Query Circuits: Explaining How Language Models Answer User Prompts

Paper Code Project Page

Tung-Yu Wu[1], Fazl Barez[1,2]
[1]University of Oxford, [2]Martian

TECHNICAL SAFETY & GOVERNANCE LAB

UNIVERSITY OF OXFORD

Motivation

CLT-based circuit discovery dominates

- Current query-level circuit discovery methods all build on surrogates like CLTs.
- CLT-based circuits place all nodes and edges on CLTs instead of the original model.
- Question: Can we identify query-level circuits that reside directly within the original model?

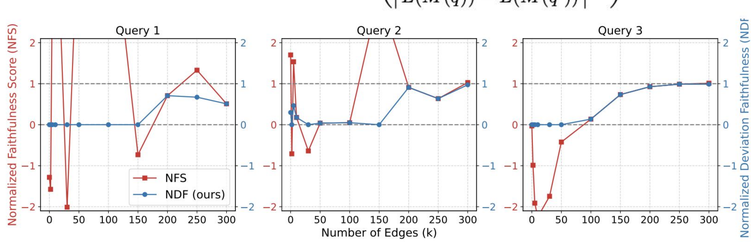
Our Goal

Query-level circuit discovery within the model

- Explore the possibilities of query-level circuits within the original model.
- We argue that gradient-attribution with SAEs can be a decent alternative to CLTs.

Method 1 - NDF Score

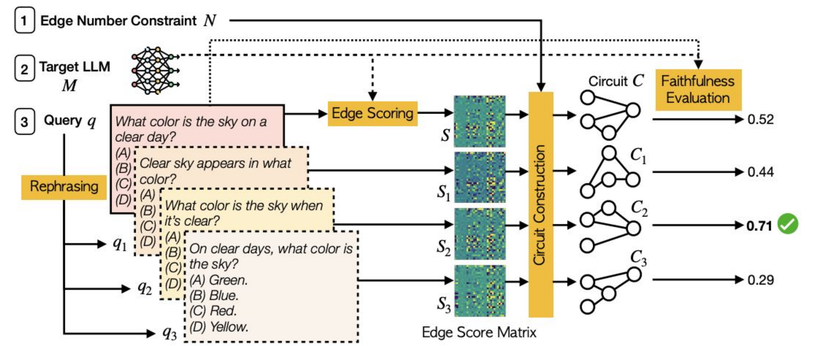
The original NFS score to measure the circuit faithfulness is unreliable. We proposed the NDF as an alternative: $NDF(C_q) = 1 - \min\left(\left|\frac{L(M(q)) - L(C_q(q))}{L(M(q)) - L(M(q'))}\right|, 1\right)$



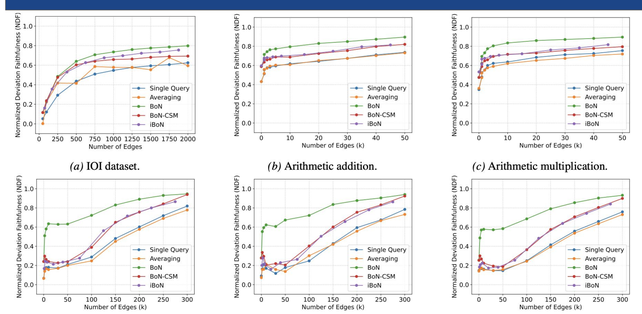
Method 2 - BoN on EAP-IG

BoN sampling via query paraphrasing and integrated gradients for edge selection can discover faithful query-level circuits within the original model.

$$a_e = (e - e')^\top \int_0^1 \nabla_e M(z' + \alpha(z - z'))\, d\alpha$$
$$\approx (e - e')^\top \frac{1}{m} \sum_{k=1}^{m} \nabla_e M\left(z' + \frac{k}{m}(z - z')\right)$$



Results



(a) IOI dataset. *(b)* Arithmetic addition. *(c)* Arithmetic multiplication.
(d) MMLU Marketing. *(e)* MMLU Astronomy. *(f)* ARC Challenge.

High-NDF query circuits reflect more actual model computations. Deactivating gender-related SAE features in the high-NDF circuits reduces the model's gender bias significantly more than deactivating them in the low-NDF ones.

Metric	Scale	Circuit	Mean ± Std	W	p-value	Rosenthal's r
Absolute Bias Reduction	Logit	Best	0.810 ± 0.581	16.0	<0.0001	0.787
		Worst	0.234 ± 0.278			
		Δ Mean = +0.576	(****)			
	Probability	Best	0.063 ± 0.057	41.0	<0.0001	0.737
		Worst	0.011 ± 0.020			
		Δ Mean = +0.052	(****)			
Avg. Bias Reduction per Gender Feature	Logit	Best	0.073 ± 0.055	11.0	<0.0001	0.836
		Worst	0.014 ± 0.017			
		Δ Mean = +0.059	(****)			
	Probability	Best	0.006 ± 0.005	21.0	<0.0001	0.803
		Worst	0.001 ± 0.001			
		Δ Mean = +0.005	(****)			

An example of a query circuit with SAEs