Oregon State University

COLLEGE OF ENGINEERING

Phoenix: Enabling Sparse Fine-tuning for Foundation Model Downstream Tasks on Cerebras

Yale

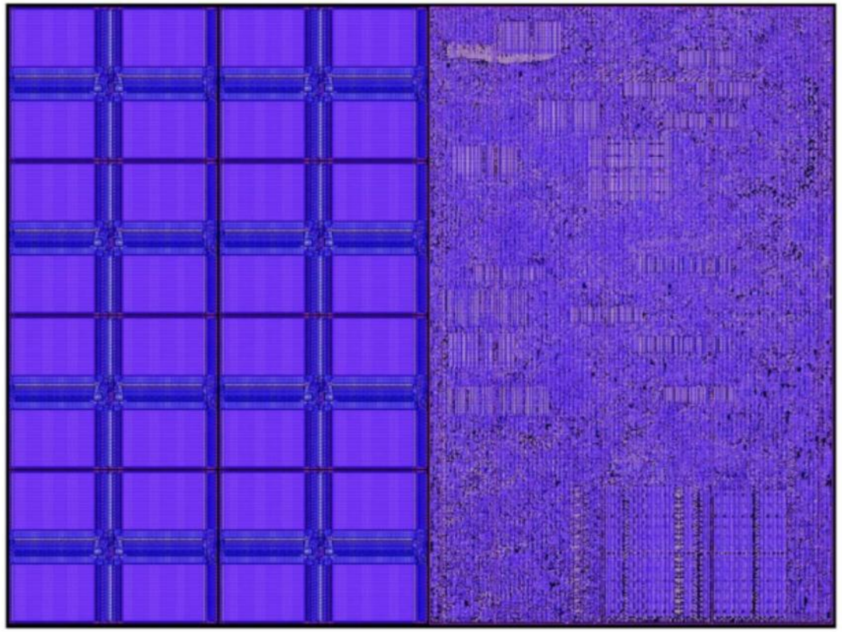
Haoyu Zheng (*OSU*), Linghao Song (*Yale*), Murali Emani (*ANL*)
Wenqian (Wendy) Dong (*OSU, Presenter*)

Argonne NATIONAL LABORATORY

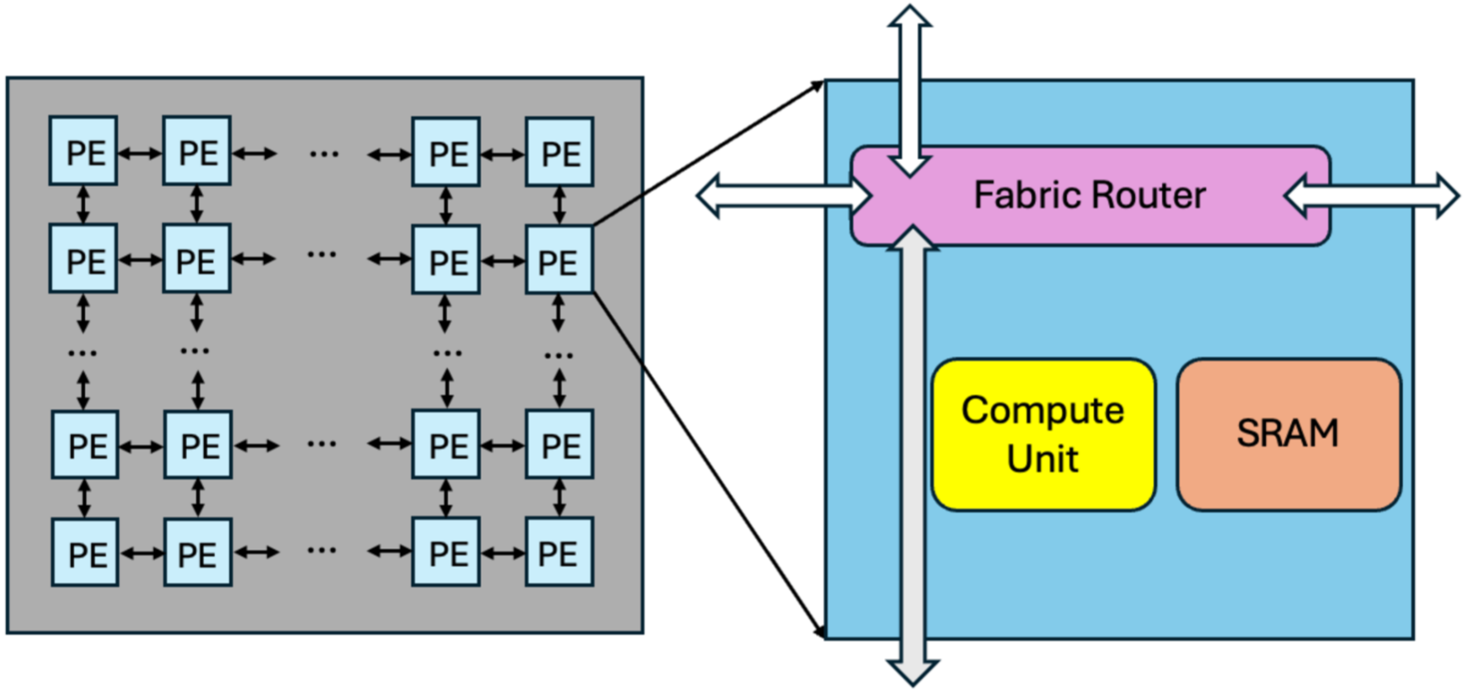
Yale

Cerebras: Wafer Scale Engine

A Wafer Scale Engine (WSE) is a type of computer chip designed to accelerate artificial intelligence and high-performance computing workloads.



The Cerebras core physical design: 50% of the area is static random-access memory (SRAM) and 50% of the area is logic.



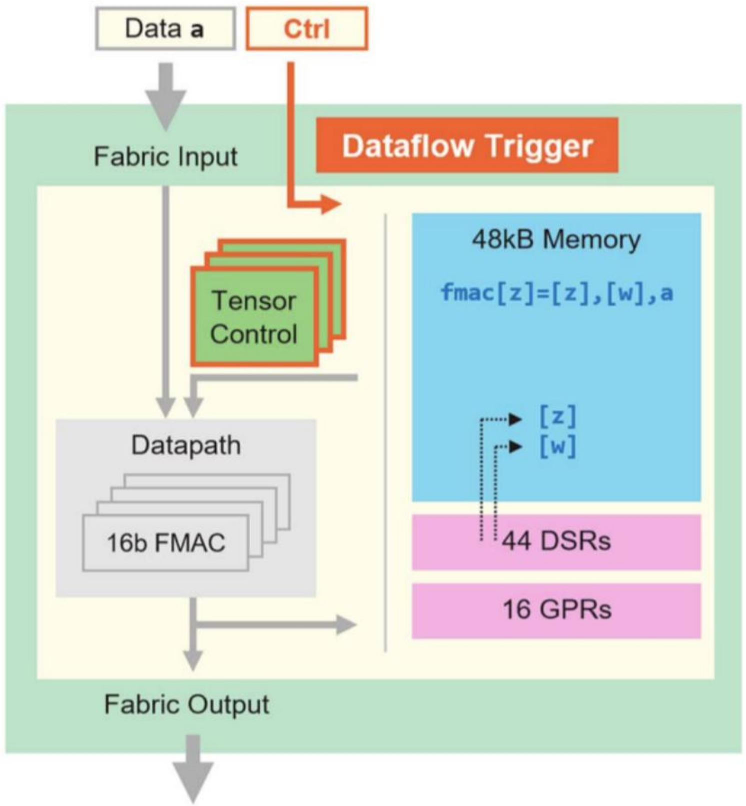
Each core includes local SRAM, compute unit, and a fabric router for direct inter-core communication within a 2D mesh network

Cerebras Architecture is Designed for Sparse Architecture

Oregon State University College of Engineering

- Fine-grained dataflow cores
  - Triggers compute only for non-zero data
- High bandwidth memory
  - Enables full datapath performance
- High bandwidth interconnect
  - Enables low overhead reductions

Only architecture capable of accelerating all types of sparsity, including **dynamic and unstructured sparsity**.



Low-Rank Adaptation (LoRA)

Oregon State University College of Engineering

- Only retrain few parameters for Down-stream Adaptation

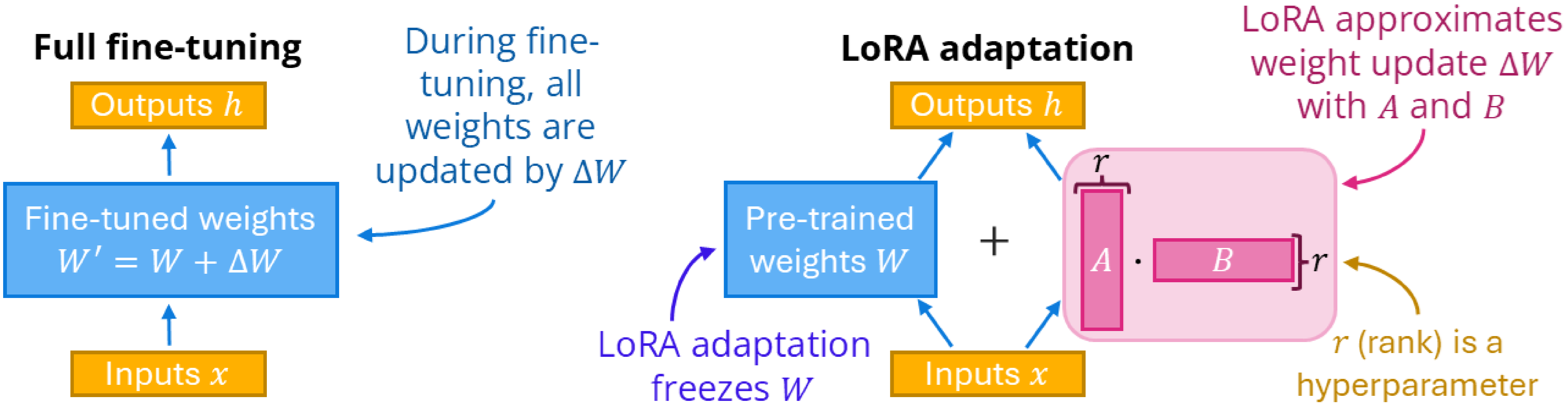


Image credit: https://oneminutenlp.substack.com/p/low-rank-adaptation

Low-Rank Adaptation (LoRA)

Oregon State University
College of Engineering

- Only retrain few parameters for Down-stream Adaptation

	Method	# of Trainable Params	E2E (BLEU)	DART (BLEU)	WebNLG (BLEU-U/S/A)
	GPT-2 M (Fine-Tune)	354.92M	68.2	46.0	30.4/**63.2**/47.6
	GPT-2 M (Adapter)	0.37M	66.3	42.4	45.1/54.5/50.2
	GPT-2 M (Prefix)	0.35M	69.7	45.7	44.1/63.1/54.4
	GPT-2 M (LoRA)	0.35M	**70.4**±.1	**47.1**±.2	**46.7**±.4/62.1±.2/**55.3**±.2
	GPT-2 L (Fine-Tune)	774.03M	68.5	46.5	41.7/**64.6**/54.2
	GPT-2 L (Adapter)	0.88M	69.1±.1	45.7±.1	**49.8**±.0/61.1±.0/56.0±.0
	GPT-2 L (Prefix)	0.77M	70.3	46.5	47.0/64.2/56.4
	GPT-2 L (LoRA)	0.77M	**70.4**±.1	**47.5**±.1	48.4±.3/**64.0**±.3/**57.0**±.1

0.22% reduction!

Phoenix Framework

Oregon State University
College of Engineering

An end-to-end framework that explore the benefit of unstructured sparsity on Cerebras for LLMs' finetuning and inference

Targets:

(1) Sparse Inference Efficiency:

- Enable true FLOPs reduction at inference time using unstructured sparsity.

(2) Cerebras Hardware Utilization:

- Leverage Cerebras CS-2's unique support for unstructured sparse computation.

(3) End-to-End Sparse Tuning Pipeline:

- Support sparsity from pruning to deployment with no conversion steps.

Challenges:

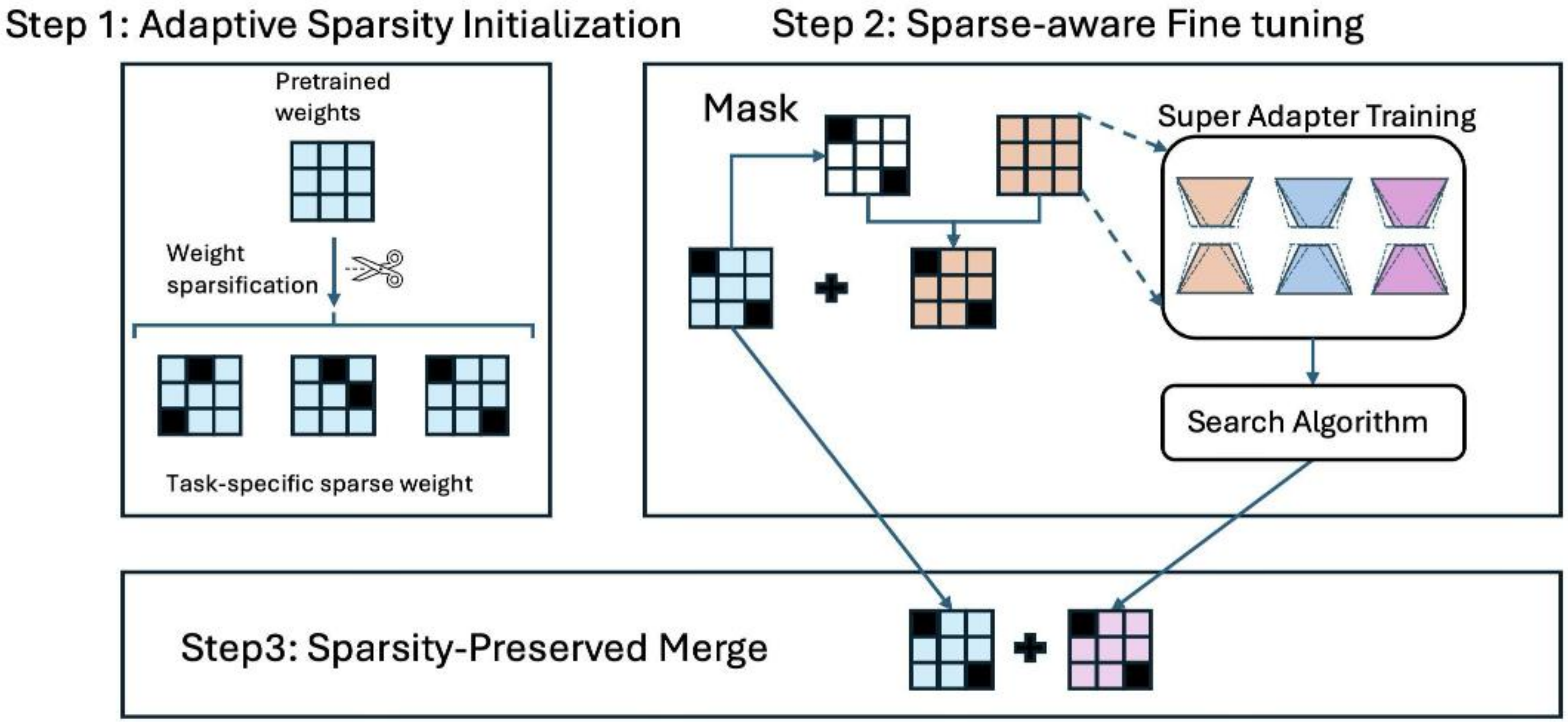
(1) LoRA Merge Breaks Sparsity:

Standard LoRA adapters (dense matrices) overwrite zeroed weights when merged.

(2) Dense Fine-Tuning after Sparse Pretraining:

Methods like SPDF reintroduce density during downstream adaptation.

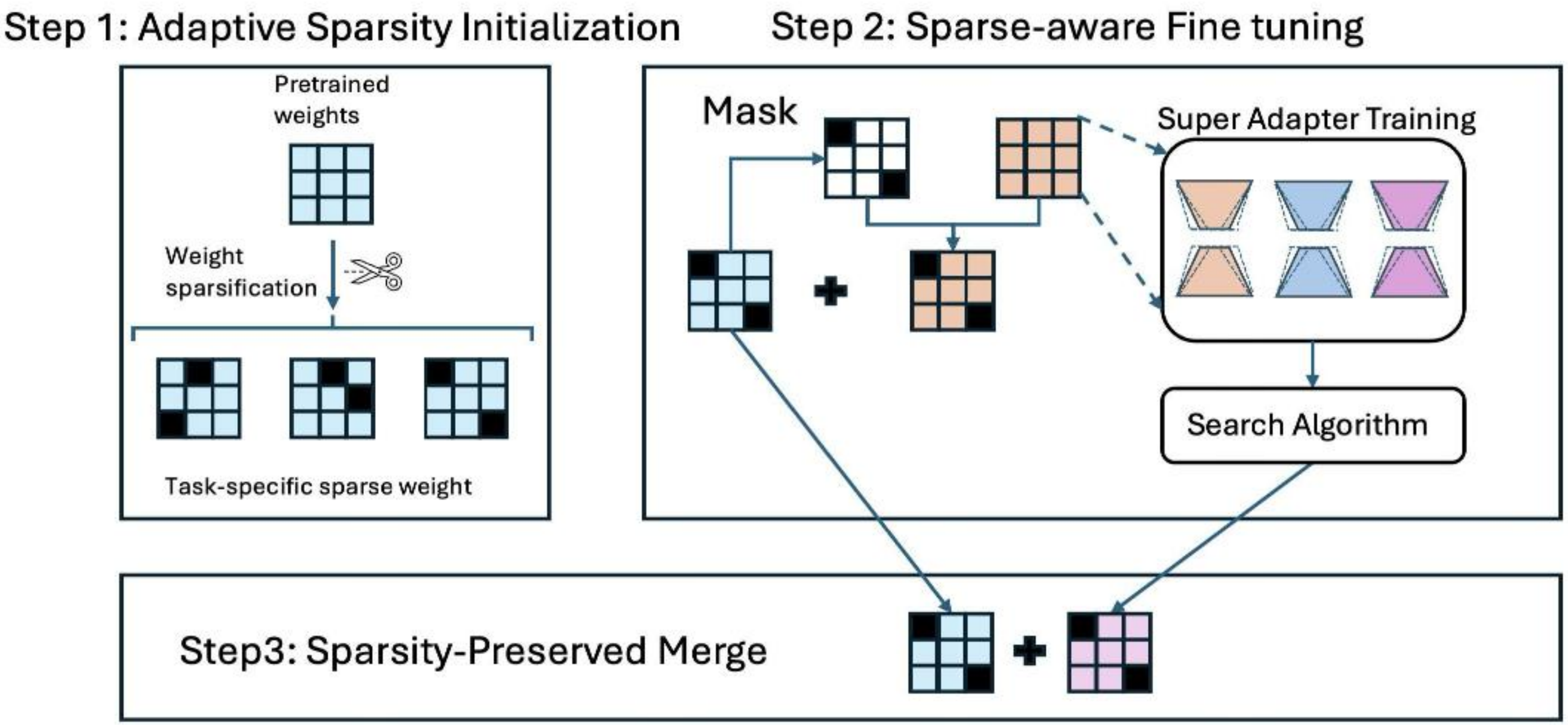
Phoenix - Enabling LLM Sparsity on Cerebras



Three steps in Phoenix: (1) Adaptive Sparsity Initialization, where task-specific sparse weights are generated from a pre-trained dense model; (2) Sparsity-Aware Fine-Tuning, Phoenix applies a binary mask to get the sparse structure derived from the sparsified pre-trained model weights and uses a search algorithm to select the high performance subadapter configuration; and (3) Sparsity-Preserved Merge, where the fine-tuned adapter is integrated back into the sparse model without breaking the original sparsity pattern.

Phoenix - Enabling LLM Sparsity on Cerebras

Oregon State University College of Engineering



Evaluation

Oregon State University
College of Engineering

Models: LLaMA-3-8B and Mistral-7B-v0.3

Downstream task:

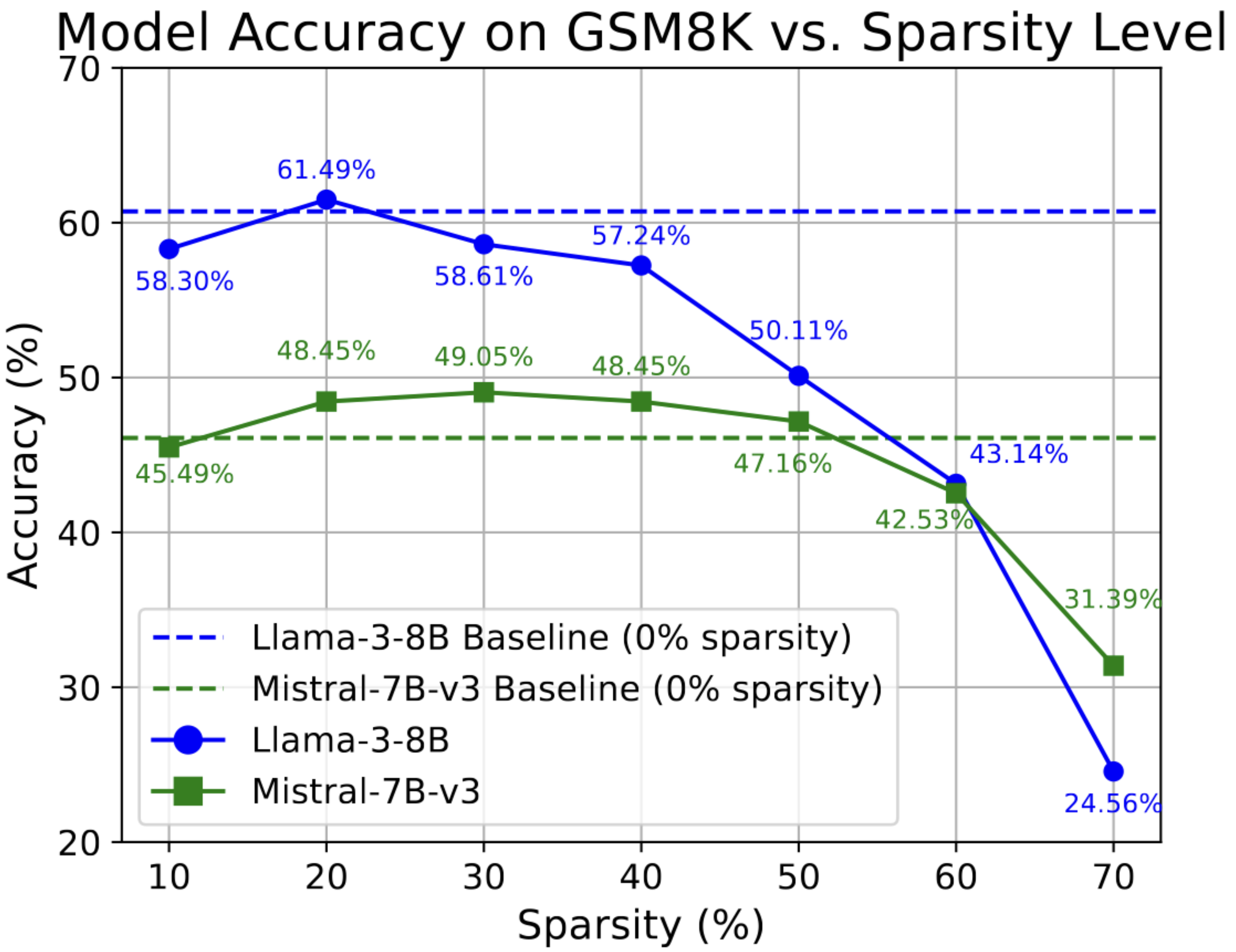
(1) Grade School Math: We benchmark performance on the GSM8K dataset, a challenging arithmetic reasoning task requiring multi-step problem solving

(2) Instruction-Tuned Math Reasoning: This includes a trio of math-focused datasets — GSM8K, Math Word Problems (MAWPS), and SVAMP

Specification	Cerebras CS-2	Nvidia A100 80GB
Chip Size	46225 mm^2	826 mm^2
Memory	40GB on-chip SRAM	80GB off-chip HBM
Memory Bandwidth	22 PB/s	2 TB/s
Compute Capacity	850000 cores	6912 CUDA cores
Process	7nm (TSMC)	7nm (TSMC)

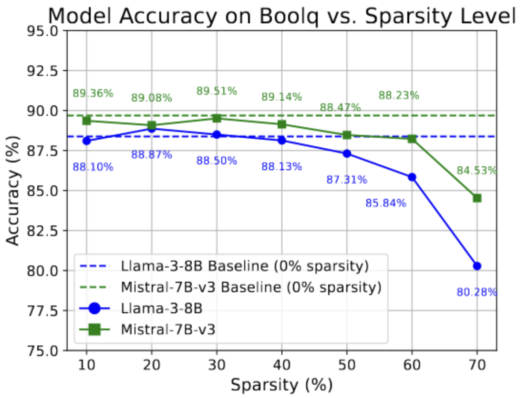
Performance – Accuracy (1)

Oregon State University
College of Engineering

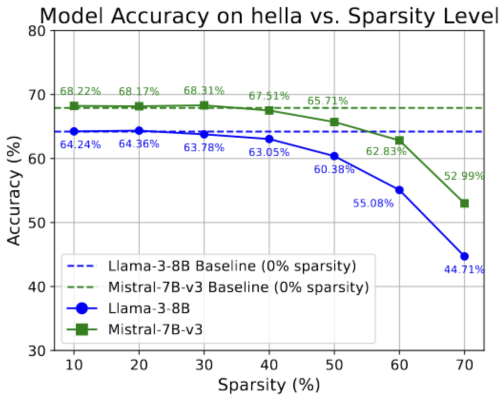


Performance – Accuracy (2)

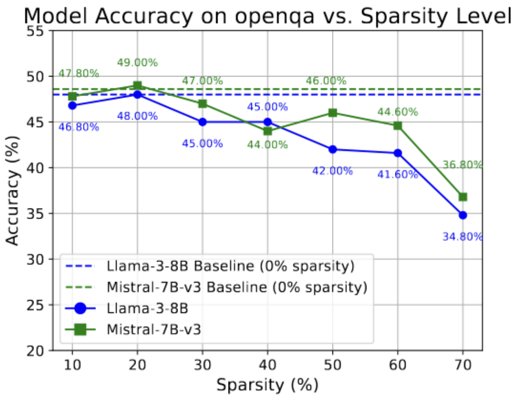
Oregon State University
College of Engineering



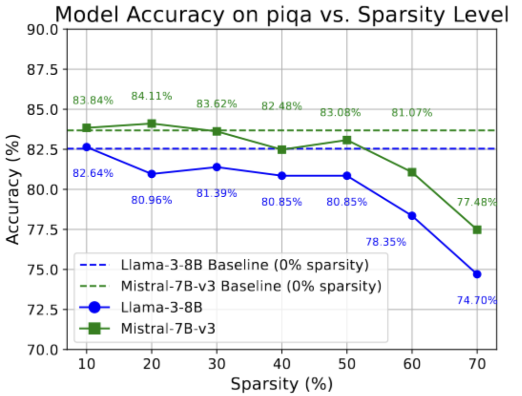
(a) BoolQ



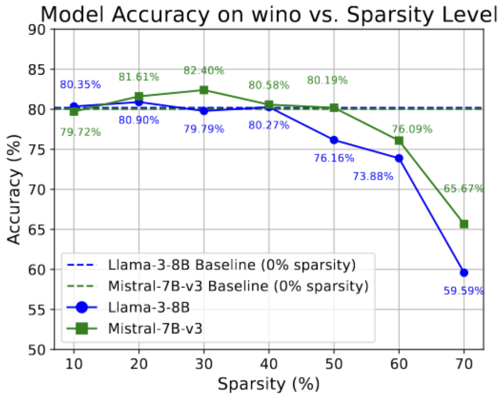
(b) HellaSwag



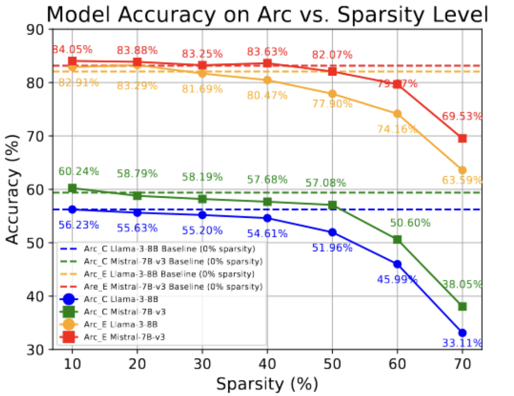
(c) OBQA



(d) PIQA



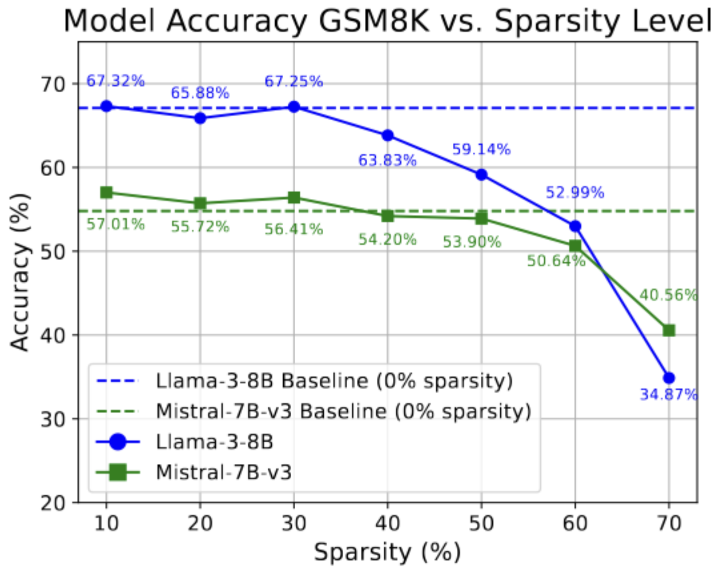
(e) WinoGrande



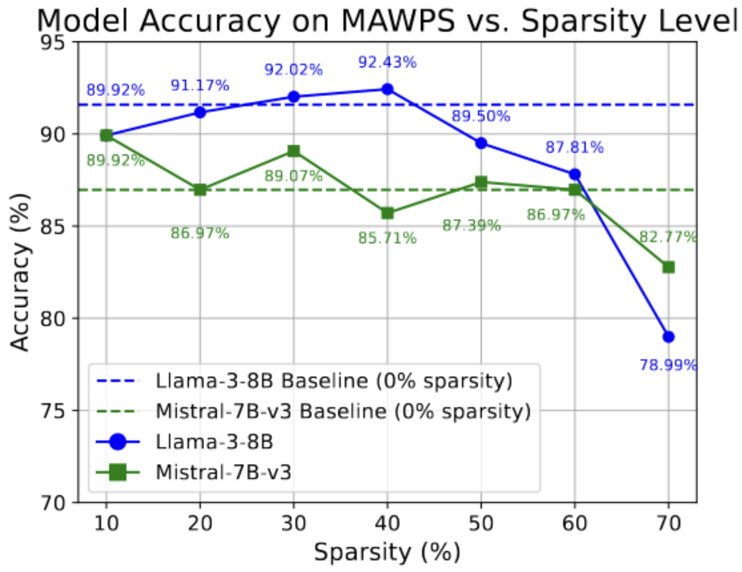
(f) ARC-Easy and ARC-Challenge

Performance – Accuracy (3)

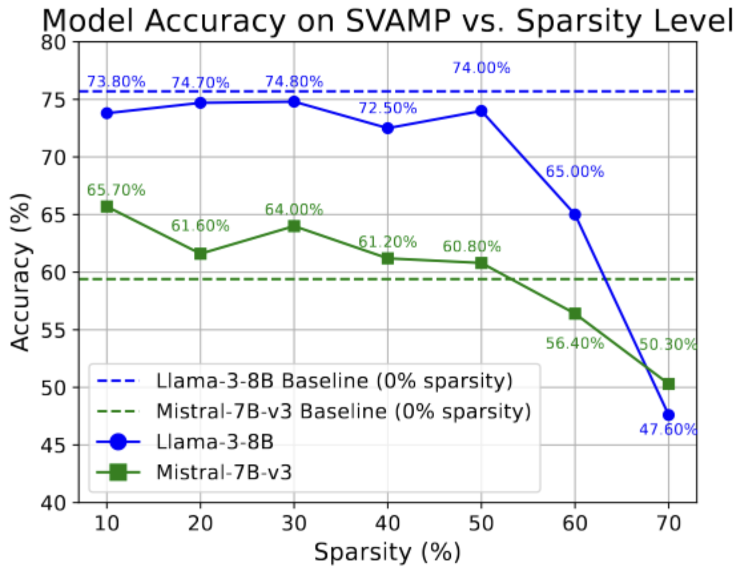
Oregon State University
College of Engineering



(a) GSM8K Accuracy

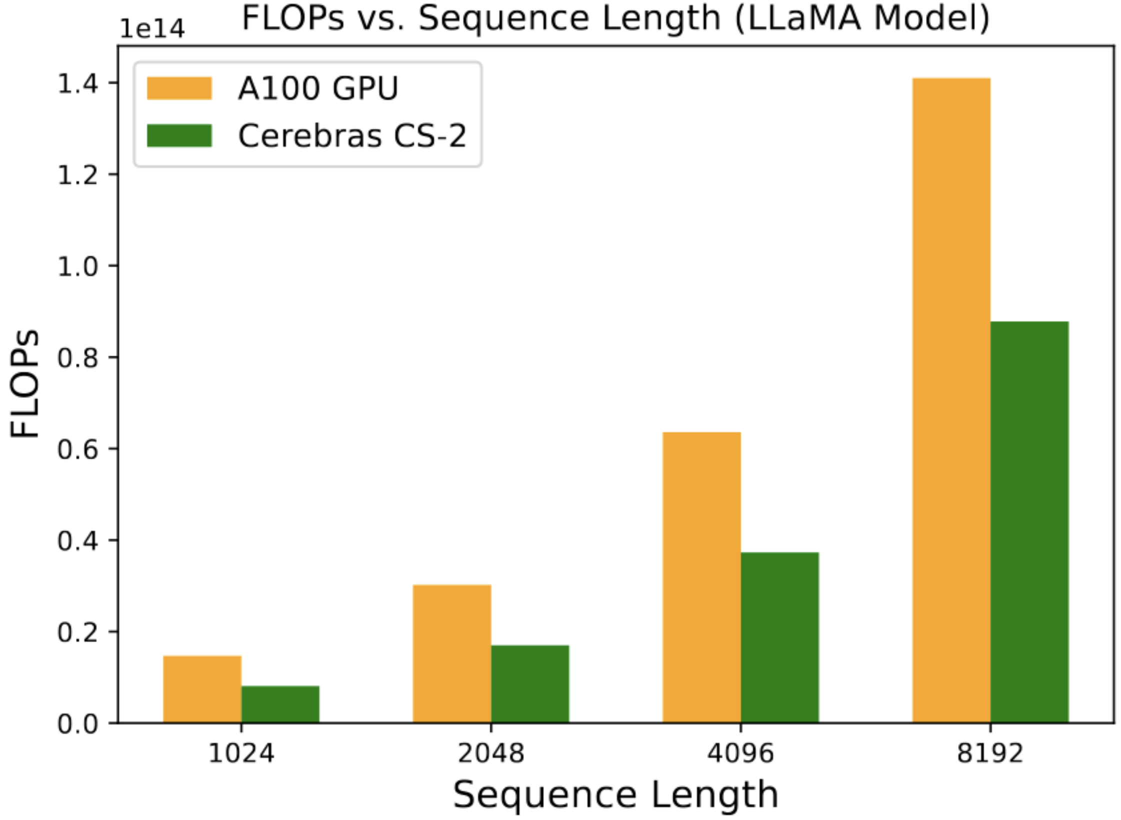


(b) MAWPS Accuracy

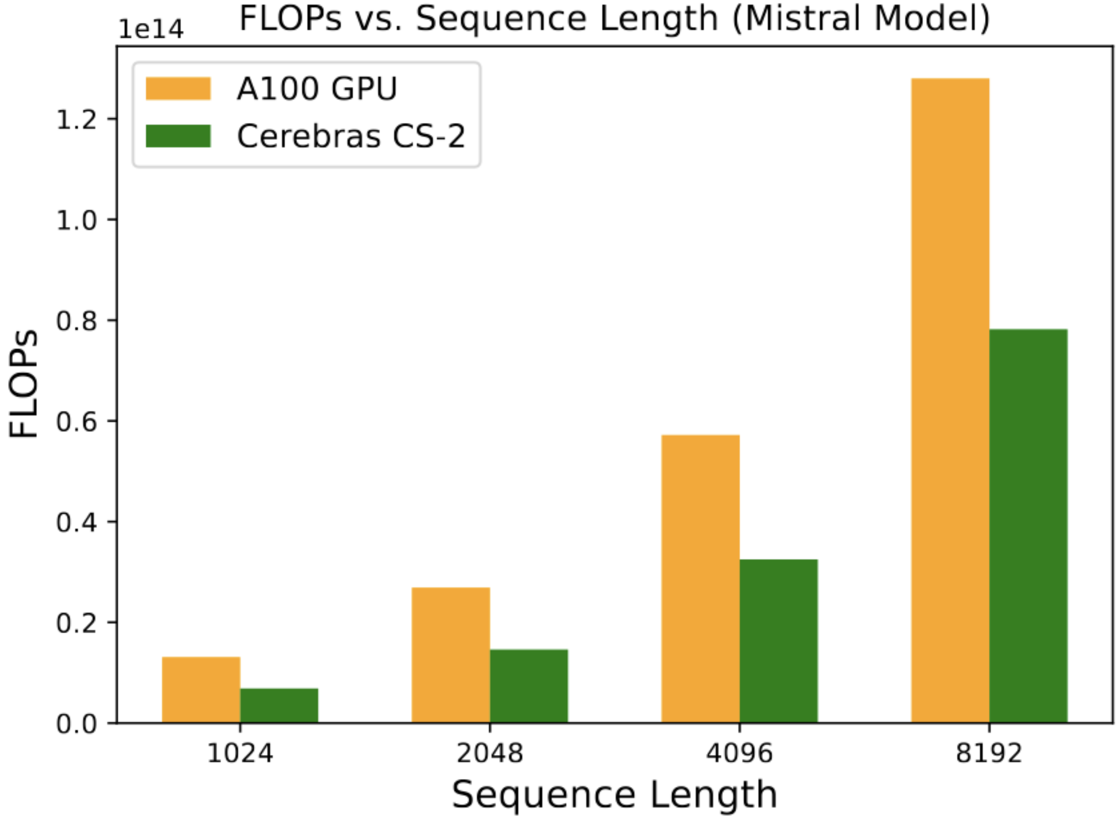


(c) SVAMP Accuracy

Performance – FLOPs Reduction



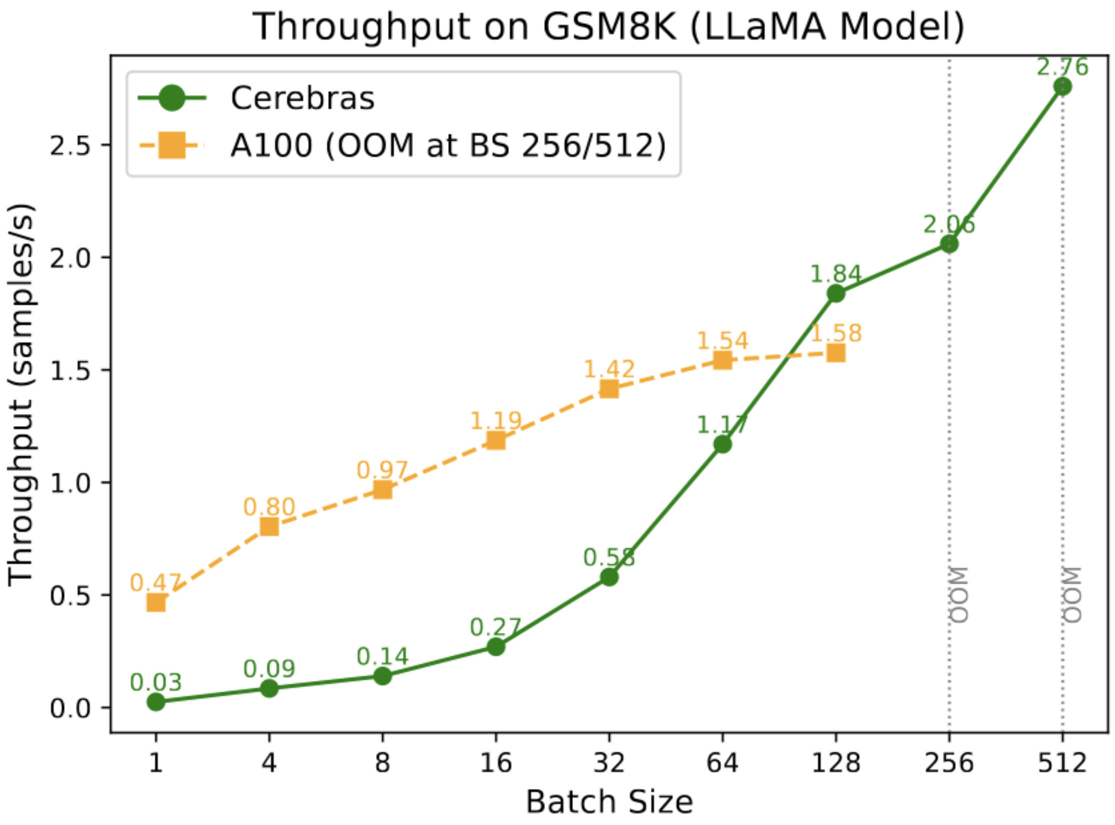
(a) Inference FLOPs reduction of Llama Model



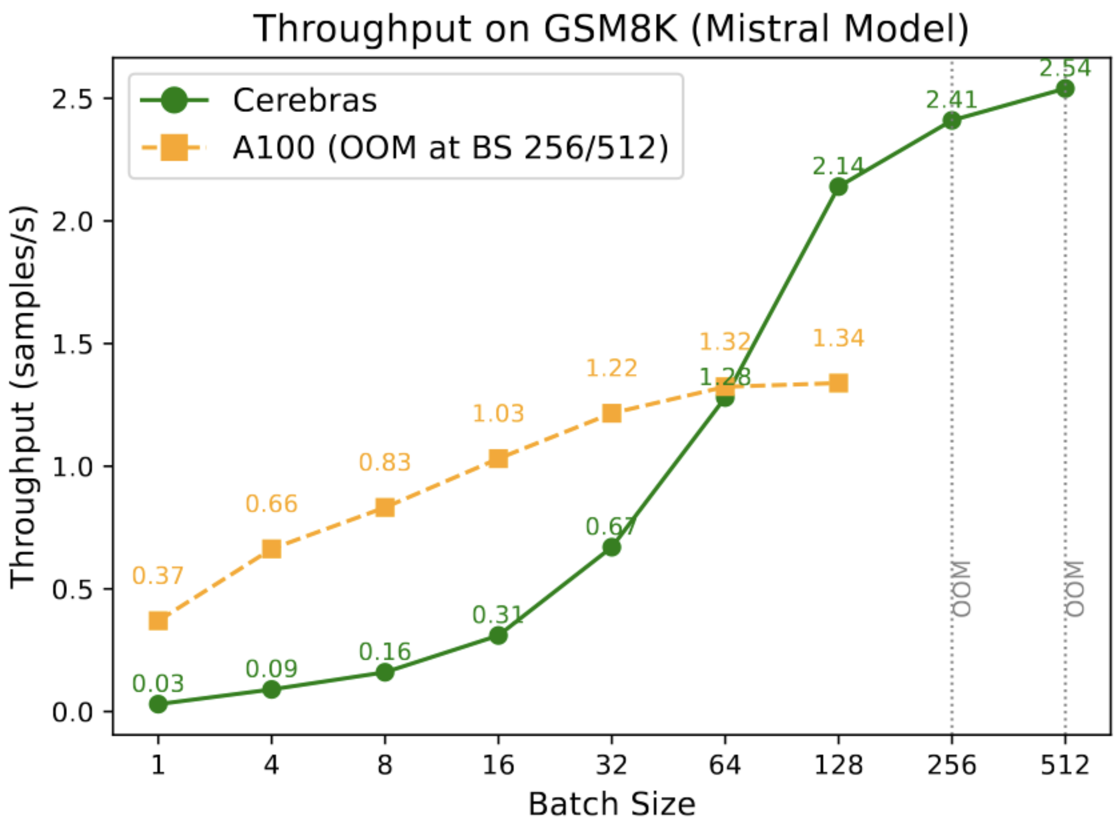
(b) Inference FLOPs reduction of Mistral Model

Performance – Speedup

Oregon State University
College of Engineering



(a) Throughput Comparison of Llama Model



(b) Throughput Comparison of Mistral Model

Oregon State University
College of Engineering

Thank you, and Question?

Wenqian “Wendy” Dong <wenqian.dong@oregonstate.edu>