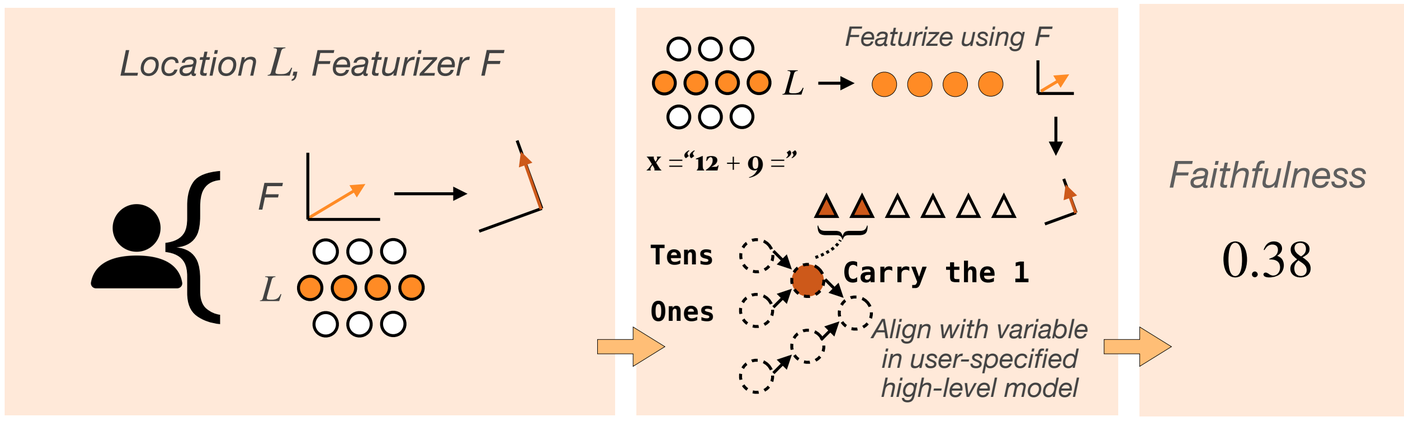
MIB

A MECHANISTIC INTERPRETABILITY BENCHMARK

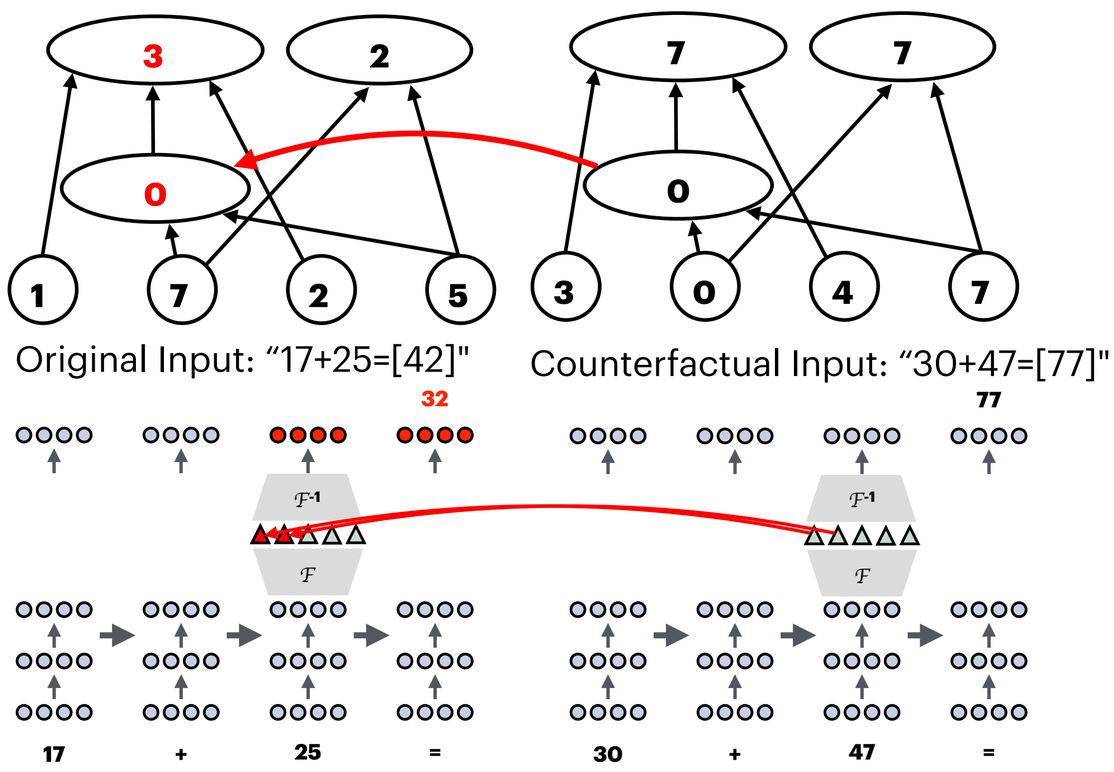
Aaron Mueller*[1], Atticus Geiger*[2], Sarah Wiegreffe[3], Dana Arad[4], Iván Arcuschin[5], Adam Belfki[6], Yik Siu Chan[7], Jaden Fiotto-Kaufman[6], Tal Haklay[4], Michael Hanna[8], Jing Huang[9], Rohan Gupta[5], Yaniv Nikankin[4], Hadas Orgad[4], Nikhil Prakash[6], Anja Reusch[4], Aruna Sankaranarayanan[10], Shun Shao[11], Alessandro Stolfo[12], Martin Tutek[4], Amir Zur[2], David Bau[6], Yonatan Belinkov[4]

Causal Variable Localization Track



Faithfulness: do interventions to the variable cause the model's behavior to change *in the expected way*?

Testing via Interchange Interventions: fixing the "carry-the-one" variable



Mean (**Best**) Interchange Intervention Accuracy for 2 Different Causal Variables across Layers

	ARC (Easy)			
	Gemma-2		Llama-3.1	
Method	O_{Answer}	X_{Order}	O_{Answer}	X_{Order}
DAS	88 (**94**)	76 (**88**)	88 (**99**)	74 (**84**)
DBM	82 (**99**)	63 (**80**)	85 (**100**)	69 (**82**)
+PCA	78 (**98**)	64 (**81**)	84 (**100**)	72 (**83**)
+SAE	70 (**89**)	54 (**70**)	74 (**94**)	55 (**67**)
Full Vector	63 (**100**)	43 (**74**)	68 (**100**)	47 (**72**)

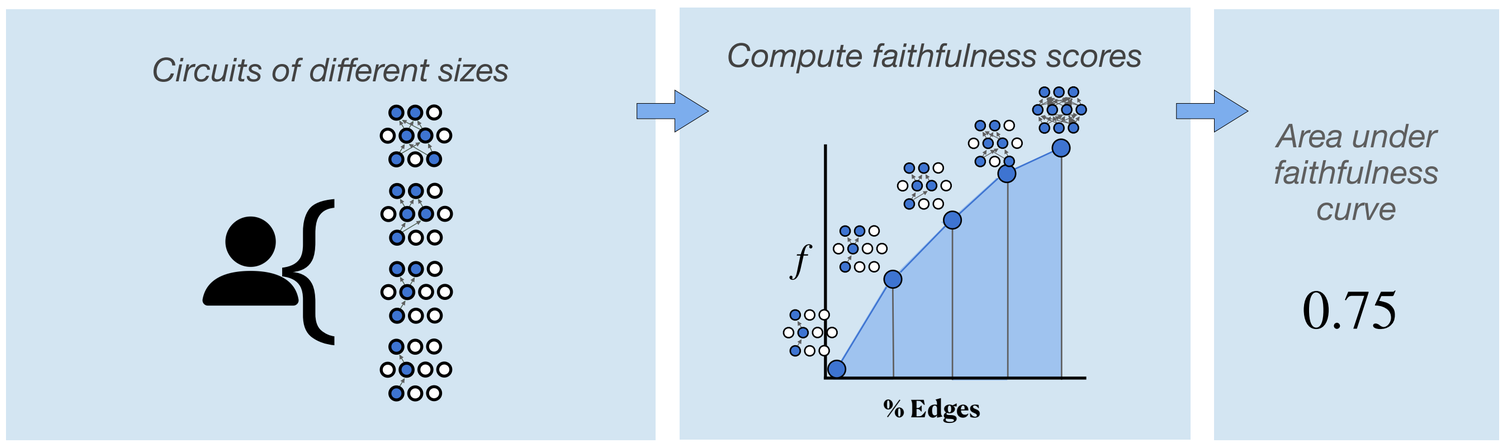
	MCQA					
	Gemma-2		Llama-3.1		Qwen-2.5	
Method	O_{Answer}	X_{Order}	O_{Answer}	X_{Order}	O_{Answer}	X_{Order}
DAS	95 (**97**)	77 (**93**)	94 (**100**)	77 (**91**)	86 (**95**)	78 (**100**)
DBM	84 (**99**)	63 (**84**)	86 (**100**)	66 (**73**)	46 (**94**)	60 (**99**)
+PCA	57 (**96**)	52 (**81**)	65 (**99**)	53 (**74**)	22 (**76**)	54 (**100**)
+SAE	73 (**90**)	51 (**65**)	80 (**99**)	58 (**65**)	–	–
Full Vector	61 (**100**)	44 (**77**)	77 (**100**)	46 (**68**)	35 (**99**)	49 (**99**)

Insights:

Supervised methods >>> unsupervised.

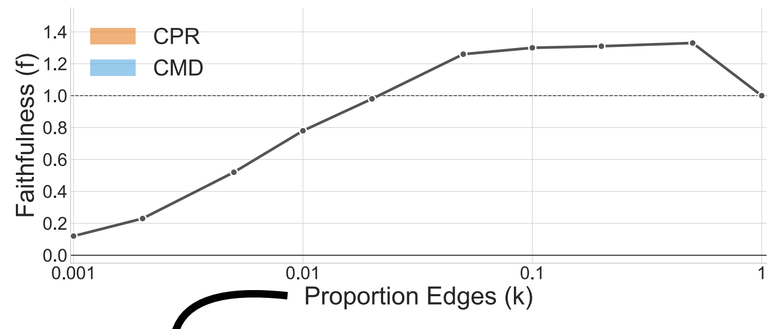
Non-basis-aligned subspaces > basis-aligned subspaces.

Circuit Localization Track



Faithfulness: does the circuit capture the model's task behavior?

Minimality: does the circuit contain as few components as is necessary?



Weighted edge count: a way to directly compare the size of neuron- and edge-based circuits

$$|C| = \sum_{(u,v)\in C} \frac{|N_u \cap N_C|}{|N_u|}$$

$|N_u \cap N_C|$ ← # neurons in circuit

$|N_u|$ ← # neurons in node

CMD scores (lower is better) for all tasks except InterpBench (AUROC; higher is better)

		IOI					Arithmetic	MCQA			ARC (E)		ARC (C)
Method (Ablation)	InterpBench (↑)	GPT-2	Qwen-2.5	Gemma-2	Llama-3.1		Llama-3.1	Qwen-2.5	Gemma-2	Llama-3.1	Gemma-2	Llama-3.1	Llama-3.1
Random	0.44	0.75	0.72	0.69	0.74	0.75	0.73	0.68	0.74	0.68	0.74	0.74	
EActP (CF)	0.28	**0.02**	0.49	-	-	-	0.36	-	-	-	-	-	
EAP (mean)	0.78	0.29	0.18	0.25	0.04	0.07	0.21	0.20	0.16	0.22	0.28	0.20	
EAP (CF)	0.73	0.03	0.15	0.06	**0.01**	0.01	0.07	0.08	**0.09**	**0.04**	**0.11**	**0.18**	
EAP (OA)	0.77	0.30	0.16	-	-	-	0.11	-	-	-	-	-	
EAP-IG-inp. (CF)	0.71	0.03	0.02	0.04	**0.01**	**0.00**	0.08	**0.06**	0.14	**0.04**	**0.11**	0.22	
EAP-IG-act. (CF)	**0.81**	0.03	**0.01**	**0.03**	**0.01**	**0.00**	**0.05**	0.07	0.13	**0.04**	0.30	0.37	
NAP (CF)	0.30	0.38	0.33	0.37	0.29	0.28	0.30	0.35	0.32	0.33	0.69	0.69	
NAP-IG (CF)	0.62	0.27	0.20	0.26	0.19	0.18	0.18	0.29	0.33	0.28	0.67	0.67	
IFR	0.71	0.42	0.69	0.75	0.83	0.22	0.60	0.62	0.48	0.66	0.64	0.76	
UGS	0.74	0.03	0.03	-	-	-	0.20	-	-	-	-	-	

Insights:

Counterfactual (CF) ablations > mean or optimal ablations.

Edge-based circuits > node-based circuits.

Attribution with integrated gradients is generally best.

1 Boston University
2
3 Ai2
4 Technion Israel Institute of Technology
5 Independent
6 Northeastern University
7 Brown
8 University of Amsterdam
9
10 MIT
11 University of Cambridge
12 ETH zürich

Funded by the European Union (ERC, Control-LM,101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Website | Code | Leaderboard | Data