RSAE: one flexible activation matches — then beats — ReLU, JumpReLU & TopK

Rational Sparse Autoencoder (RSAE)

Naiyu Yin · Yue Yu — Department of Mathematics, Lehigh University

1 TL;DR

MAIN 3 TAKEAWAYS

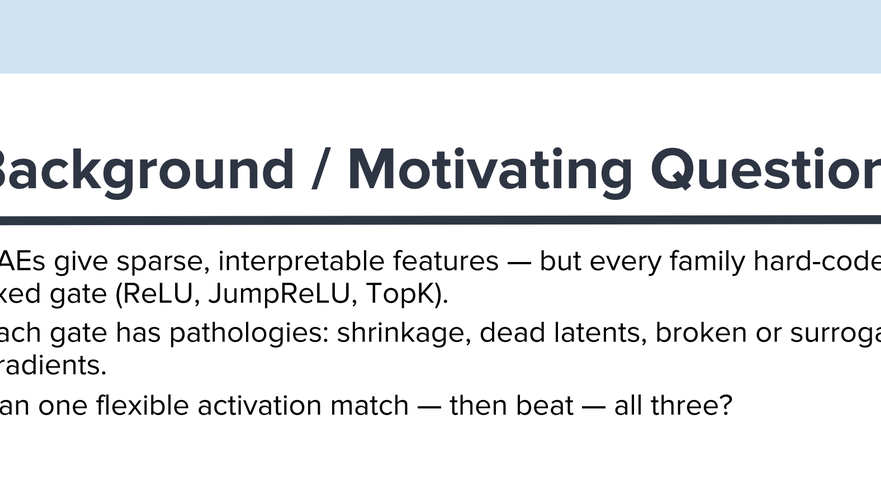
- Replace the SAE's fixed gate (ReLU / JumpReLU / TopK) with a **trainable rational activation**.
- **Strictly beats every baseline** on reconstruction & downstream metrics at matched sparsity.
- **Drop-in upgrade**: inherits weights, adds a few scalars, fine-tunes in minutes.

2 Background / Motivating Question

- SAEs give sparse, interpretable features — but every family hard-codes a fixed gate (ReLU, JumpReLU, TopK).
- Each gate has pathologies: shrinkage, dead latents, broken or surrogate gradients.
- Can one flexible activation match — then beat — all three?

3 Our Approach: RSAE

- Learnable rational $r(t) = \frac{P(t)}{Q(t)} = \frac{\sum_{i=0}^{p} b_i t^i}{\sum_{j=0}^{q} b_j t^j}$, applied element-wise, with input/output scales.
- Approximates all three gates at low degree — theory: rationals need $\mathcal{O}(1)$ coords vs piecewise-affine $\Omega(\frac{1}{\sqrt{\varepsilon}})$.
- Two-step: initialization from baseline weights with *Remez fit* of rational functions, then fine-tune under the reconstruction loss with ℓ_1 sparsity penalty.

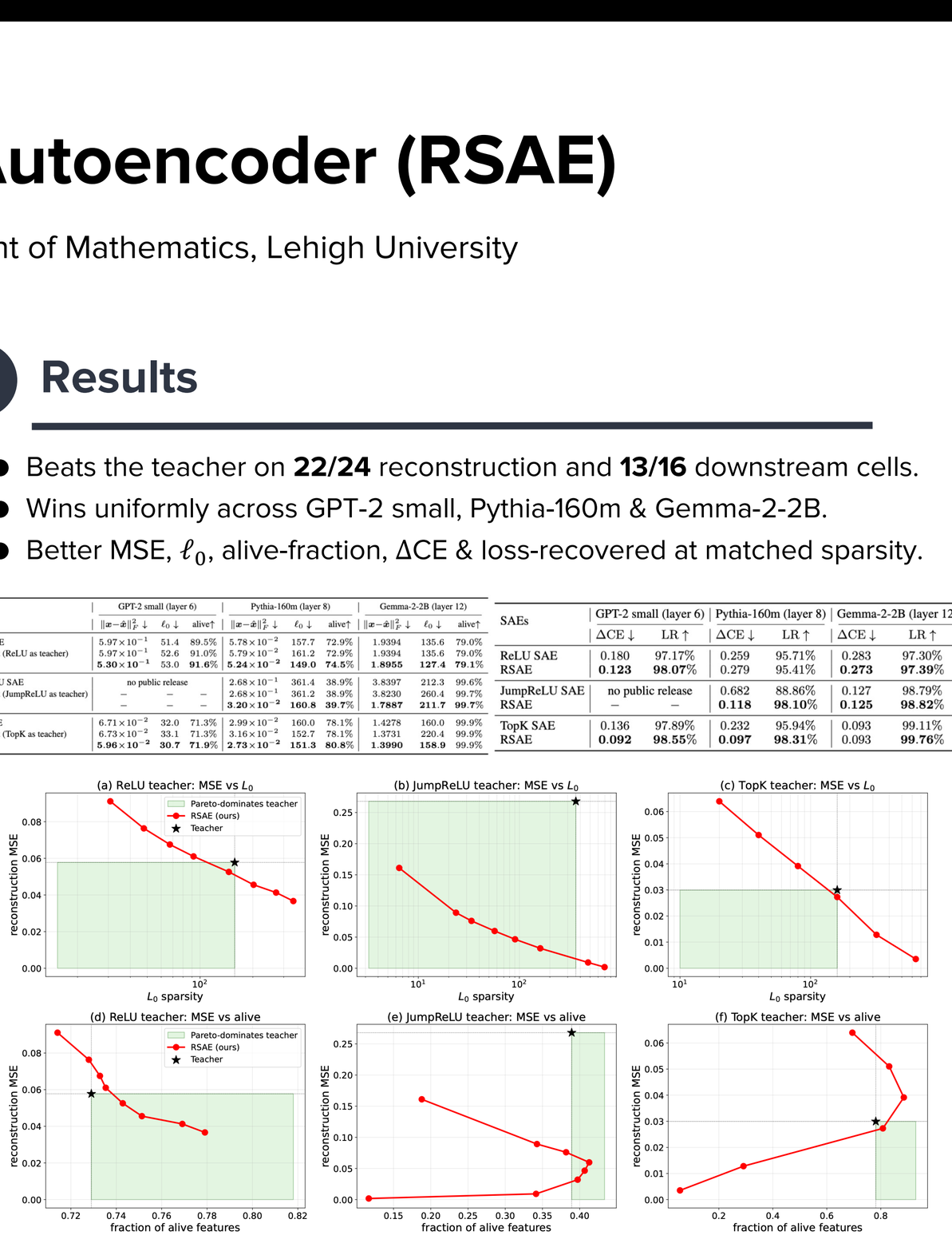


4 Results

- Beats the teacher on **22/24** reconstruction and **13/16** downstream cells.
- Wins uniformly across GPT-2 small, Pythia-160m & Gemma-2-2B.
- Better MSE, ℓ_0, alive-fraction, ΔCE & loss-recovered at matched sparsity.

SAEs	GPT-2 small (layer 6)			Pythia-160m (layer 8)			Gemma-2-2B (layer 12)		
	$\|x-\hat{x}\|_F^2 \downarrow$	$\ell_0 \downarrow$	alive↑	$\|x-\hat{x}\|_F^2 \downarrow$	$\ell_0 \downarrow$	alive↑	$\|x-\hat{x}\|_F^2 \downarrow$	$\ell_0 \downarrow$	alive↑
ReLU SAE	5.97×10^{-1}	51.4	89.5%	5.78×10^{-2}	157.7	72.9%	1.9394	135.6	79.0%
RSAE init (ReLU as teacher)	5.97×10^{-1}	52.6	91.0%	5.79×10^{-2}	161.2	72.9%	1.9394	135.6	79.0%
RSAE	$\mathbf{5.30\times10^{-1}}$	53.0	**91.6%**	$\mathbf{5.24\times10^{-2}}$	**149.0**	**74.5%**	**1.8955**	**127.4**	**79.1%**
JumpReLU SAE	no public release			2.68×10^{-1}	361.4	38.9%	3.8397	212.3	99.6%
RSAE init (JumpReLU as teacher)	–	–	–	2.68×10^{-1}	361.2	38.9%	3.8230	260.4	99.7%
RSAE	–	–	–	$\mathbf{3.20\times10^{-2}}$	**160.8**	**39.7%**	**1.7887**	**211.7**	**99.7%**
TopK SAE	6.71×10^{-2}	32.0	71.3%	2.99×10^{-2}	160.0	78.1%	1.4278	160.0	99.9%
RSAE init (TopK as teacher)	6.73×10^{-2}	33.1	71.3%	3.16×10^{-2}	152.7	78.1%	1.3731	220.4	99.9%
RSAE	$\mathbf{5.96\times10^{-2}}$	**30.7**	**71.9%**	$\mathbf{2.73\times10^{-2}}$	**151.3**	**80.8%**	**1.3990**	**158.9**	99.9%

SAEs	GPT-2 small (layer 6)		Pythia-160m (layer 8)		Gemma-2-2B (layer 12)	
	ΔCE ↓	LR ↑	ΔCE ↓	LR ↑	ΔCE ↓	LR ↑
ReLU SAE	0.180	97.17%	0.259	95.71%	0.283	97.30%
RSAE	**0.123**	**98.07%**	0.279	95.41%	**0.273**	**97.39%**
JumpReLU SAE	no public release		0.682	88.86%	0.127	98.79%
RSAE	–	–	**0.118**	**98.10%**	**0.125**	**98.82%**
TopK SAE	0.136	97.89%	0.232	95.94%	0.093	99.11%
RSAE	**0.092**	**98.55%**	**0.097**	**98.31%**	0.093	**99.76%**



Dataset	Teacher full ↑	RSAE full ↑	Δ
bias_in_bios_set1	95.28	**95.30**	+0.02
bias_in_bios_set2	**93.12**	93.04	−0.08
bias_in_bios_set3	90.94	**91.22**	+0.28
amazon_reviews	87.98	**88.00**	+0.02
amazon_sentiment	91.40	91.40	0.00
github-code	**96.64**	96.50	−0.14
ag_news	94.18	**94.25**	+0.07
europarl	99.96	99.96	0.00
mean	93.69	**93.71**	+0.02

- **Ablations:** sweeping the sparsity coefficient enters the strict-domination “sweet zone” vs every teacher.
- **Interpretability:** sparse-probing accuracy ties or improves on 6/8 tasks.

5 Limitations and Discussion

- **Summary:** One flexible, trainable rational activation matches all three SAE families (ReLU, JumpReLU, TopK) and then strictly beats them — a drop-in upgrade with better fidelity at matched sparsity for a handful of scalar parameters and minutes on a single GPU.
- **Limitations:** 1) Evaluation covers only 3 open-weight LLMs and 3 activation families — gains at larger frontier models remain unexplored. 2) RSAE’s interpretability remains to be more fully evaluated, beyond the single-model sparse-probing study reported here.

PAPER

Mechanistic Interpretability Workshop, ICML 2026

Naiyu Yin: nay224@lehigh.edu