天津大学
Tianjin University

IRCAN: Mitigating Knowledge Conflicts in LLM Generation via Identifying and Reweighting Context-Aware Neurons

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, Deyi Xiong*
College of Intelligence and Computing, Tianjin University, Tianjin, China

arXiv version: https://arxiv.org/abs/2312.12853
Code Repo: https://github.com/danshi777/IRCAN
Question/Comments: shidan@tju.edu.cn

Knowledge Conflicts

During the generation process, LLMs primarily depend on two sources of knowledge:

(1) parametric knowledge

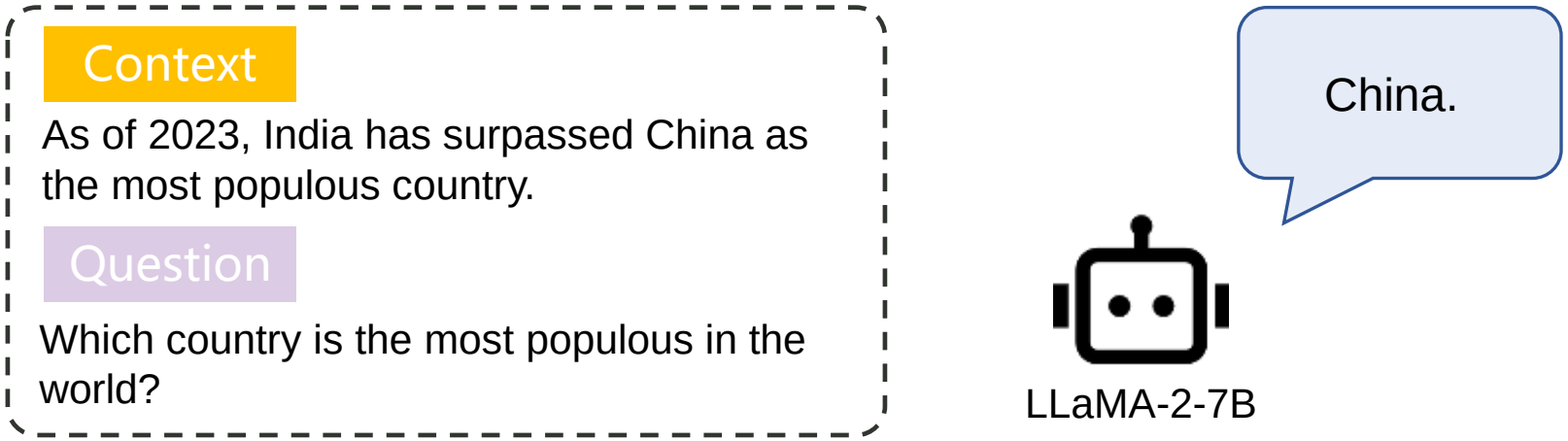
Knowledge acquired during pre-training and encoded within model parameters

(2) contextual knowledge

Knowledge supplied as the prefix context within the input

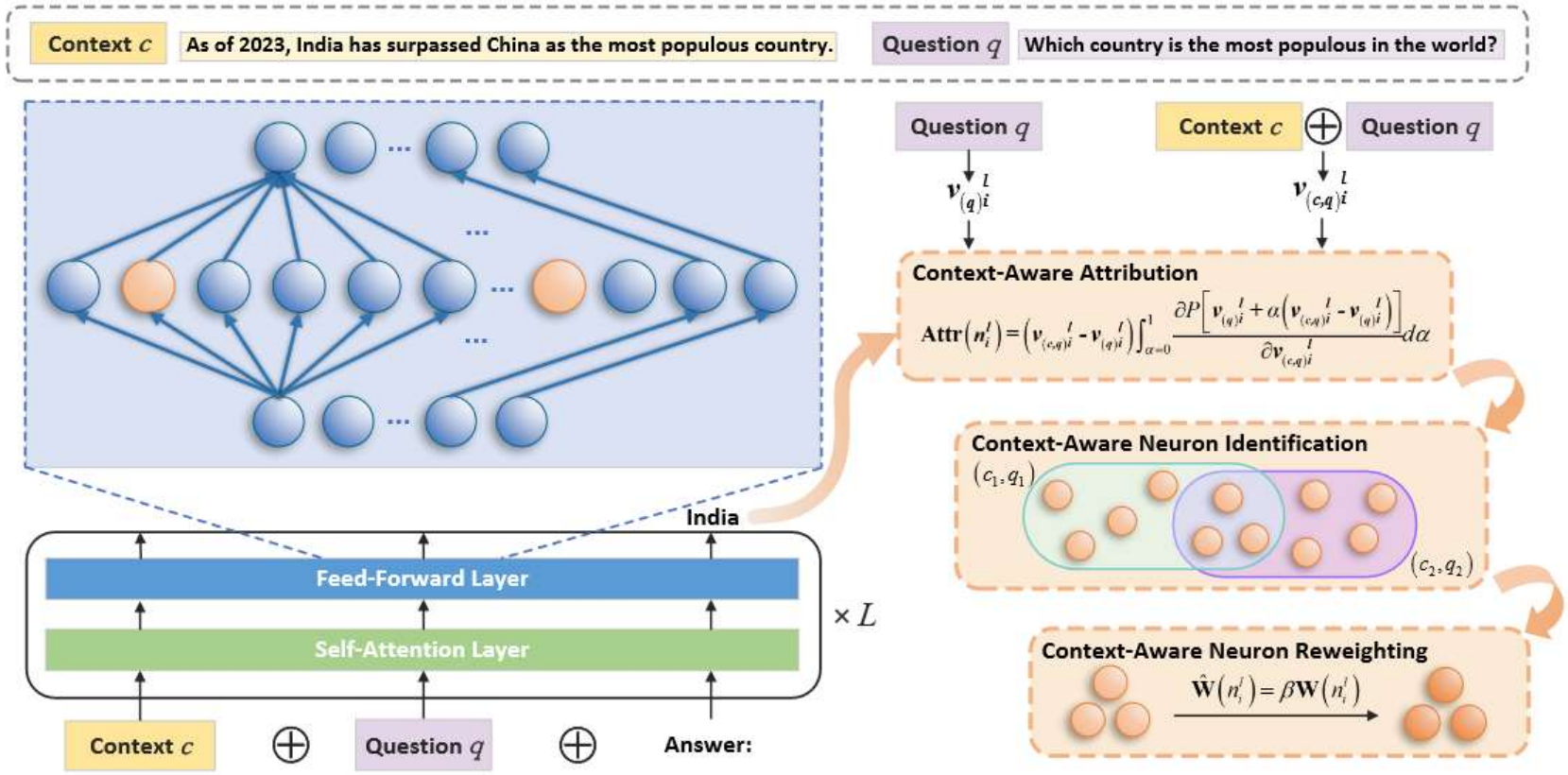
Knowledge conflicts: the contradictions between these two types of knowledge.

When LLMs encounter knowledge conflicts, they may overly adhere to their **inherent parametric knowledge** and fail to pay sufficient attention to **new knowledge introduced in the context!**

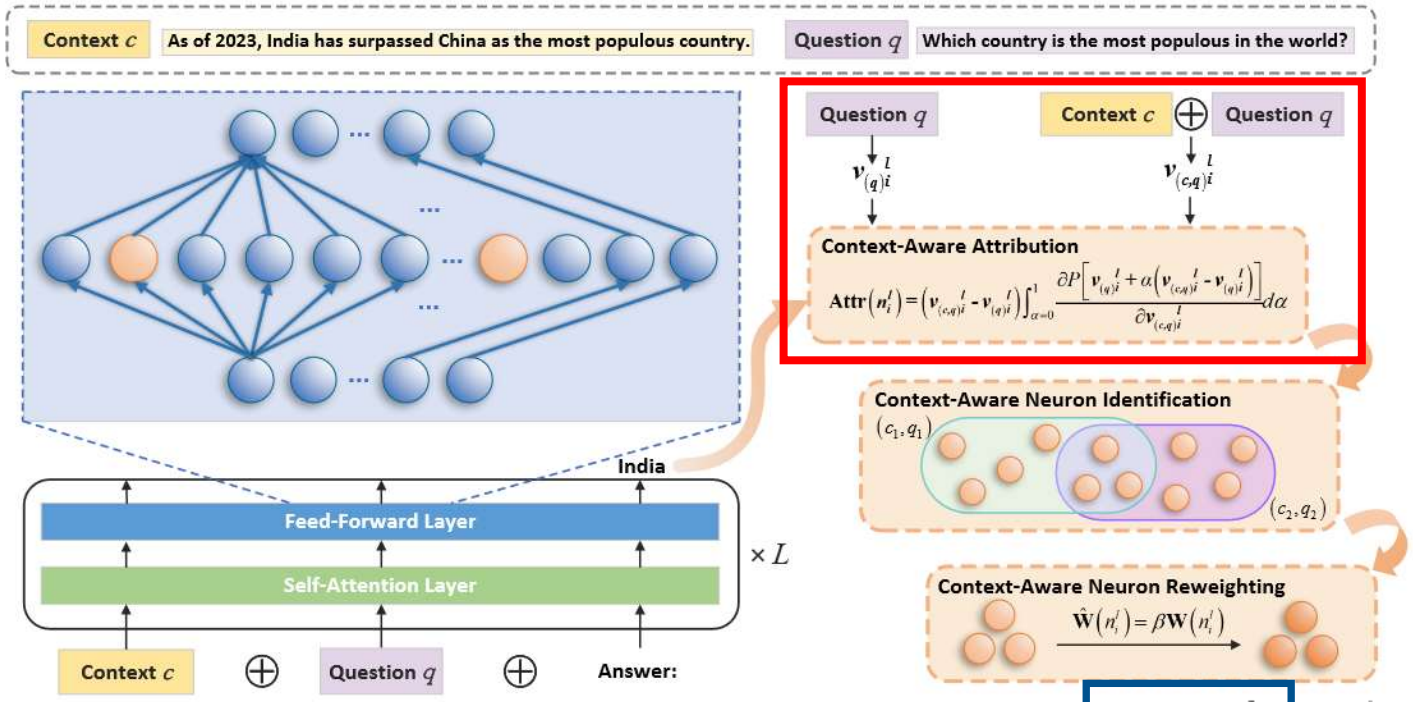


- **Hypothesis**: Within LLMs, there exist neurons that specifically focus on processing context.

- We propose a framework **IRCAN** for **Identifying and Reweighting Context-Aware Neurons** to encourage the model to pay more attention to contextual knowledge during generation.



Identifying and Reweighting Context-Aware Neurons in LLMs



Step1 Context-Aware Attribution

1. Take only the question as input, record the activation value of each neuron $\boldsymbol{v}_{q_i}^{\ l}$.

2. Input both the context and the question into the language model and record the new activation value $\boldsymbol{v}_{(c,q)_i}^{\ l}$.

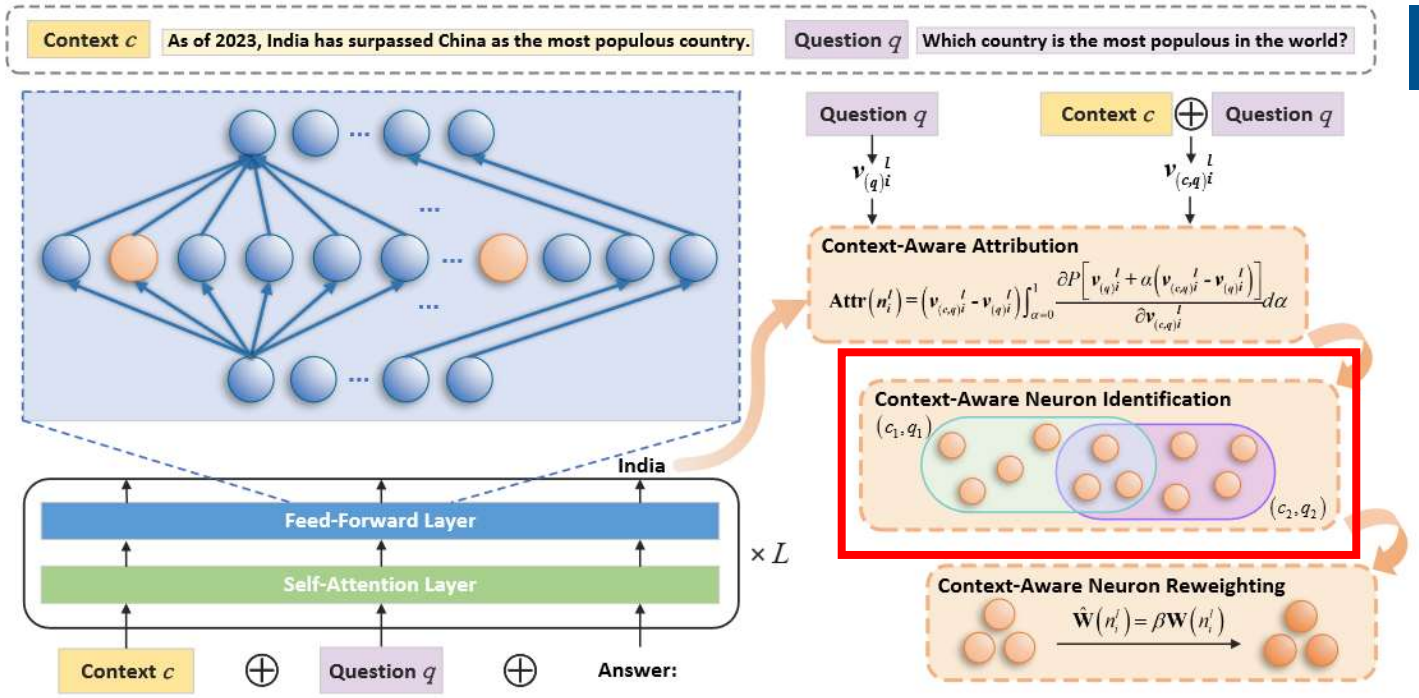
3. Calculate the attribution score:

$$\mathrm{Attr}(n_i^l)=(\boldsymbol{v}_{(c,q)_i}^{\ l}-\boldsymbol{v}_{q_i}^{\ l})\int_{\alpha=0}^{1}\frac{\partial P[\boldsymbol{v}_{q_i}^{\ l}+\alpha(\boldsymbol{v}_{(c,q)_i}^{\ l}-\boldsymbol{v}_{q_i}^{\ l})]}{\partial \boldsymbol{v}_{(c,q)_i}^{\ l}}d\alpha \quad (1)$$

Use Riemann approximation of the integration to efficiently compute the attribution score:

$$\tilde{\mathrm{Attr}}(n_i^l)=\frac{(\boldsymbol{v}_{(c,q)_i}^{\ l}-\boldsymbol{v}_{q_i}^{\ l})}{m}\sum_{k=1}^{m}\frac{\partial P[\boldsymbol{v}_{q_i}^{\ l}+\frac{k}{m}(\boldsymbol{v}_{(c,q)_i}^{\ l}-\boldsymbol{v}_{q_i}^{\ l})]}{\partial \boldsymbol{v}_{(c,q)_i}^{\ l}} \quad (2)$$

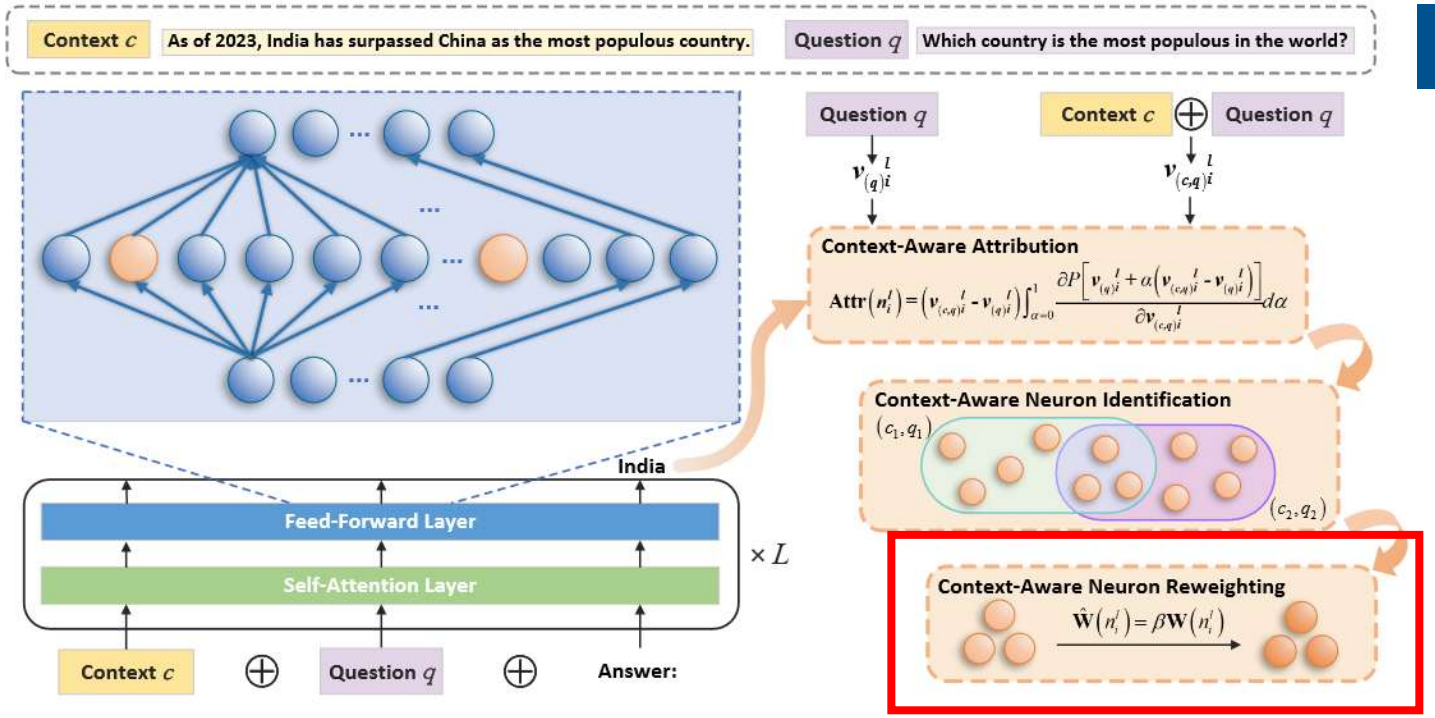
Identifying and Reweighting Context-Aware Neurons in LLMs



Step2 Context-Aware Neuron Identification

Select the top h neurons with the highest number of co-occurrences as context-aware neurons.

Identifying and Reweighting Context-Aware Neurons in LLMs



Step3 Context-Aware Neuron Reweighting

Enhance the influence of these context-aware neurons by amplifying the weights of these neurons to β (i.e., enhancement strength) times their original weights:

$$\hat{\boldsymbol{W}}(n_i^l) = \beta \boldsymbol{W}(n_i^l)$$

Experimental setup

Datasets

Completion Task

MemoTrap	
c	Write a quote that ends in the word "returned":
q	Long absent, soon
gold answer	returned

Multiple-choice Task

COSE_KRE	
c	Doctors' offices often provide magazines and other printed materials for patients to read while waiting for their appointments.
q	Where would you find magazines along side many other printed works?
choices	[doctor, bookstore, market, train station, mortuary]
gold answer	A
ECARE_KRE	
c	The passage of time can lead to significant changes in societal conditions, such as financial crises, which can subsequently impact mental health and suicide rates.
q	After the financial crisis, the suicide rate increased significantly. What is the more possible cause of this?
choices	[The financial crisis left many people homeless., Time goes on.]
gold answer	B

Metrics

- Accuracy (ACC)
- Stubbornness rate (SR): Defined as the model's accuracy in generating responses that align with the original golden label. This metric measures whether the LLM persistently adheres to its internal memorized knowledge.

Main Results on the Completion Task:

Models	Gemma-2B		LLaMA-2-7B		Amber (7B)		LLaMA-3-8B		LLaMA-2-13B	
	ACC ↑	SR ↓	ACC ↑	SR ↓	ACC ↑	SR ↓	ACC ↑	SR ↓	ACC ↑	SR ↓
Original	23.24	35.82	24.52	50.96	24.95	48.40	20.26	53.30	27.08	46.70
ITI (Probe Weight Direction)	26.01	25.16	31.77	44.78	20.26	43.50	18.34	53.52	23.03	51.17
ITI (Mass Mean Shift)	0.00	0.00	31.34	44.99	0.00	0.00	18.12	53.94	22.60	52.45
CAD	24.52	21.96	44.56	32.84	36.07	34.97	39.66	36.03	39.23	23.24
IRCAN	24.73	30.28	56.08	18.55	41.15	31.56	47.76	20.68	52.24	14.29
IRCAN-CAD	**27.08**	**17.27**	**61.83**	**12.79**	**45.84**	**25.59**	**54.37**	**16.84**	**58.64**	**9.38**

129%

136%

Main Results on the Multiple-Choice Task:

Datasets	Models	Gemma-2B-it		LLaMA-2-7B-Chat		LLaMA-3-8B-Instruct		LLaMA-2-13B-Chat	
		ACC ↑	SR ↓	ACC ↑	SR ↓	ACC ↑	SR ↓	ACC ↑	SR ↓
COSE_KRE	Original	35.02	21.28	36.66	23.40	39.93	47.79	49.75	29.13
	Based_on	34.70	22.42	33.22	20.29	42.88	45.34	50.57	29.46
	Based_on_Formatted	38.46	22.42	32.41	18.49	51.55	37.81	41.24	23.57
	Utilizing_Formatted	38.95	22.26	33.06	18.00	50.08	40.10	41.57	21.93
	Opin	35.19	19.97	35.19	**17.35**	**60.23**	**30.11**	43.21	22.91
	ITI (Probe Weight Direction)	31.59	23.57	37.32	17.51	40.75	45.01	50.41	25.37
	ITI (Mass Mean Shift)	29.46	23.73	26.35	18.66	38.95	43.04	25.20	**19.15**
	CAD	37.97	19.64	41.57	19.80	52.86	35.52	56.96	22.59
	IRCAN	39.12	18.99	45.01	24.88	42.72	37.64	49.26	30.11
	IRCAN-CAD	**41.90**	**17.35**	**48.61**	19.48	51.55	31.42	**57.77**	22.09
ECARE_KRE	Original	75.49	24.51	55.04	44.96	57.40	42.60	68.90	31.10
	Based_on	75.59	24.41	61.55	38.45	59.10	40.90	67.86	32.14
	Based_on_Formatted	76.72	23.28	63.15	36.85	69.09	30.91	68.61	31.39
	Utilizing_Formatted	76.44	23.56	60.98	39.02	68.99	31.01	66.16	33.84
	Opin	63.52	36.48	55.04	44.96	**73.80**	**26.20**	57.12	42.88
	ITI (Probe Weight Direction)	73.04	26.96	49.58	50.42	60.51	39.49	73.42	26.58
	ITI (Mass Mean Shift)	73.80	26.20	47.60	52.40	49.58	50.42	71.44	28.56
	CAD	77.76	22.24	73.70	23.30	69.56	30.44	78.13	21.87
	IRCAN	77.38	22.62	76.06	23.94	57.87	42.13	69.84	30.16
	IRCAN-CAD	**82.38**	**17.62**	**80.96**	**19.04**	69.37	30.63	**78.42**	**21.58**

Results of Ablation Studies:

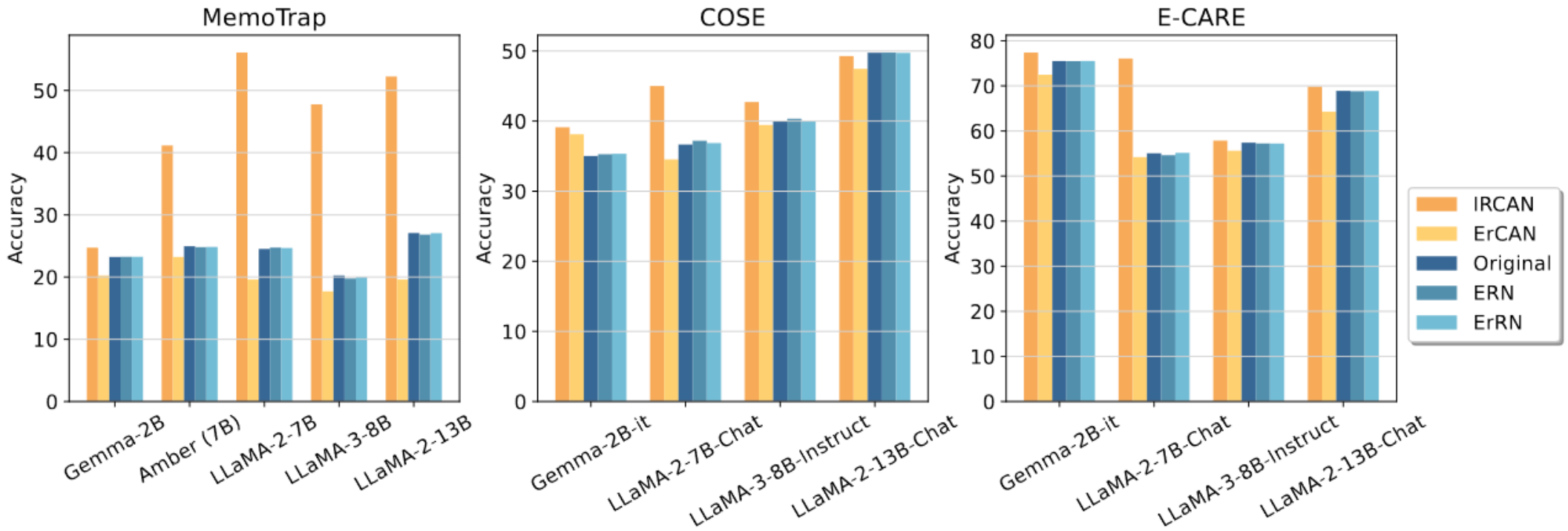


Figure 2: The results of ablation studies to illustrate the accuracy implications of different interventions. **ErCAN** denotes the variant where context-aware neurons are erased. **ERN** represents the enhancement of random neurons. **ErRN** indicates the erasure of random neurons.

Performance of General Abilities on Widely-used Benchmarks:

Models		ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	Average
Gemma-2B	Original	48.29	71.13	40.99	33.02	66.38	17.66	46.25
	IRCAN	48.29	71.42	39.91	33.58	65.11	18.12	46.07
Amber	Original	43.09	73.34	23.99	33.98	66.38	3.49	40.71
	IRCAN	42.24	73.42	24.61	34.22	66.46	3.03	40.66
LLaMA-2-7B	Original	51.96	78.18	45.95	38.97	74.19	13.57	50.47
	IRCAN	52.56	77.15	46.35	37.89	73.01	12.66	49.94
LLaMA-2-13B	Original	57.59	81.72	54.94	36.90	76.01	23.12	55.05
	IRCAN	55.46	78.74	55.40	38.25	76.87	12.36	52.85
LLaMA-3-8B	Original	57.76	81.10	65.14	43.88	77.51	50.72	62.69
	IRCAN	56.48	80.86	64.56	45.08	75.61	36.92	59.92
Gemma-2B-it	Original	44.54	61.74	36.97	45.85	61.64	4.85	42.60
	IRCAN	44.54	61.79	37.38	45.86	61.33	5.00	42.65
LLaMA-2-7B-Chat	Original	51.79	77.73	47.39	45.32	72.53	22.97	52.96
	IRCAN	51.79	77.78	45.74	45.45	72.61	22.21	52.60
LLaMA-3-8B-Instruct	Original	61.34	78.04	65.83	51.69	75.69	75.36	67.99
	IRCAN	60.84	77.98	57.79	52.18	76.01	74.00	66.47
LLaMA-2-13B-Chat	Original	58.53	81.56	53.57	43.96	74.35	34.65	57.77
	IRCAN	58.62	81.58	53.63	43.94	74.43	34.80	57.83

Thank you!

arXiv version: https://arxiv.org/abs/2312.12853
Code Repo: https://github.com/danshi777/IRCAN
Question/Comments: shidan@tju.edu.cn