Using **sparse feature circuits**, we can *discover* and *edit* **unanticipated mechanisms** in language models.

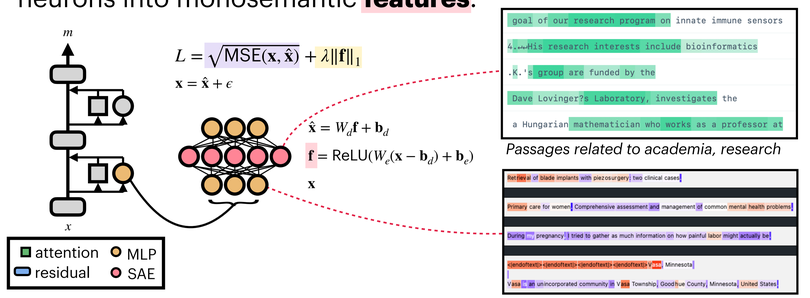
Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models

Samuel Marks[1], Can Rager[2], Eric J. Michaud[3], Yonatan Belinkov[4], David Bau[2], Aaron Mueller[2,4]

1 2 3 4 Northeastern University · TECHNION Israel Institute of Technology

Discovering Sparse Feature Circuits

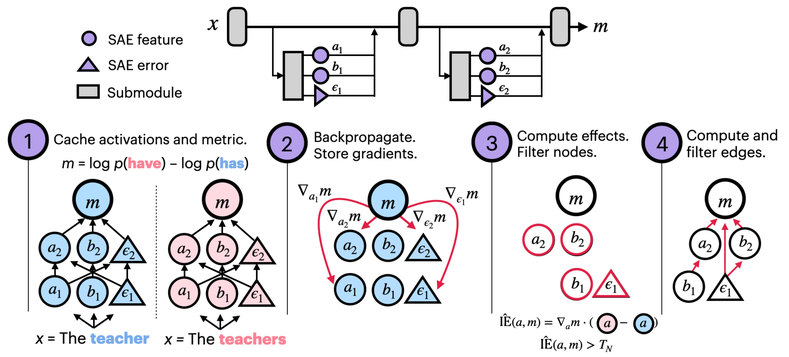
Sparse autoencoders let us decompose polysemantic neurons into monosemantic **features**:



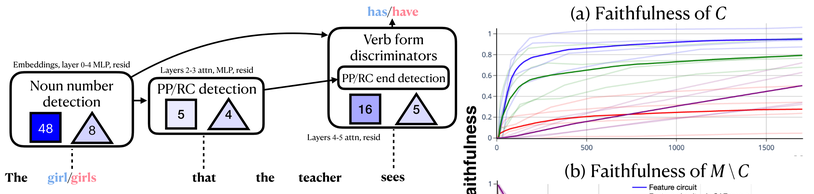
Passages related to academia, research

Sentence boundaries, healthcare, places, i.a.?

We fold SAEs into the forward pass, and attribute graphs of features in performing a behavior:



Case Study: Subject–Verb Agreement



This matches the human intuition!

Feature circuits achieve greater faithfulness f and completeness with fewer nodes than neuron circuits: $f = \frac{m(C) - m(\emptyset)}{m(M) - m(\emptyset)}$

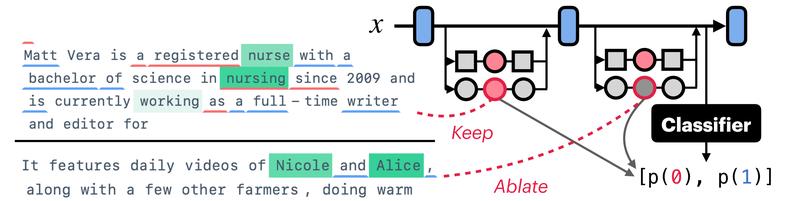
SHIFT: Improving Generalization

1. Train a probe to classify profession on a biased dataset:

"**He** was previously an **assistant professor** at the University of Arizona..." — Man: 0, Professor: 0

"**She** graduated in 2005 with honors, and has 11 years of experience as a **nurse practitioner**" — Woman: 1, Nurse: 1

2. Discover a sparse feature circuit for the classifier
3. Inspect high-effect features; ablate spurious ones



Method	Pythia-70M ↑Profession	Pythia-70M ↓Gender	Pythia-70M ↑Worst group	Gemma-2-2B ↑Profession	Gemma-2-2B ↓Gender	Gemma-2-2B ↑Worst group
Original	61.9	87.4	24.4	67.7	81.9	18.2
CBP	83.3	60.1	67.7	90.2	**50.1**	86.7
Random	61.8	87.5	24.4	67.3	82.3	18.0
SHIFT	88.5	54.0	76.0	76.0	51.5	50.0
SHIFT + retrain	**93.1**	**52.0**	**89.0**	**95.0**	52.4	**92.9**
Neuron skyline	75.5	73.2	41.5	65.1	84.3	5.6
Feature skyline	88.5	54.3	62.9	80.8	53.7	56.7
Oracle	93.0	49.4	91.9	95.0	50.6	93.1

SHIFT outperforms a neuron-based approach that has an unfair advantage.

Our judgments about feature relevance are largely informative.

*We get close to the performance of a classifier trained on **unbiased** data!*

A Fully Unsupervised Interpretability Pipeline

Most interpretability pipelines require us to know which behaviors we're looking for. What about unanticipated behaviors?

Given large corpus of $\{(x_i, y_i)\}$, encode into feature activations $\mathbf{v}(x_i, y_i)$, cluster $\mathbf{v}$, discover sparse feature circuits for each cluster.

Cluster 382: Incrementing sequences

var input = [1, 2, 3, 4, 5, 6, 7, 8

Step 1. Download the latest CompsNY 3.49 Full
Step 2. Double click the Setup file and follow the prompts […]
Step 3. After the main install closes, click OK […]
Step 4

Example features involved:

Succession: Chapter 1, Chapter 2, Chapter 3 | A, B, C | I, II, III, IV

Narrow induction: A3 … A → 3 or III or 4 … | A7 … A → 7 or VII or 8 …

Cluster 475: "to" as infinitive object

At issue, whether the defendant should be allowed to

British Prime Min David Cameron says in televised remarks he would like Britain to

Reader bloggers are asked to

Example features involved:

Objects which can precede object complements: Direct the user to | It's up to you to

Other words which precede infinitive objects: According to | This infection leads to