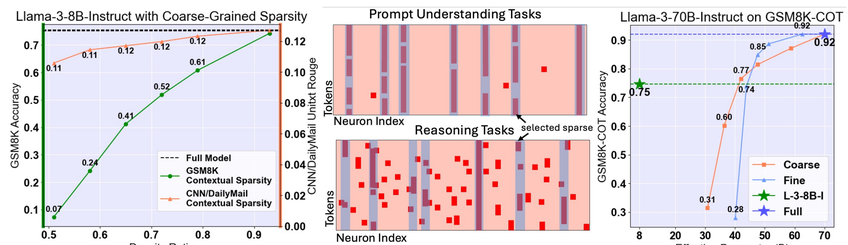
# Sirius: Contextual Sparsity with Correction for Efficient LLMs

Yang Zhou[1], Zhuoming Chen[1], Zhaozhuo Xu[2], Xi Victoria Lin[3], Beidi Chen[1,3]
Carnegie Mellon University[1], Stevens Institute of Technology[2], Meta[3]

NEURAL INFORMATION PROCESSING SYSTEMS

## Problem: Contextual Sparsity Struggles at Complex Generation



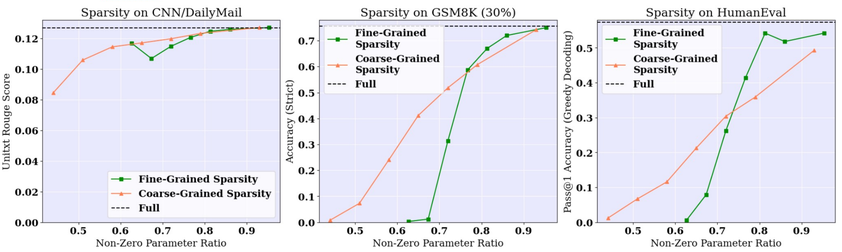

(a) Comparing Sparsity Method on CNN/DailyMail and GSM8K




(b) Visualization of Activation Maps across the generation sequence




(b) Visualization of Activation Maps across the generation sequence





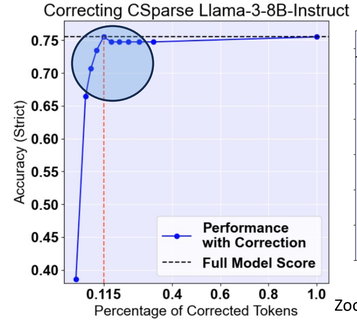


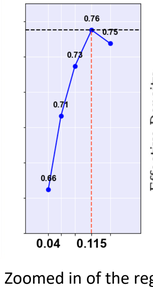


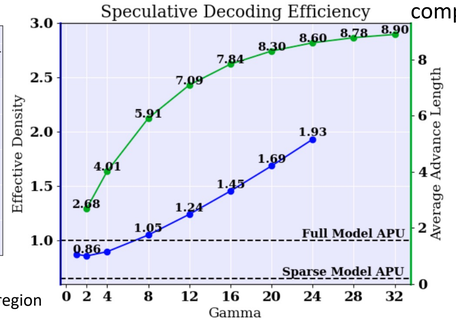

| **Where CS Succeeds** Experiment Settings | **CNN/DailyMail** Unitxt Rouge | **CoQA** EM/F1 | **TruthfulQA** Rouge-1/2 ACC |
|---|---|---|---|
| **Llama-3-8B-Instruct** | 0.1237 | 0.6153/0.7825 | 0.4945/0.3647 |
| **Llama-3-8B-Instruct-CSparse** | 0.1144 | 0.6633/0.7977 | 0.4725/0.3403 |
| **Llama-3-8B-Instruct-FSparse** | 0.1166 | 0.6625/0.7984 | 0.5043/0.3305 |
| **Llama-2-7B-Chat** | 0.1489 | 0.5982/0.7580 | 0.4480/0.3831 |
| **Llama-2-7B-Chat-CSparse** | 0.1448 | 0.6117/0.7639 | 0.4529/0.3843 |
| **Llama-2-7B-Chat-FSparse** | 0.1521 | 0.5898/0.7540 | 0.4565/0.3660 |
| **Where CS Fails** Experiment Settings | **GSM8K** ACC (strict/flexible) | **HumanEval** Pass@1 (GD) | **MMLU*** Accuracy |
| **Llama-3-8B-Instruct** | 0.7551/0.7544 | 0.560 | 0.6231 |
| **Llama-3-8B-Instruct-CSparse** | 0.3859/0.3874 | 0.207 | 0.5558 |
| **Llama-3-8B-Instruct-FSparse** | 0.5868/0.5891 | 0.457 | 0.5304 |
| **Llama-2-7B-Chat** | 0.2396/0.2462 | 0.140 | 0.492 |
| **Llama-2-7B-Chat-CSparse** | 0.1334/0.1380 | 0.067 | 0.4637 |
| **Llama-2-7B-Chat-FSparse** | 0.1979/0.2017 | 0.134 | 0.4768 |

* **MMLU** is a classification task, not generation tasks. We use **MMLU-FLAN-COT**

## Observation: Oracle Correction Mechanism + SD Discussion




## Error Analysis: Three Categories of the Accuracy Crashes

**Question:** Marie ordered one chicken meal that costs $12, 5 packs of milk that costs $3 each, 4 apples that cost $1.50 each, and some boxes of pizza. Marie paid a total of $50. How many boxes of pizza did Marie order if each box costs $8.50?

**Answer:** The total cost of the chicken meal, milk, and apples is 12 + 5 * 3 + 4 * 1.5 = <<12+5*3+4*1.5=27>>27 dollars *(33 dollars)* .....

Miscalculation

**Question:** A merchant wants to make a choice of purchase between 2 purchase plans: jewelry worth $5,000 or electronic gadgets worth $8,000. His financial advisor speculates that the jewelry market will go up 2.5% while the electronic gadgets market will rise 1.2% within the same month. If the merchant is looking to maximize profit at the end of this month by making a choice, how much profit would this be?

**Answer:** The jewelry will increase in value by 2.5% which is 5,000 * 0.025 = **$1250** *($125)* .....

Miscalculation

**Question:** Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

**Answer (Sparse):**
Janet eats 3 + 4 = <<3+4=7>>7 eggs per day.
She lays 16 - 7 = <<16-7=9>>9 eggs per day.
She sells 9 - 7 = <<9-7=2>>2 eggs per day.
She makes 2 * 2 = <<2*2=4>>4 dollars per day.
#### 4

***Answer (Full):***
Janet eats 3 eggs and bakes 4 eggs, so she has 16 - 3 - 4 = <<16-3-4=9>>9 eggs left.
She sells 9 eggs for $2 each, so she makes 9 * 2 = <<9*2=18>>18 dollars.
#### 18

Different Reasoning Path

**Question:** Raymond and Samantha are cousins. Raymond was born 6 years before Samantha. Raymond had a son at the age of 23. If Samantha is now 31, how many years ago was Raymond's son born?

**Answer:** Samantha is 31, ***so she is 31 - 6 = 25 years old***. Raymond is 25 - 23 = 2 years older than Samantha, so he is 25 years old .....

Unreasonable Statement

**Question:** A candle melts by 2 centimeters every hour that it burns. How many centimeters shorter will a candle be after burning from 1:00 PM to 5:00 PM?

**Answer:** The candle will be 4 centimeters shorter after 5:00 PM because it will be 4 hours x 2 centimeters = <<4*2=8>>8 centimeters shorter.
#### 4
**Remarks:** There is conflicting statement in reasoning, leading to the wrong end result.

Unreasonable Statement

The error happens in the middle and then propagate forwards towards the wrong end-results.
Correcting these wrong tokens (roughly 10% of them) recovers the Sparse models performance fully.

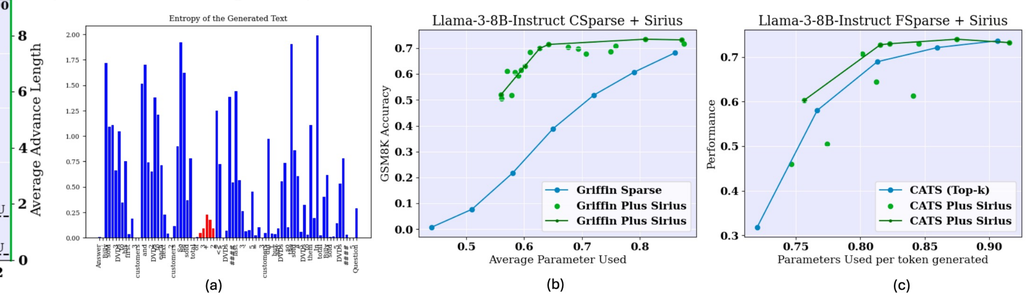
## Overview of Sirius: An Efficient Correction System for Contextual Sparsity




- Full Model Weights in GPU Memory (as Contextual Sparsity Requires)
- Sparse Model Signals **unreliable**
- Verification with **Long** Period (16+)
- During check, direct rewrite of KV Cache (Shared) + Interleave with the Correct Tokens
- Rejects unlikely tokens (**based on Confidence threshold**, not frequent on average accepting 15/16 tokens)
- The efficiency is further **boosted by building Efficient Decoding Tree** on the sparse model side

## Sparse Models Output not Trust-worthy + Sirius as a Compression Method

(a) Sparse Model often gets too confident when making the mistakes; (b) and (c) studies Sirius as a compression method against sparse methods;




## Experiment Results

Table 3: We show SIRIUS effectiveness and efficiency in the following table. We select GSM8K for Arithmetic Reasoning, CSQA for Commonsense Reasoning, and HumanEval for code generation. Under the "SIRIUS Perf. " column, A(B) is shown. A denotes the accuracy after SIRIUS correction in the dataset evaluated, while (B) represents the optimal treewidth selected under the current model dataset settings. Under the column of "AAL", X/Y is shown, where X is the AAL, while Y is the period.

**GSM8K**

| Model | Full Perf. | CSparse Perf. | CSparse Density | SIRIUS Perf. | AAL | Effective Density |
|---|---|---|---|---|---|---|
| Llama-3-8B-Instruct | 0.7536 | 0.3844 | 0.65 | 0.7051 (8) | 15.22/16 | 0.706 |
| Llama-3-8B | 0.4966 | 0.2085 | 0.65 | 0.4177 (8) | 15.29/16 | 0.703 |
| Llama-2-7B-Chat | 0.2403 | 0.1334 | 0.69 | 0.2244 (8) | 15.00/16 | 0.757 |
| Llama-2-7B | 0.1357 | 0.0758 | 0.69 | 0.1183 (6) | 15.87/16 | 0.715 |
| Llama-2-13B-Chat | 0.3548 | 0.2714 | 0.68 | 0.3381 (4) | 15.34/16 | 0.730 |
| Llama-2-13B | 0.2282 | 0.1759 | 0.68 | 0.2418 (1) | 15.34/16 | 0.730 |

| Model | Full Perf. | FSparse Perf. | FSparse Density | SIRIUS Perf. | AAL | Effective Density |
|---|---|---|---|---|---|---|
| Llama-3-8B-Instruct | 0.7536 | 0.5868 | 0.76 | 0.7278 (4) | 15.37/16 | 0.807 |
| Llama-3-8B | 0.4966 | 0.3199 | 0.76 | 0.4579 (2) | 15.03/16 | 0.825 |
| Llama-2-7B-Chat | 0.2403 | 0.1971 | 0.79 | 0.2388 (6) | 15.69/16 | 0.819 |
| Llama-2-7B | 0.1357 | 0.1137 | 0.79 | 0.1410 (4) | 15.91/16 | 0.807 |
| Llama-2-13B-Chat | 0.3548 | 0.3222 | 0.78 | 0.3533 (1) | 15.08/16 | 0.842 |
| Llama-2-13B | 0.2282 | 0.2191 | 0.78 | 0.2372 (4) | 15.92/16 | 0.797 |

**CSQA**

| Model | Full Perf. | CSparse Perf. | CSparse Density | SIRIUS Perf. | AAL | Effective Density |
|---|---|---|---|---|---|---|
| Llama-3-8B-Instruct | 0.7073 | 0.6470 | 0.58 | 0.7076 (8) | 14.76/16 | 0.657 |
| Llama-3-8B | 0.6437 | 0.5585 | 0.58 | 0.6429 (8) | 15.43/16 | 0.628 |
| Llama-2-7B-Chat | 0.6248 | 0.5200 | 0.62 | 0.6175 (8) | 15.07/16 | 0.683 |
| Llama-2-7B | 0.4742 | 0.4414 | 0.62 | 0.4742 (8) | 15.80/16 | 0.652 |
| Llama-2-13B-Chat | 0.6879 | 0.5536 | 0.61 | 0.6691 (4) | 11.43/12 | 0.674 |
| Llama-2-13B | 0.6109 | 0.5601 | 0.61 | 0.6060 (4) | 15.72/16 | 0.645 |

| Model | Full Perf. | FSparse Perf. | FSparse Density | SIRIUS Perf. | AAL | Effective Density |
|---|---|---|---|---|---|---|
| Llama-3-8B-Instruct | 0.7073 | 0.6158 | 0.72 | 0.7043 (8) | 15.66/16 | 0.753 |
| Llama-3-8B | 0.6437 | 0.533 | 0.72 | 0.6388 (1) | 15.00/16 | 0.786 |
| Llama-2-7B-Chat | 0.6248 | 0.6167 | 0.75 | 0.6380 (4) | 15.09/16 | 0.811 |
| Llama-2-7B | 0.4742 | 0.4717 | 0.75 | 0.5012 (6) | 15.89/16 | 0.771 |
| Llama-2-13B-Chat | 0.6879 | 0.533 | 0.74 | 0.6691 (4) | 14.30/16 | 0.846 |
| Llama-2-13B | 0.6109 | 0.5700 | 0.74 | 0.5864 (4) | 15.72/16 | 0.770 |

**HumanEval**

| Model | Full Perf. | CSparse Perf. | CSparse Density | SIRIUS Perf. | AAL | Effective Density |
|---|---|---|---|---|---|---|
| Llama-3-8B-Instruct | 0.561 | 0.207 | 0.65 | 0.524 (8) | 14.67/16 | 0.733 |
| Llama-3-8B | 0.262 | 0.067 | 0.65 | 0.243 (8) | 15.10/16 | 0.691 |
| Llama-2-7B-Chat | 0.140 | 0.067 | 0.69 | 0.159 (8) | 10.88/12 | 0.789 |
| Llama-2-7B | 0.116 | 0.079 | 0.69 | 0.128 (8) | 14.84/16 | 0.765 |
| Llama-2-13B-Chat | 0.189 | 0.122 | 0.68 | 0.171 (8) | 11.12/12 | 0.762 |
| Llama-2-13B | 0.262 | 0.067 | 0.68 | 0.244 (8) | 15.10/16 | 0.741 |

| Model | Full Perf. | FSparse Perf. | FSparse Density | SIRIUS Perf. | AAL | Effective Density |
|---|---|---|---|---|---|---|
| Llama-3-8B-Instruct | 0.561 | 0.457 | 0.76 | 0.616 (6) | 15.42/16 | 0.804 |
| Llama-3-8B | 0.262 | 0.189 | 0.76 | 0.298 (6) | 15.54/16 | 0.797 |
| Llama-2-7B-Chat | 0.140 | 0.134 | 0.79 | 0.165 (6) | 15.27/16 | 0.841 |
| Llama-2-7B | 0.116 | 0.116 | 0.79 | 0.165 (6) | 15.86/16 | 0.810 |
| Llama-2-13B-Chat | 0.189 | 0.146 | 0.78 | 0.183 (6) | 15.34/16 | 0.827 |
| Llama-2-13B | 0.246 | 0.233 | 0.78 | 0.259 (4) | 15.85/16 | 0.801 |

## Efficient Implementation + Hardware Speedup

Table 3: Performance and Speedup Ratios on GSM8K-COT with Different Hardware Configurations.

| Settings | ACC | A40 | Ratio | L40 | Ratio | A100 | Ratio | H100 | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| **CSparse** | 0.3601 | 20.7 ms | 0.66 | 15.6 ms | 0.67 | 9.6 ms | 0.72 | 6.6 | 0.76 |
| **Sirius** | 0.7127 | 24.1 ms | 0.78 | 18.2 ms | 0.78 | 11.1 ms | 0.83 | 7.7 ms | 0.88 |
| **Full** | 0.7612 | 30.9 ms | 1.0 | 23.2 ms | 1.0 | 13.3 ms | 1.0 | 8.6 ms | 1.0 |

Table 4: Llama-3-70B-Instruct with Offloading.

| Settings | Sparse | Sirius | Full |
|---|---|---|---|
| **Performance** | 0.7407 | 0.8719 | 0.9014 |
| **Latency (s)** | 3.57 s | 3.68 s | 5.72 s |
| **Ratio to Full** | 0.6241 | 0.6434 | 1.0 |

Offloading (PCIe bandwidth 25G/s)

**Checkout the project page for more details and Code!**

For laptop users search for:
https://infini-ai-lab.github.io/Sirius/