FROST: Filtering Reasoning Outliers with Attention for Efficient Reasoning

Haozheng Luo, Zhuolin Jiang, Md Zahid Hasan, Yan Chen, Soumalya Sarkar

Northwestern University & RTX Technology Research Center (RTRC)

https://openreview.net/forum?id=a9dngZLqGS

ICLR2026

Problem: Large Reasoning Models (LRMs) often generate numerous irrelevant steps, which we term **reasoning outliers**.

Proposal: Filtering Reasoning Outliers with Attention (termed FROST) via outlier-removal architecture and supervised finetuning.

- Serves as an outlier-free model structure that mitigates outliers induced by reasoning traces.
- Improves the efficiency of LRMs during reasoning generation and formalizes the concept of reasoning outliers.
- We theoretically prove that $\mathrm{Softmax}_1$ suppresses low attention weights while preserving high ones at the sentence level.
- Achieves up to a **26.70%** accuracy gain while reducing reasoning path length by **69.68%** relative to base models, cutting inference time by at least **28.6%**, and reduces training time by **42.2%** compared to SFT baselines.

Motivation: Efficiency in Large Reasoning Models

Large Reasoning Models: [Guo et al., 2025] introduce DeepSeek-R1, demonstrating strong reasoning performance, particularly on mathematical and logical tasks.

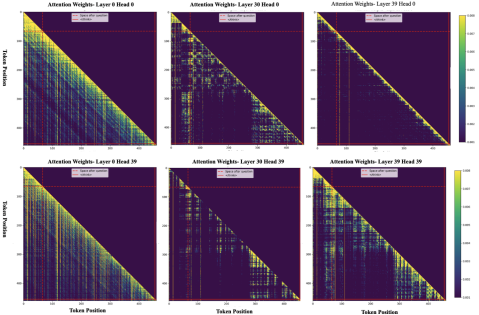
- However, large reasoning models (LRMs) often generate many irrelevant reasoning steps, which degrade both performance and efficiency.

Efficient Reasoning Methods: [Sui et al., 2025] categorize efficient reasoning methods into three types:

- **Prompt-based:** `Chain-of-Draft`, `TALE`, `Thought Manipulation`
- **SFT-based**: `DRP`, `CoT-Valve`, `TokenSkip`.
- **Reinforcement Learning**: `SelfBudgeter`, `Thinkprune`, `ThinkLess`.

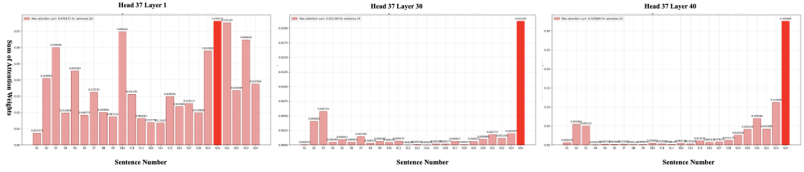
Attention Heatmap of Reasoning Tokens.

- Use the Phi-4-Reasoning model to generate a reasoning trace for a representative GSM8K example.
- In shallow layers, token contributions to the final answer are nearly uniform, whereas deeper layers and later attention heads concentrate on a small set of tokens with substantially higher influence.

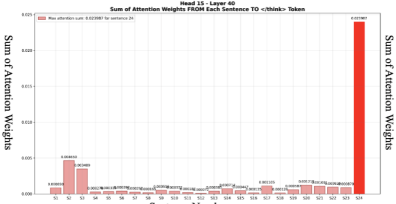


Total Attention Weight Distribution to the Final Answer

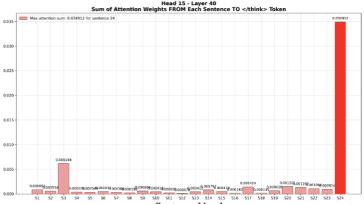
- Only a small number of reasoning traces contribute strongly to the final *</think>* token, while many traces have negligible influence.



- The attention weight distributions before and after outlier removal show that the model's focus on critical reasoning traces is preserved or enhanced, while the influence of outliers is significantly reduced.



Vanilla

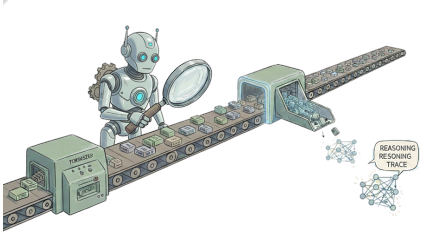


FROST

- We propose an efficient reasoning training framework with supervised finetuning (SFT) that replaces $\mathrm{Softmax}$ in the attention mechanism with $\mathrm{Softmax}_1$ to enable efficient reasoning.

$$\mathrm{Softmax}_1(S) := \frac{\exp(S)}{1 + \sum_{i=1}^{L} \exp(S_i)},$$

- Few fine-tuning steps from existing pretrained checkpoints.



Compare FROST with vanilla attention across three LRMs on several math tasks.

Type	Method	GSM8K		MATH500		AIME24		Minerva		$\Delta_{\text{Pass@1}}$	$\Delta_{\#\text{Tk}}$
		Pass@1	#Tk	Pass@1	#Tk	Pass@1	#Tk	Pass@1	#Tk		
Phi-4 -Reasoning	Base	0.9242	1017.70	0.5480	1721.95	0.0667	1017.70	0.2500	1898.86	0.000	0.00
	TALE	**0.9500**	1716.60	0.5800	1874.43	0.2900	2069.97	0.2627	2093.17	**+0.074**	+524.49
	DRP	0.8340	721.00	**0.6200**	2122.00	**0.3333**	6135.00	0.2701	1289.50	+0.067	+1152.69
	SelfBudgeter	0.9189	1507.14	0.5347	1195.18	0.1342	1372.83	0.2357	2618.23	+0.009	+259.30
	ThinkLess	0.9279	1421.90	0.5414	1101.21	0.1608	1405.40	0.2575	1708.70	+0.025	-4.75
	Ours	0.9311	**154.33**	0.5980	**344.37**	0.2667	**899.80**	**0.2716**	**401.19**	+0.070	**-964.13**
GPT-OSS-20B	Base	0.8704	1275.23	0.5400	1575.36	0.1333	**1003.57**	0.2574	1586.95	0.000	0.00
	TALE	0.8283	2664.41	0.5454	3878.87	0.2000	1354.67	0.2700	3262.47	+0.011	+1430.33
	DRP	0.7880	902.50	**0.6146**	4137.00	**0.2245**	4983.00	0.2715	1885.15	+0.024	+1616.64
	SelfBudgeter	0.8610	1850.00	0.5340	2285.00	0.1320	1256.00	0.2550	1298.00	-0.005	+312.47
	ThinkLess	0.8740	1785.00	0.5410	2206.00	0.1600	1205.00	0.2580	1220.00	+0.008	+244.22
	Ours	**0.8764**	**377.17**	0.5800	**680.89**	0.1667	1009.60	**0.2794**	**691.71**	**+0.025**	**-669.94**
Magistral -Small-1.1	Base	0.6075	2664.41	0.1480	1389.89	0.0000	537.13	0.0699	1288.04	0.000	0.00
	TALE	0.7146	1516.86	**0.3040**	723.91	0.0333	967.43	0.1544	748.18	**+0.095**	-480.77
	DRP	0.6500	902.50	0.2100	1680.33	0.0450	1350.77	0.1120	1604.22	+0.048	-85.41
	SelfBudgeter	0.6900	1850.00	0.2300	1520.00	0.0520	1256.00	0.1300	1298.00	+0.069	+11.13
	ThinkLess	0.7200	1785.00	0.2500	1405.00	0.0600	1205.00	0.1450	1220.00	+0.087	-66.12
	Ours	**0.7551**	**137.55**	**0.3040**	**98.20**	**0.0974**	**149.93**	**0.1551**	**109.23**	**+0.122**	**-1346.14**

Results: FROST improves accuracy by an average of **26.70%** and reduces token usage by **69.68%** on the three base models, GPT-OSS-20B, Magistral- Small-1.1 and Phi-4-reasoning.

Efficiency of Different Activation Functions

We compare FROST with different $\mathrm{Softmax}$ activation functions on Phi-4-Reasoning.

Method	GSM8K		MATH500		AIME24		Minerva		$\overline{\text{Pass@1}}$	$\overline{\#\text{Tk}}$
	Pass@1	#Tk	Pass@1	#Tk	Pass@1	#Tk	Pass@1	#Tk		
Base	0.9242	1017.70	0.5480	1721.95	0.0667	1017.70	0.2500	1898.86	0.4472	1414.05
Softmax	0.8317	1160.63	0.4880	1379.52	0.1333	1909.07	0.2390	1934.72	0.4230	1595.99
Sparsemax	0.8188	160.99	0.5120	451.59	0.1667	948.60	0.2647	580.84	0.4406	535.26
Entmax15	0.8984	163.75	0.5520	406.97	0.1667	**876.63**	**0.2831**	439.48	0.4751	471.71
$\mathrm{Softmax}_1$ (FROST)	**0.9311**	**154.33**	**0.5980**	**344.37**	**0.2667**	899.80	0.2716	**401.19**	**0.5169**	**449.92**

Results: FROST achieves the best overall performance in both Pass@1 accuracy and token usage. Specifically, the average accuracy increases by **15.65%**, while the number of tokens decreases by **68.18%** compared to the base model.

Compare FROST on Phi-4-Reasoning using outlier metrics and sentence entropy.

Method	Maximum Infinity Norm $\|\mathbf{x}\|_{\infty}\downarrow$	Average Kurtosis ↓	Average Sentence Entropy ↑	Pass@1 ↑	#Tk ↓
Base	35.31	241.72	2.71	0.0667	1017.70
Softmax	34.53	189.36	2.79	0.1333	1909.07
Sparsemax	34.06	152.18	<u>2.93</u>	<u>0.1667</u>	948.60
Entmax15	<u>30.39</u>	<u>43.72</u>	2.92	<u>0.1667</u>	**876.63**
FROST	**29.67**	**21.54**	**3.07**	**0.2667**	<u>899.80</u>

Results: FROST reduces average kurtosis by **91+%** and the maximum infinity norm of reasoning outputs by **15+%** relative to the base model. And FROST achieves an average accuracy gain of **15.65%** while reducing the number of generated tokens by **68.18%** compared to the base model.

Compare FROST with vanilla attention on other non-mathematical reasoning tasks.

Method	Leetcode		LiveCodeBench		UGPhysics		$\overline{\text{Pass@1}}$	$\overline{\#\text{Tk}}$
	Pass@1	#Tk	Pass@1	#Tk	Pass@1	#Tk		
Base	0.3222	2755.13	0.3248	3154.80	0.3172	2603.00	0.3214	2837.64
Softmax	0.3778	2106.85	0.3538	2909.07	0.3011	2622.52	0.3442	2546.15
FROST	**0.3889**	**1163.06**	**0.3777**	**1967.56**	**0.3473**	**805.77**	**0.3713**	**1312.13**

Results: FROST preserves—and even improves—generalization to unseen reasoning tasks while reducing token cost during reasoning.

Summary

- Filtering Reasoning Outliers with Attention
  - Manages outliers in transformer-based LRMs.
  - Remove outliers in reasoning generation.
- Theoretical Enhancements
  - Provide an expressive guarantee for deployment-time suppression of low-attention sentences.
  - Prove that the outlier-suppression property of $\mathrm{Softmax}_1$ extends from token-level attention weights to sentence-level aggregation.
- New Outlier Concept
  - Show that low-attention regions correspond to non-critical reasoning traces, which we define as reasoning outliers.
- Empirical Performance of FROST
  - Achieves up to a **26.70%** accuracy gain while reducing reasoning path length by **69.68%** relative to base models,
  - Cuts inference time by at least **28.6%**, and reduces training time by **42.2%** compared to SFT baselines.

Thank You!

Haozheng Luo, Zhuolin Jiang, Md Zahid Hasan, Yan Chen, Soumalya Sarkar

hluo@u.northwestern.edu
zhuolin.jiang@rtx.com
zahid@iastate.edu
ychen@northwestern.edu
soumalya.sarkar@rtx.com