Scalable circuit discovery: as accurate as intervention methods, at a fraction of the cost.

CircuitLasso: Scalable Circuit Learning for Interpreting Large Language Models

Naiyu Yin, Dennis Wei, Tian Gao, Amit Dhurandhar, Karthikeyan Natesan Ramamurthy, Yue Yu

1 TL;DR

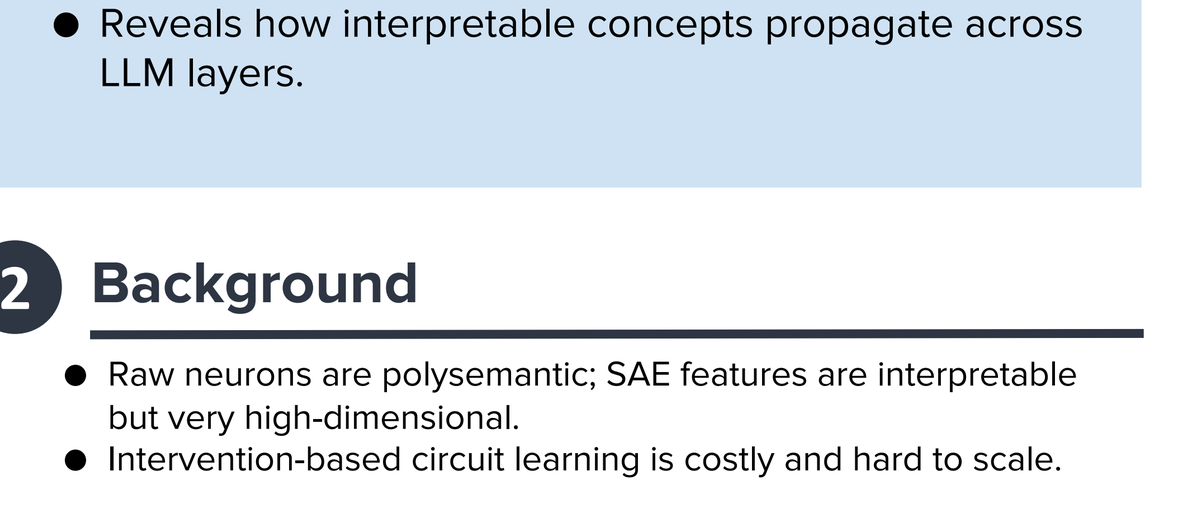
- Recasts circuit discovery as sparse (Lasso) regression on observational activations — no interventions.
- Matches SOTA accuracy at **2–3× lower cost**; scales to high-dimensional SAE features.
- Reveals how interpretable concepts propagate across LLM layers.

2 Background

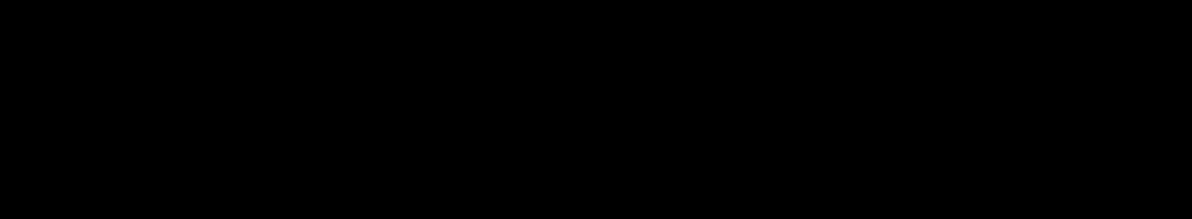
- Raw neurons are polysemantic; SAE features are interpretable but very high-dimensional.
- Intervention-based circuit learning is costly and hard to scale.

3 Method

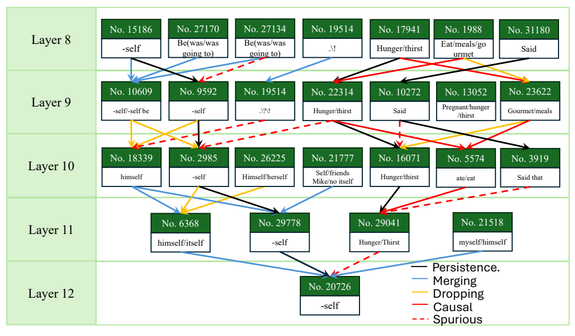
- Models activations as a sparse linear SEM, solved with an L1 (Lasso) penalty.
- Block upper-triangular structure follows the model's computation order — acyclic by construction.
- Reduces to scalable sparse regression on neurons or SAE features.



4 Results



- Comparable accuracy to state-of-the-art baselines (EAP-ig, EAP).
- **3.0× faster** than EAP-ig, **2.1× faster** than EAP.



- Scales to SAE features (GPT-2 small).
- Recovers interpretable concept circuits.

Method	Pythia-70M # of features	Pythia-70M Runtime (s) ↓	Gemma-2-2b # of features	Gemma-2-2b Runtime (s) ↓	Gemma-2-9b # of features	Gemma-2-9b Runtime (s) ↓
SHIFT	49	257.6	65	371.2	71	908.4
SHIFT-retrain		356.3		476.8		1056.0
CircuitLasso	41	36.5	55	47.2	59	107.4
CircuitLasso-retrained		61.9 (17.37%↑)		72.5 (15.20%↑)		125.2 (11.98% ↑)

Method	Pythia-70M	Gemma-2-2B	Gemma-2-9B
ORIGINAL	61.9	69.6	70.8
CBP	83.3	90.1	94.7
ORACLE	93.0	95.1	95.7
LINEAR PROBING	84.0	90.5	95.0
SHIFT	88.5	72.8	77.1
SHIFT-retrain	93.1	94.2	96.0
CircuitLasso	90.5	77.5	81.5
CircuitLasso-retrain	**94.2**	**95.1**	**96.9**

- Matches SHIFT debiasing at lower cost.

5 Discussion & Limitations

- **Discussion:** CircuitLasso scales circuit discovery to SAE features via sparse regression — SOTA accuracy at a fraction of the cost.
- Complementary to per-prompt attribution; enables model editing & domain generalization.
- **Limitations:** linear surrogate; some recovered paths can appear anti-causal.
- Nonlinear variant gives similar structure at higher cost.

PAPER

Naiyu Yin: nay224@lehigh.edu