SJTU Cross Media
Language Intelligence Lab
上海交通大學跨媒體語言智能實驗室

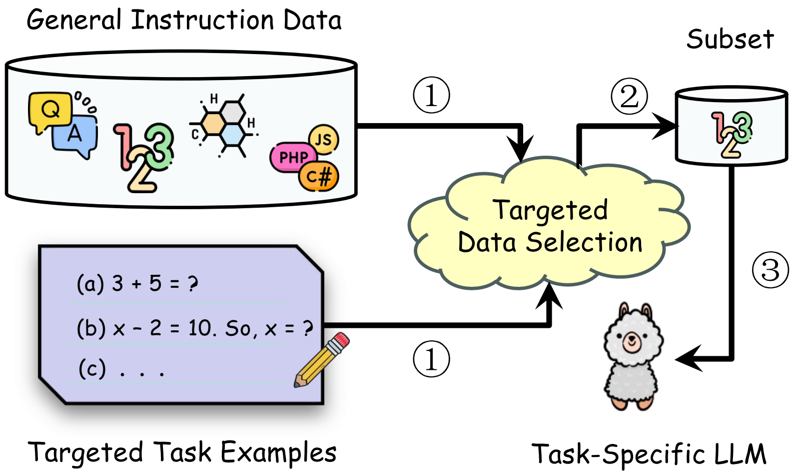
Task-Specific Data Selection for Instruction Tuning via Monosemantic Neuronal Activations

Da Ma, X-LANCE Lab, SJTU

Directory

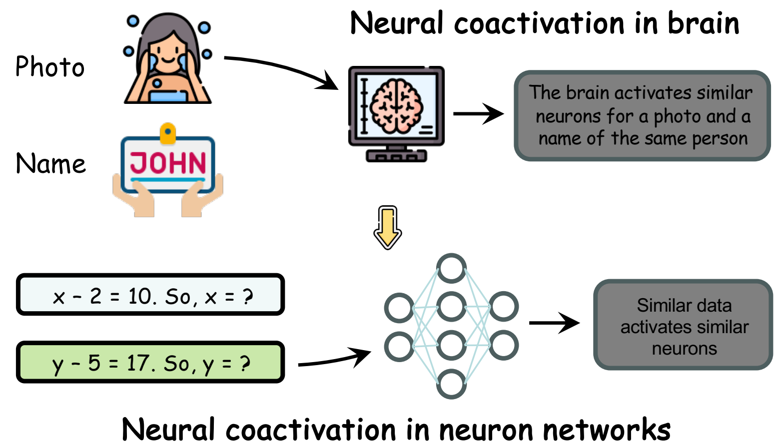
Problem & Motivation

Problem



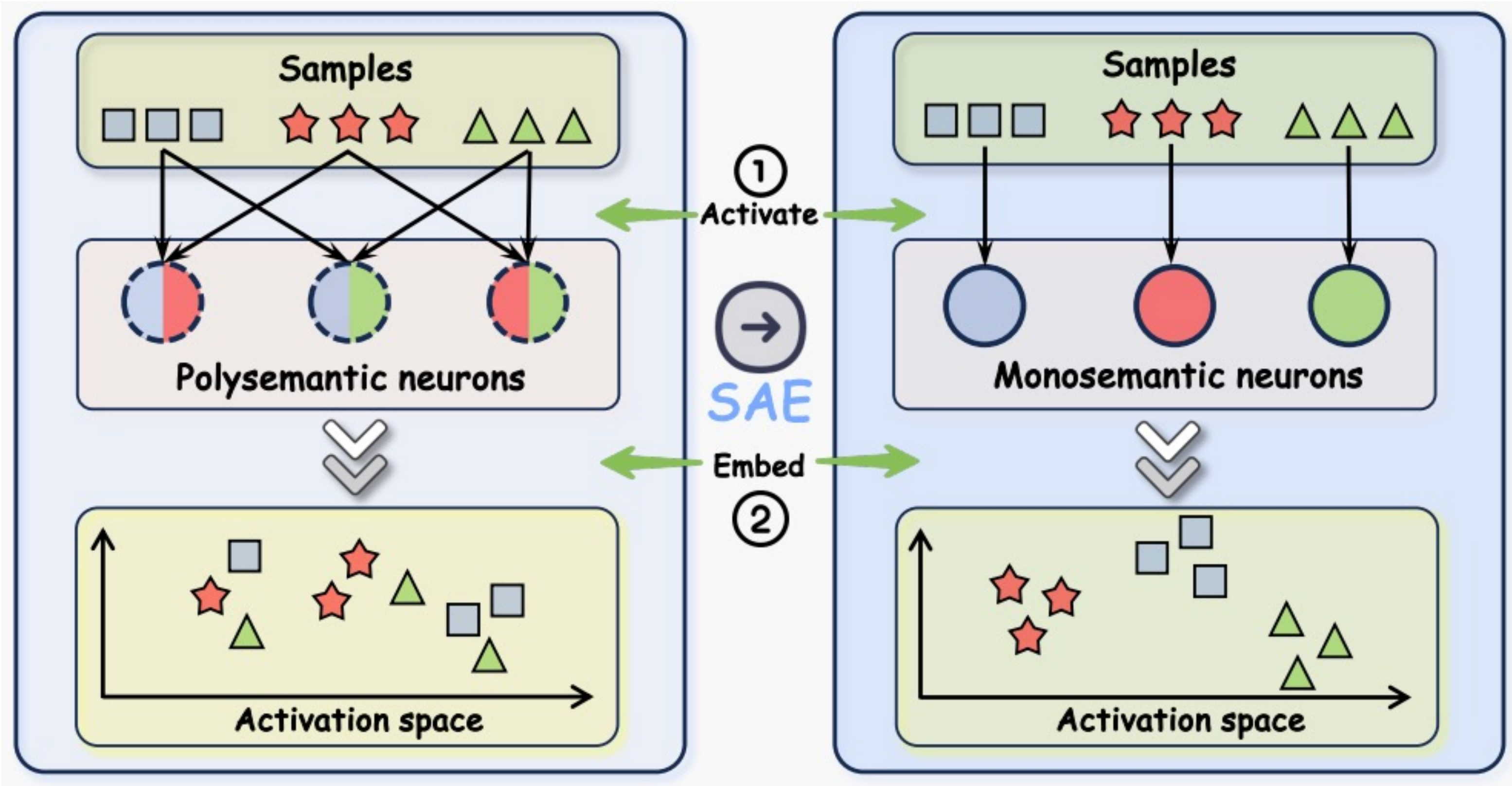
How can we efficiently select data subsets from general instruction datasets to maximize LLM performance on targeted tasks, given only a handful of examples

Motivation



- Existing methods rely on textual or embedding-based features, which overlook how the model internally processes information.
- We propose a model-centric method that selects data based on their neuronal activation patterns in a pretrained model.

Challenge

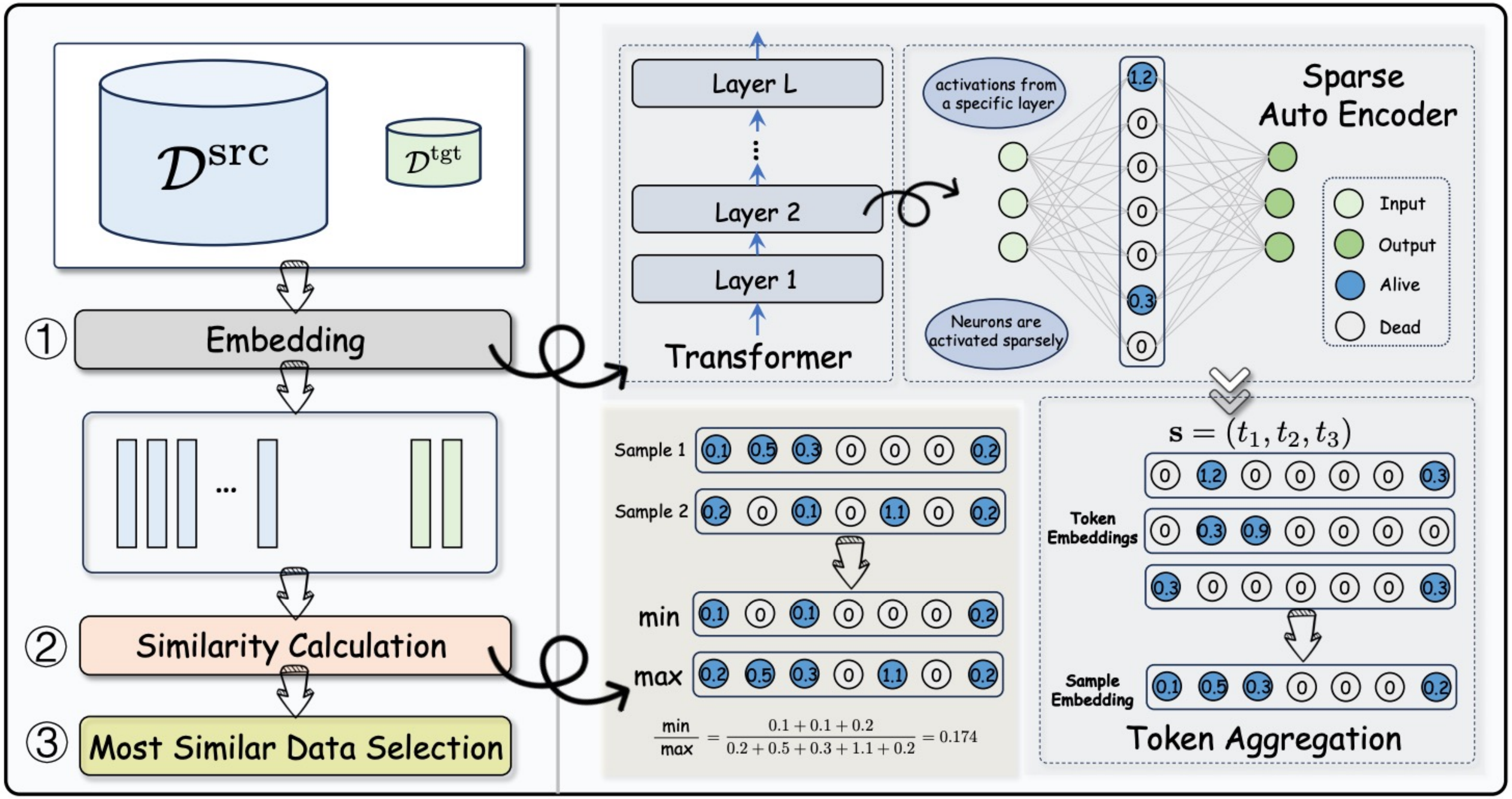


Directly using raw neuron activations can lead to inaccurate similarity estimates because **polysemantic neurons** often respond to multiple unrelated concepts, causing **spurious similarities** between unrelated samples.

目录

SJTU Cross Media
Language Intelligence Lab
上海交通大学跨媒体语言智能实验室

Methods



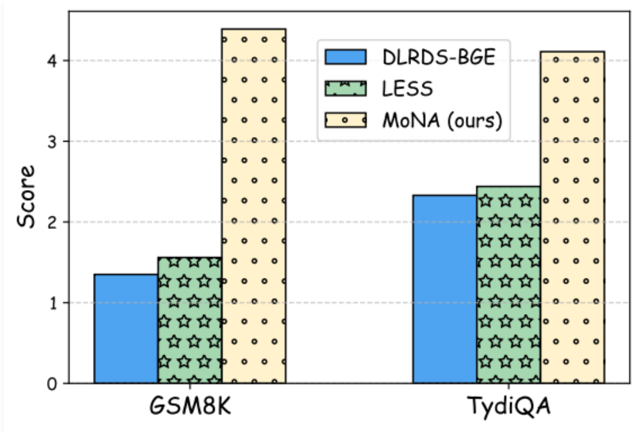
SJTU Cross Media
Language Intelligence Lab

目录

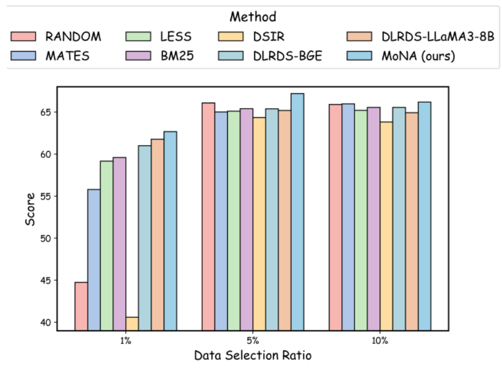
SJTU Cross Media Language Intelligence Lab

MoNA outperforms baselines across datasets and models

Method	$\mathcal{D}^{\mathrm{src}}$ =OpenHermes-2.5						$\mathcal{D}^{\mathrm{src}}$ =Less			
	MMLU	*GSM8K*	*BBH*	*MBPP*	*GPQA*	**Avg.**	*MMLU*	*BBH*	*TydiQA*	**Avg.**
LLaMA3.1-8B										
BASE	65.30	55.50	63.08	46.40	28.12	51.68	65.30	63.08	71.26	66.55
FULL	64.60	65.35	64.31	49.00	27.90	54.23	64.60	64.31	72.66	67.19
RANDOM	64.02	58.65	63.70	46.73	30.36	52.69	64.16	**64.29**	69.78	66.08
Influence-based										
MATES	64.11	54.28	65.38	47.60	28.12	51.90	63.62	63.68	67.74	65.01
LESS	64.34	66.87	63.00	47.80	**31.47**	54.70	62.51	62.11	70.68	65.10
Distribution alignment										
BM25	64.14	66.64	65.23	48.40	27.90	54.46	64.41	63.74	68.07	65.41
DSIR	63.95	66.94	64.29	48.60	29.91	54.74	64.25	63.19	65.61	64.35
DLRDS-BGE	64.45	64.82	64.20	48.60	31.25	54.66	64.06	61.82	70.30	65.39
DLRDS-LLaMA3-8B	64.31	64.75	63.97	**48.80**	29.46	54.26	62.11	61.54	71.91	65.19
LLM2Vec	64.29	63.53	65.55	48.40	30.13	54.38	62.06	62.03	68.11	64.07
MoNA (ours)	**64.49**	**67.93**	**66.44**	48.40	**31.47**	**55.75**	**64.78**	64.21	**72.60**	**67.20**
OLMo-7B										
BASE	28.42	7.35	29.96	21.40	26.56	22.74	28.42	29.96	31.67	30.02
FULL	45.05	31.96	33.13	26.40	26.56	32.62	39.31	28.86	33.43	33.87
RANDOM	36.96	16.00	31.47	19.47	27.38	26.26	28.60	30.82	31.93	30.45
Influence-based										
MATES	30.27	13.72	32.33	16.40	27.01	23.95	29.57	30.46	31.02	30.35
LESS	**46.15**	26.91	33.68	20.20	25.89	30.57	37.21	30.07	33.20	33.49
Distribution alignment										
BM25	42.34	31.08	**34.30**	**26.80**	25.45	31.99	35.74	28.95	**34.40**	33.03
DSIR	36.48	29.26	34.08	19.40	27.23	29.29	29.54	**32.87**	33.25	31.89
DLRDS-BGE	42.77	32.30	33.40	**26.80**	23.88	31.83	35.22	25.65	33.28	31.38
DLRDS-LLaMA3-8B	38.16	31.39	33.30	22.80	**30.13**	31.16	**40.64**	26.08	31.08	32.60
LLM2Vec	37.24	30.10	33.57	23.40	28.35	30.53	39.72	28.58	32.26	33.52
MoNA (ours)	44.74	**32.83**	33.51	26.00	25.00	**32.42**	40.14	30.19	33.80	**34.71**



LLM as a Data Analyst



Performance under different selection ratios

SJTU Cross Media
Language Intelligence Lab

Neuron Activation Visualization

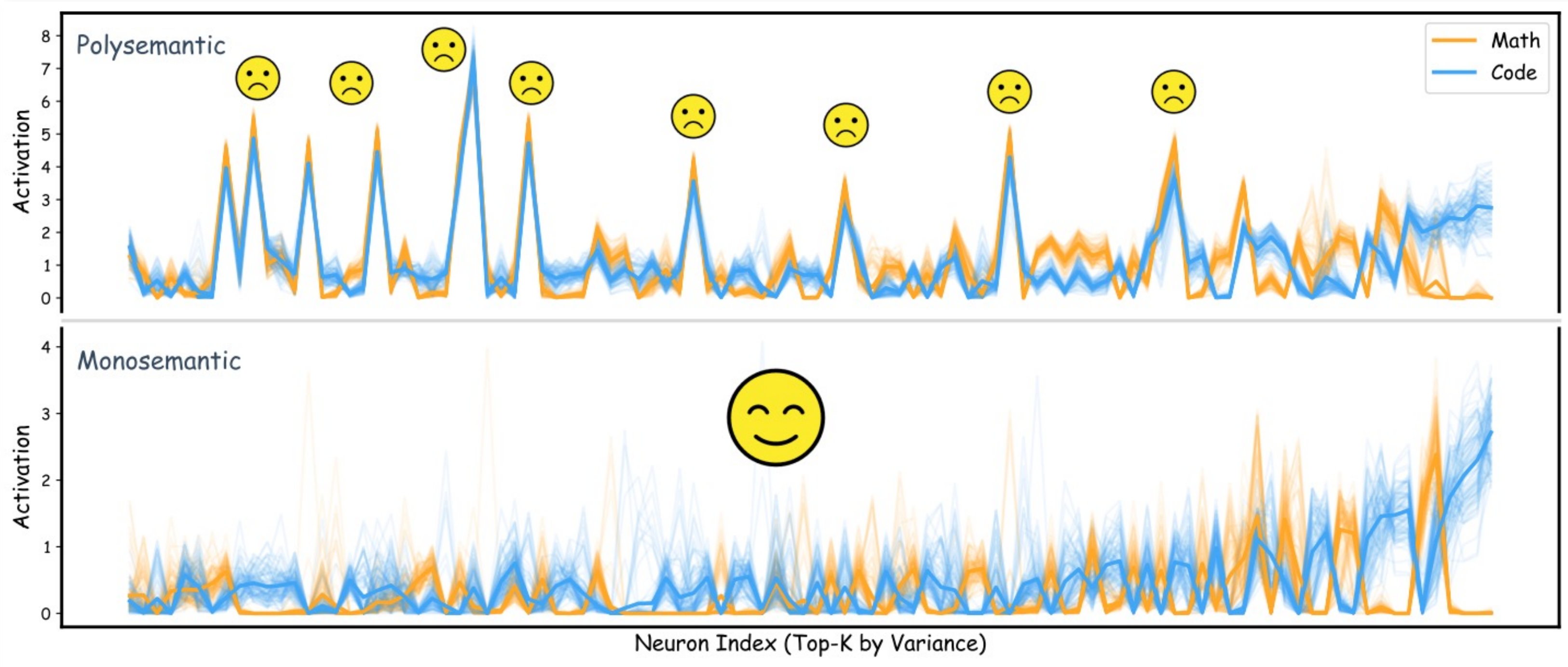


Figure 4: Neuron activation profiles for 100 Math and 100 Code samples on the top-100 most variant neurons. Faint lines show individual samples; bold lines show task means. In the polysemantic (top) plot, many neurons, especially those with high activation peaks (marked by weeping face), are simultaneously activated by both tasks, reflecting pronounced overlap and limited task specificity. In contrast, the monosemantic (bottom) plot reveals clear task-specific activation patterns.

目录

Conclusion

- Monosemantic neuronal activations from sparse autoencoders captures internal model computation and enables more semantically aligned and interpretable data selection.
- Future work includes extending MONA to other training stages, such as pre-training data selection, and applying it to multimodal scenarios, for example, image-text tasks.

Thank you for your listening!