ICML
International Conference
On Machine Learning

香港城市大學
City University of Hong Kong

Li Auto

Step-Level Sparse Autoencoder for Reasoning Process Interpretation

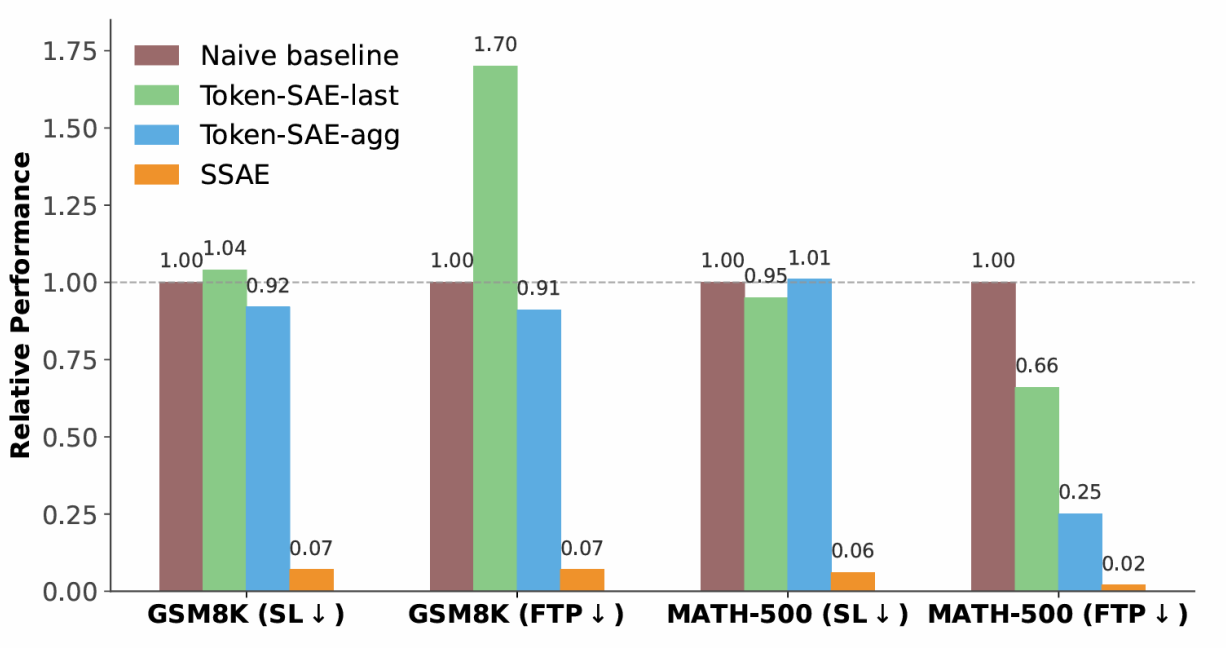
Xuan Yang[*,1,2], Jiayu Liu [*,2], Yuhang Lai [*,1,2], Hao Xu[3], Zhenya Huang[4], Ning Miao[1,2]

Department of Data Science, City University of Hong Kong;
Hong Kong Institute of AI for Science, City University of Hong Kong;
Li Auto Inc;
State Key Laboratory of Cognitive Intelligence ,University of Science and Technology of China (USTC)

Background

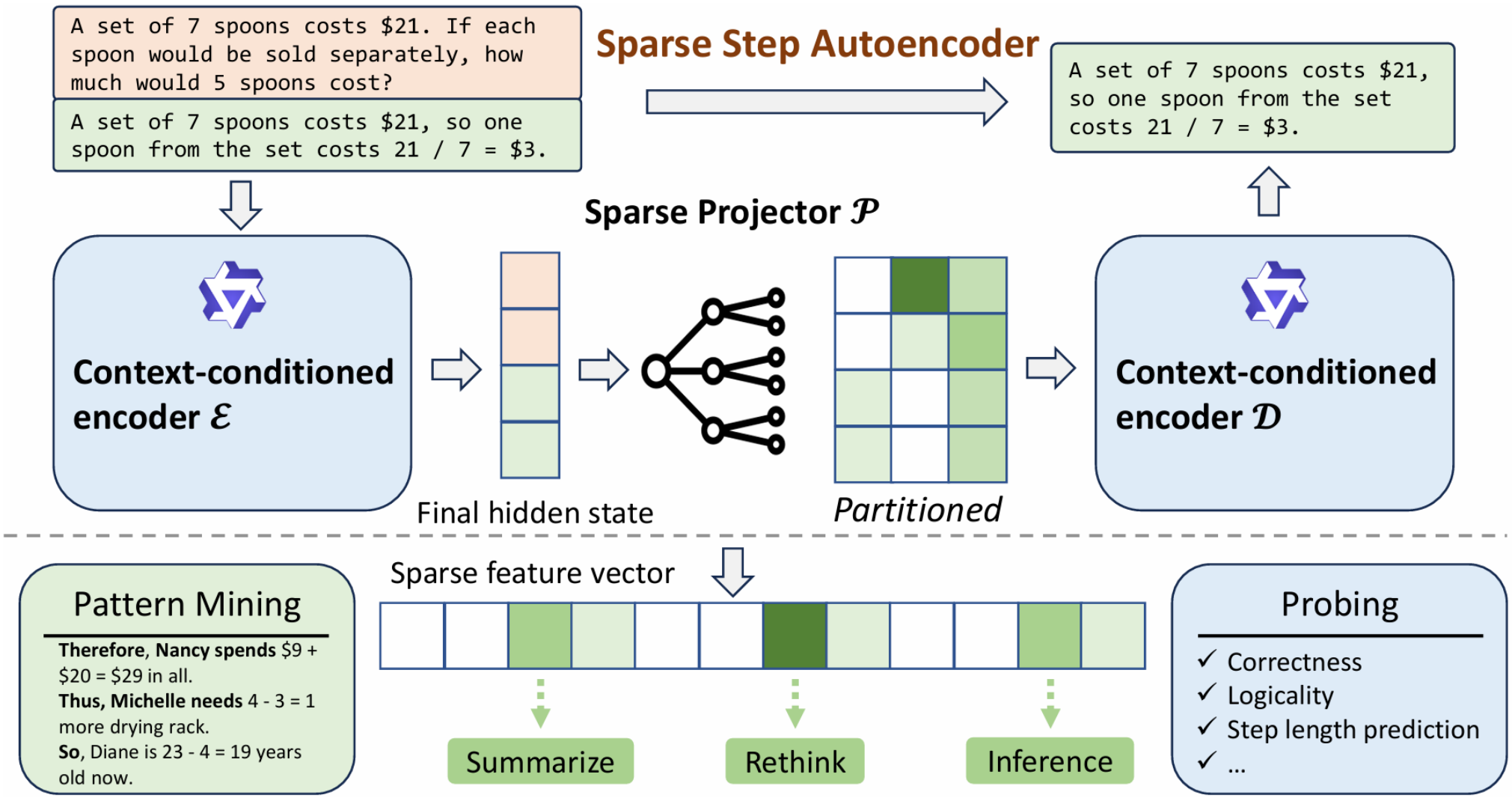
- Traditional SAEs mainly focus on token-level features.
- Step-level features are important to the analysis of LLMs' behaviors.

Granularity mismatch between token-level and step-level!



Traditional token-based SAEs fail to capture such step-level information, such as sentence-length or first-token prediction.

SSAE

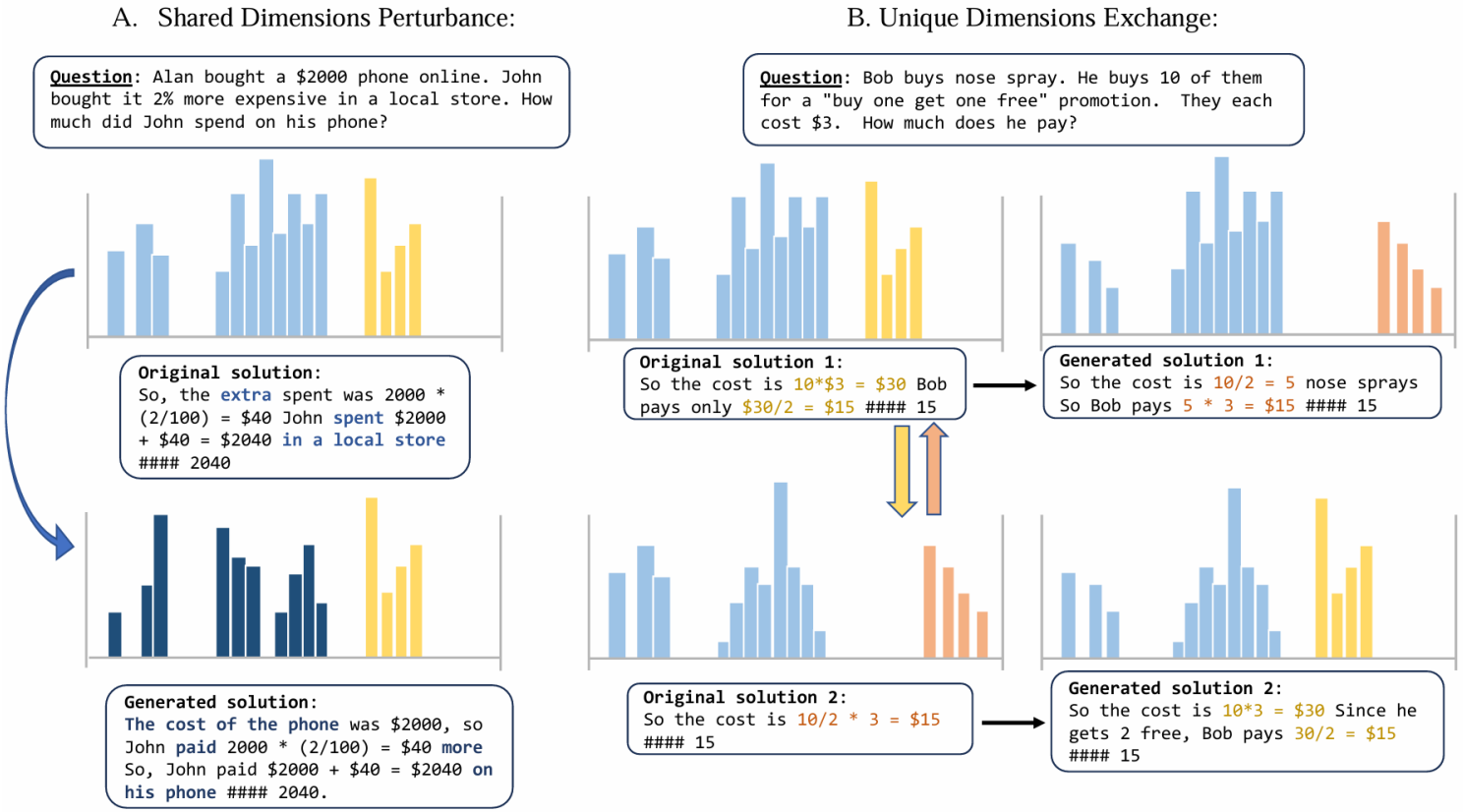

A set of 7 spoons costs $21. If each spoon would be sold separately, how much would 5 spoons cost?
A set of 7 spoons costs $21, so one spoon from the set costs 21 / 7 = $3.
Sparse Step Autoencoder
A set of 7 spoons costs $21, so one spoon from the set costs 21 / 7 = $3.
Sparse Projector $\mathcal{P}$
Context-conditioned encoder $\mathcal{E}$
Final hidden state
Partitioned
Context-conditioned encoder $\mathcal{D}$
Pattern Mining
Therefore, Nancy spends $9 + $20 = $29 in all.
Thus, Michelle needs 4 - 3 = 1 more drying rack.
So, Diane is 23 - 4 = 19 years old now.
Sparse feature vector
Summarize
Rethink
Inference
Probing
✓ Correctness
✓ Logicality
✓ Step length prediction
✓ ...

Probing with Classifier

Model	Feature	Correctness Acc ↑		Logicality Acc ↑	Step Length Error ↓		First Token Perplexity ↓	
		GSM8K	MATH-500	MATH-500	GSM8K	MATH-500	GSM8K	MATH-500
Naive baseline	-	70.49	70.65	55.06	28.04	33.30	61.01	74.96
Token-SAE-last	$\hat{\mathbf{h}}_k$	72.44	<u>86.79</u>	60.56	29.06	31.58	103.54	49.17
Token-SAE-agg	$\hat{\mathbf{h}}_k$	74.38	**86.88**	67.43	25.79	30.33	61.55	18.63
SSAE-Qwen	$\hat{\mathbf{h}}_k$	<u>78.58</u>	82.74	**76.56**	<u>2.10</u>	<u>1.94</u>	<u>4.09</u>	**1.46**
	$\mathbf{h}_k$	72.30	74.52	70.35	22.71	30.35	16.75	15.91
SSAE-Llama	$\hat{\mathbf{h}}_k$	**80.55**	86.24	<u>71.91</u>	**2.02**	**1.59**	**2.66**	<u>1.62</u>
	$\mathbf{h}_k$	73.42	79.11	63.35	20.55	29.17	19.00	15.95

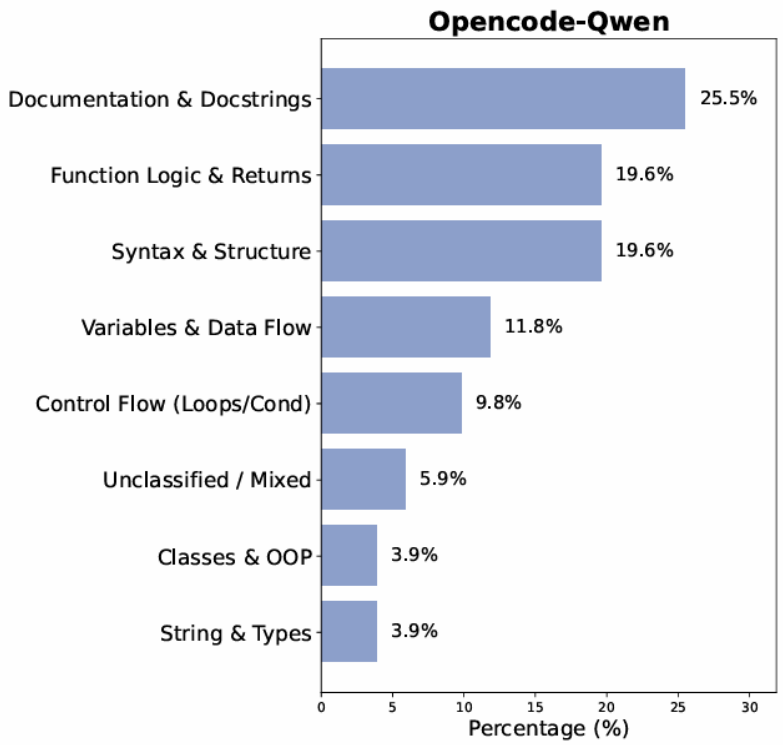
- SSAE features contains rich step-level information and it is linearly decodable.
- Sparse features are more effective in all downstream tasks than dense features.

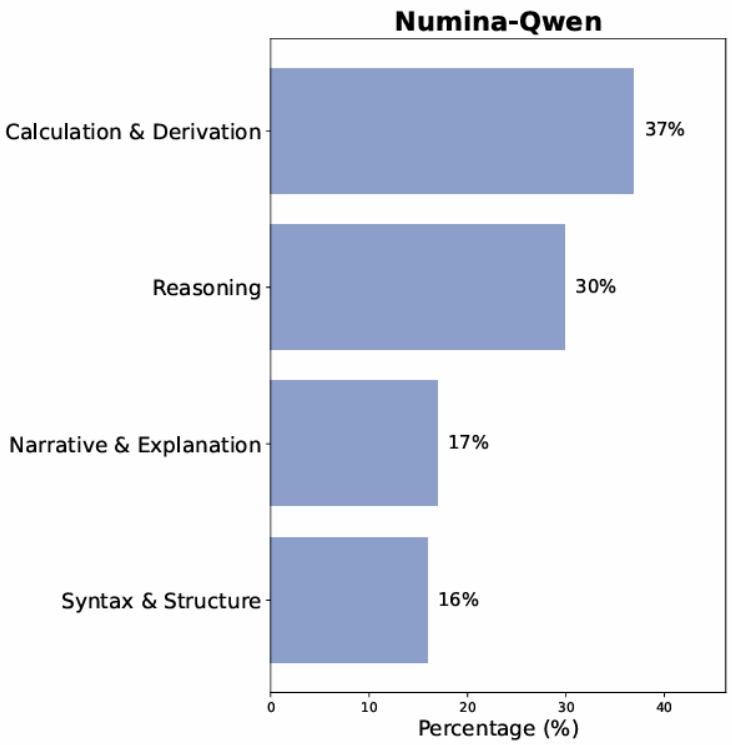
N2G Pattern Mining

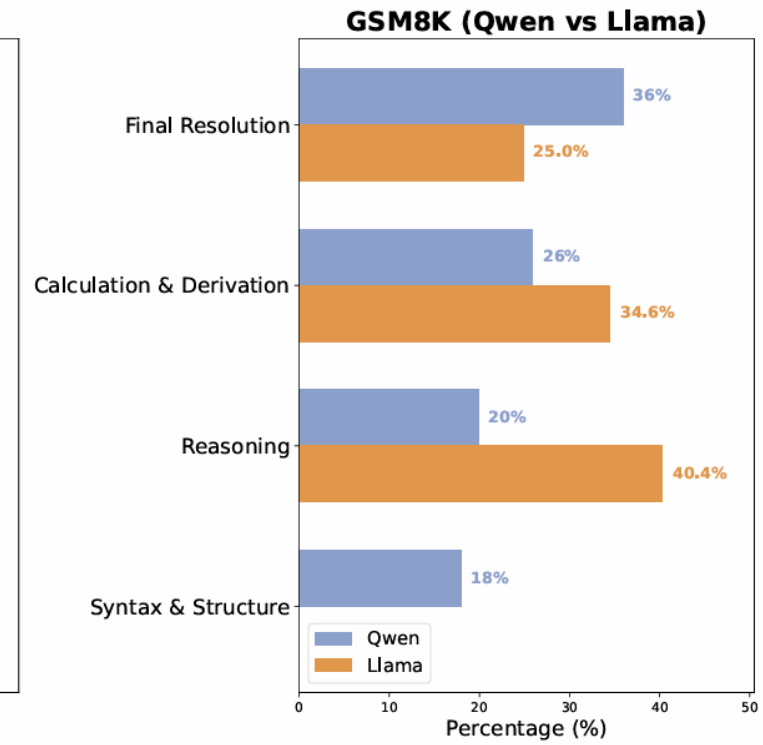


- Shared dimensions mainly encode general surface-level stylistic attributes.
- Unique dimensions focus on deep, reasoning-specific attributes.

N2G Pattern Mining







- The semantic distribution of learned features is highly domain-dependent.
- There is a divergence in latent attention between model families.

Disentanglement of Incremental Information

Dataset	Model	GPT-5	Gemini-2.5-pro	Deepseek-R1	Observed Agreement	Gwet's AC1
GSM8K	SSAE-Qwen	4.96	4.95	4.91	94.87%	0.96
	SSAE-Llama	4.99	4.99	4.98	98.11%	0.98
NuminaMath-CoT	SSAE-Qwen	4.92	4.94	4.94	97.86%	0.98
	SSAE-Llama	4.95	4.97	4.97	98.75%	0.99

Table 4. Quantitative results of the latent-swap test. Scores (1–5) evaluate semantic independence, where 5 means fully independent and 1 means fully redundant.

◆ SSAE only extracts the incremental information of the current reasoning step and squeezes background information

Probing Guided Weighted Voting

Model	Strategy	GSM8K	SVAMP	MultiArith
Qwen2.5-0.5B	Avg@16	30.40	19.33	35.56
	SC@16	46.20	29.67	59.44
	PG@16 (ours)	**46.80**	**33.00**	**61.67**
Llama-3.2-1B	Avg@16	12.00	32.33	67.22
	SC@16	16.60	48.00	82.78
	PG@16 (ours)	**19.40**	**50.33**	**83.89**

Model	Strategy	MATH-500	AIME 2024	AIME 2025
Qwen2.5-7B-Instruct	Avg@16	74.00	12.08	6.46
	SC@16	80.20	16.67	13.33
	PG@16 (Ours)	**80.80**	16.67	13.33
Deepseek-R1-Distill-Qwen32B	Avg@16	94.30	72.60	42.92
	SC@16	95.60	86.67	66.67
	PG@16 (Ours)	95.60	**90.00**	66.67

- SSAE captures useful reasoning signals and can be used to guide inference-time scaling.
- Features learned on small models can transfer to larger models without additional training.

ICML
International Conference
On Machine Learning

CityU
香港城市大學
City University of Hong Kong

Li Auto

Thanks for your listening!

For more details, please refer to our paper

Welcome to discuss with us

Xuan Yang

xyang753-c@my.cityu.edu.hk

https://github.com/TorresYangX/SSAE