	79.75	72.50