GEORGE MASON UNIVERSITY

FT2: First-Token-Inspired Online Fault Tolerance on Critical Layers for Generative Large Language Models

Yu Sun[§], Zhu Zhu[§], Cherish Mulpuru[§],

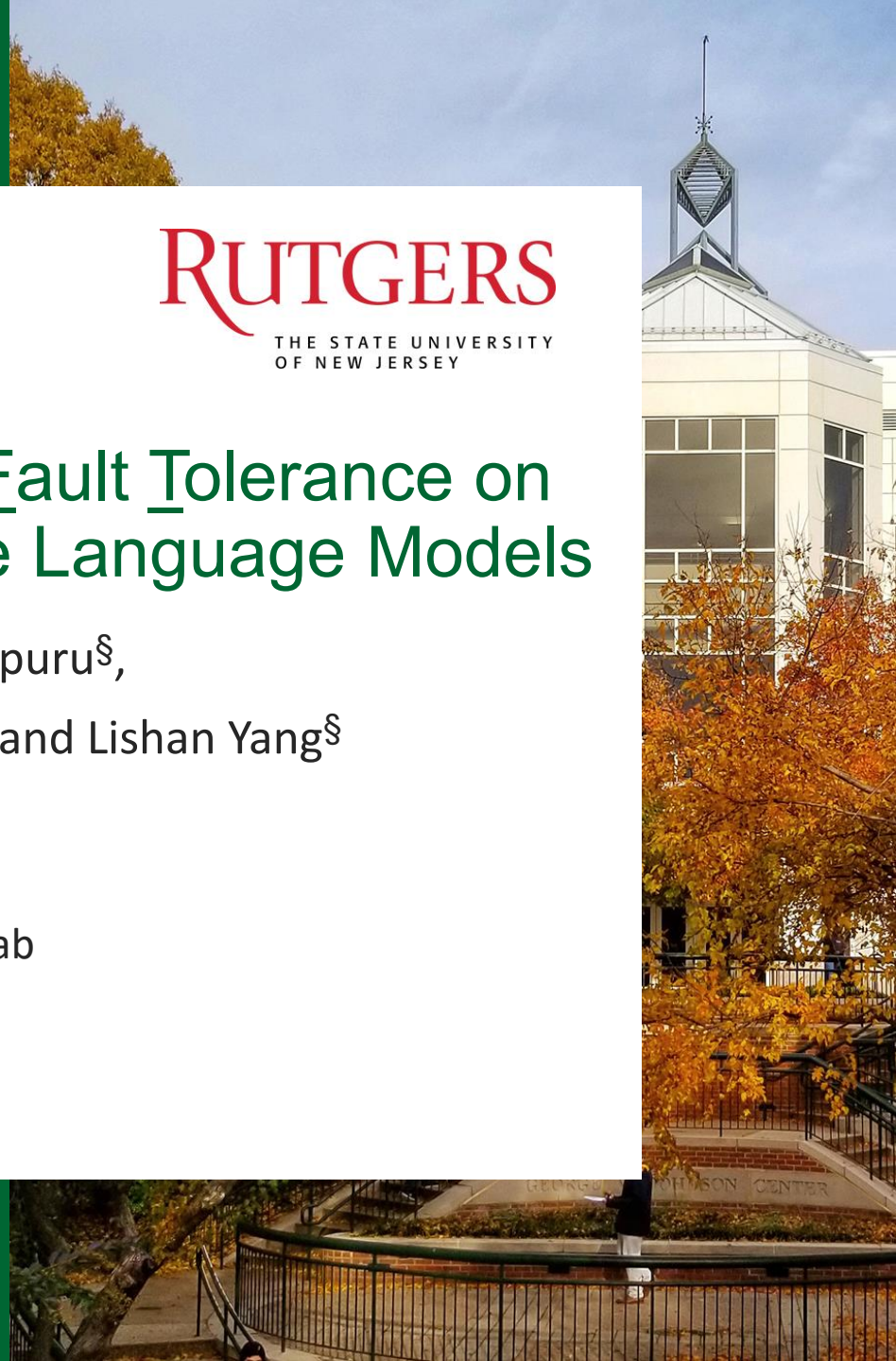
Roberto Gioiosa[‡], Zhao Zhang[†], Bo Fang[‡], and Lishan Yang[§]

[§]George Mason University

[‡]Pacific Northwest National Lab

[†]Rutgers University

Email: ysun23@gmu.edu

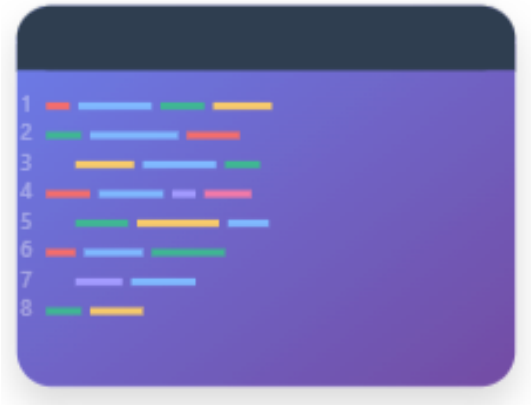


Generative Large Language Models (LLM)

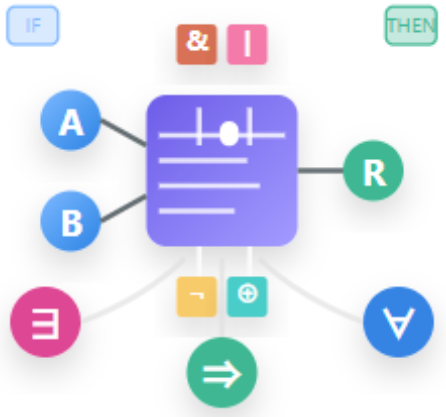
Question Answering

Translation

ChatGPT



Coding

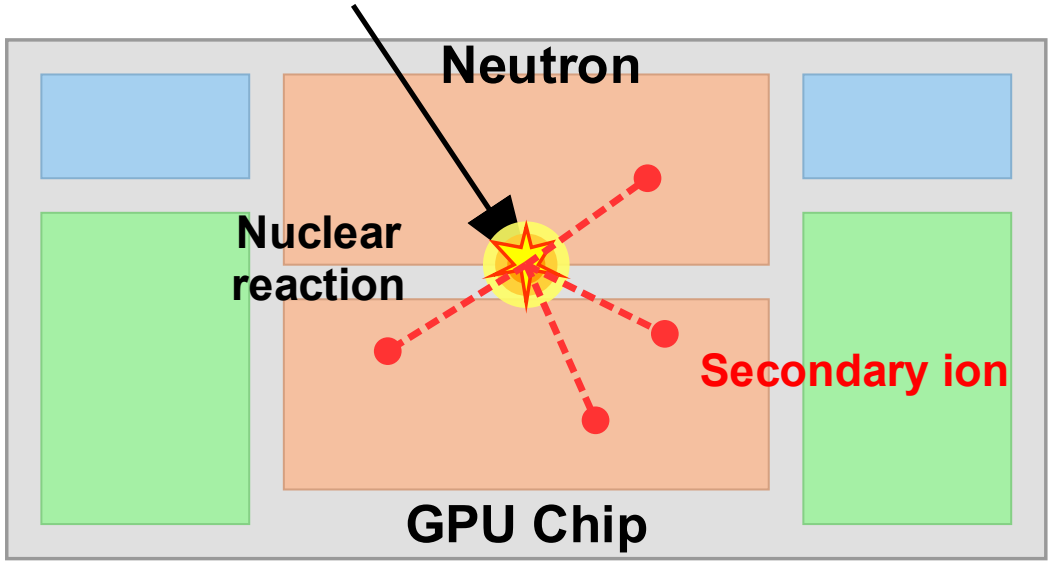


Logic Reasoning

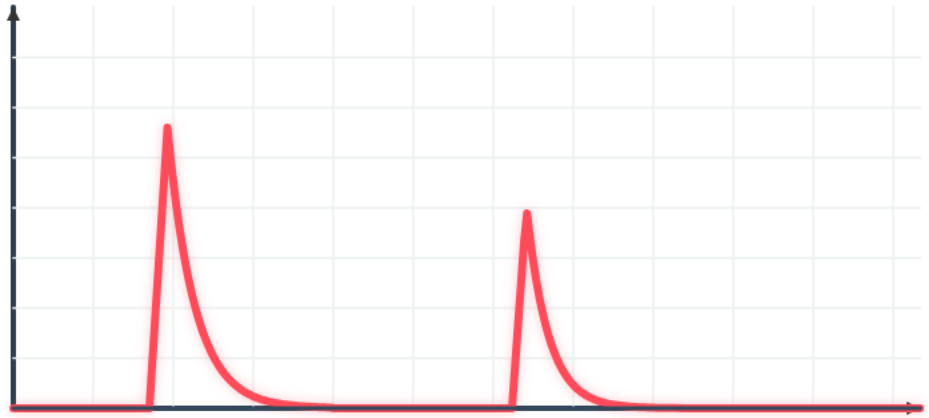
Claude

Soft Error

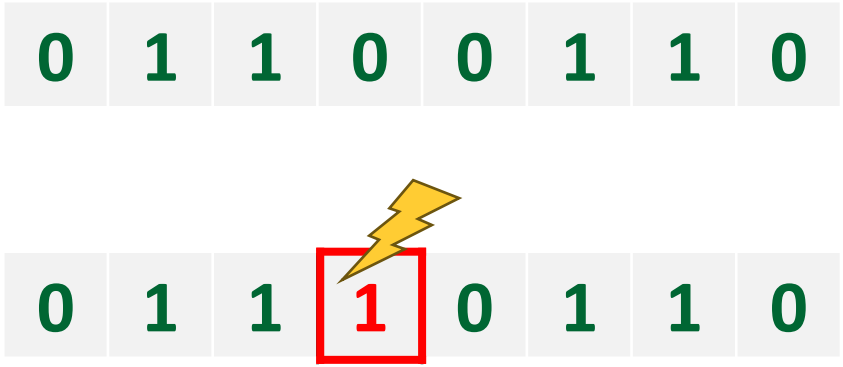
Cosmic ray



Neutrons hit silicon die



Current pulses



Bit flips

Soft Error

There are several consequences…

Text: As of August 2010, Victoria had 1,548 public schools, 489 Catholic schools and 214 independent schools…… Victoria has about 63,519 full-time teachers.

Question: How many full-time teachers does Victoria have?

Reference: 63,519

Fault-free Answer: Victoria has about 63,519 full-time teachers.

Answer with Fault Injection: The number of full-time teachers in Victoria is 63,519. → **Benign**

Answer with Fault Injection: The number is 1548.

SDC! (Silent Data Corruption)

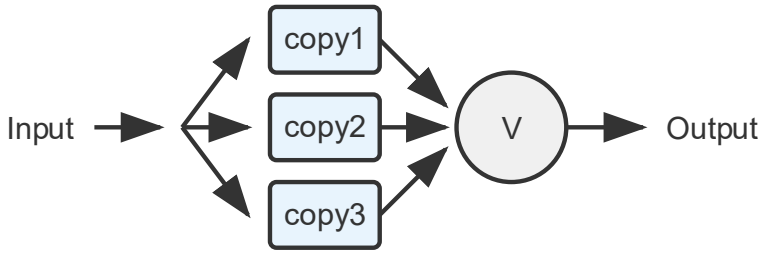
Harmful and hard to detect

Existing Protection

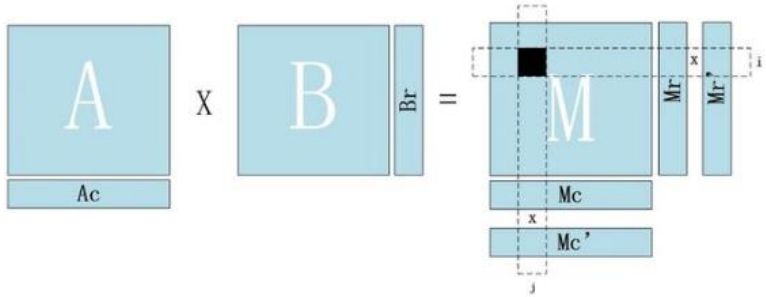
TMR: Triple Modular Redundancy

ABFT: Algorithm-Based Fault Tolerance

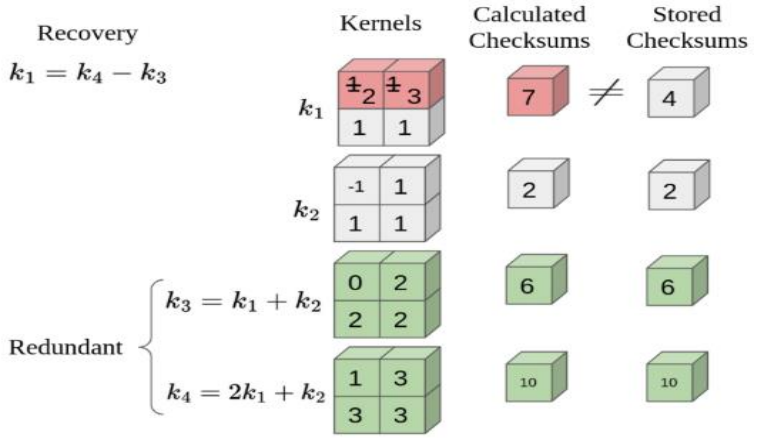
TMR



ABFT



Structural Coding

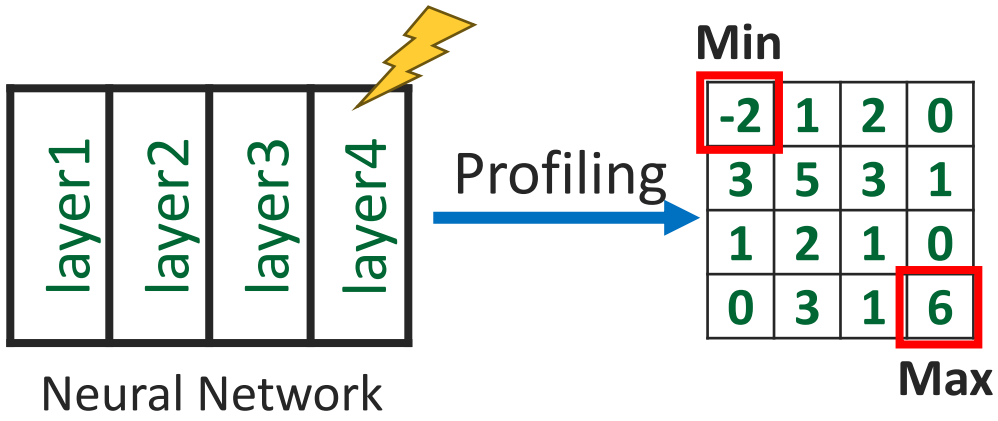


High overhead for LLMs → Larger

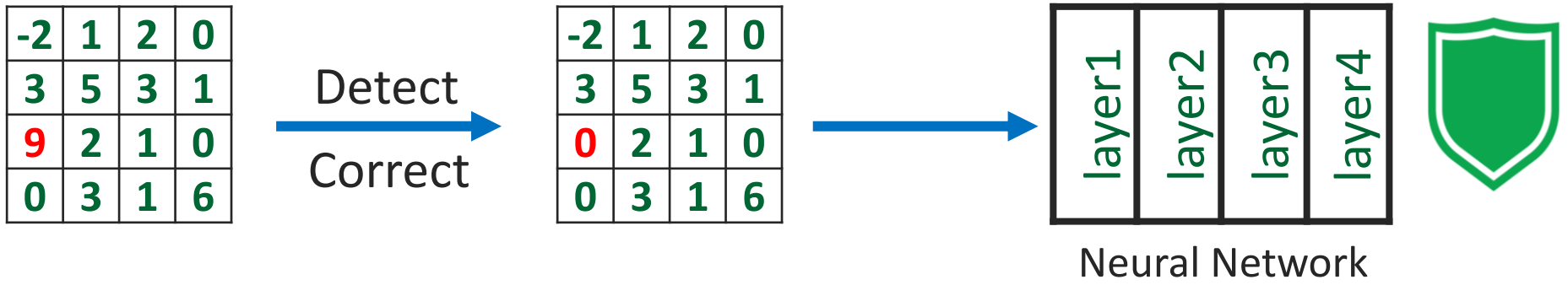
Low overhead solution?

Ranger!

Ranger



Layer	Max	Min
1	-4	3
2	-3	8
3	-1	5
4	-2	6



Two limitations to apply on LLMs

Insufficient protection
Require bound profiling

Limitation 1:

Insufficient Protection

Different layers

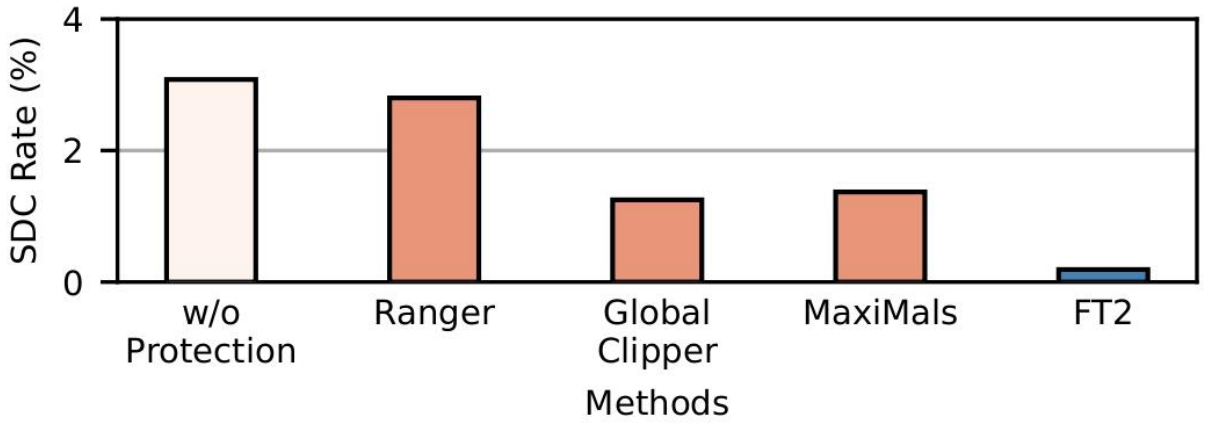
CNN
conv
conv
pooling
ReLU
FC
FC

LLM
k_proj
q_proj
v_proj
o_proj
fc1
fc2

?

If choose layer unwisely

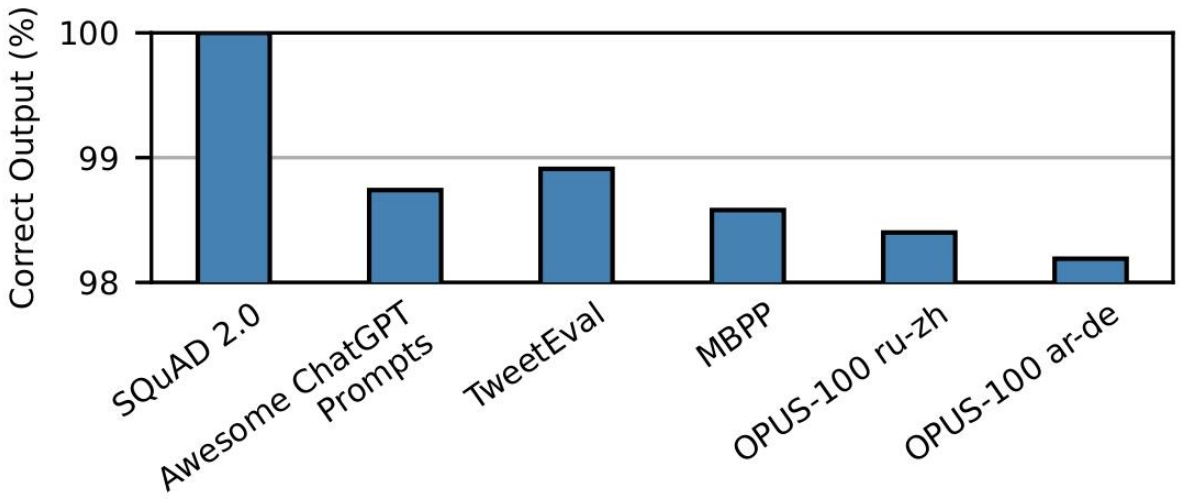
Undesirable SDC rate reduction



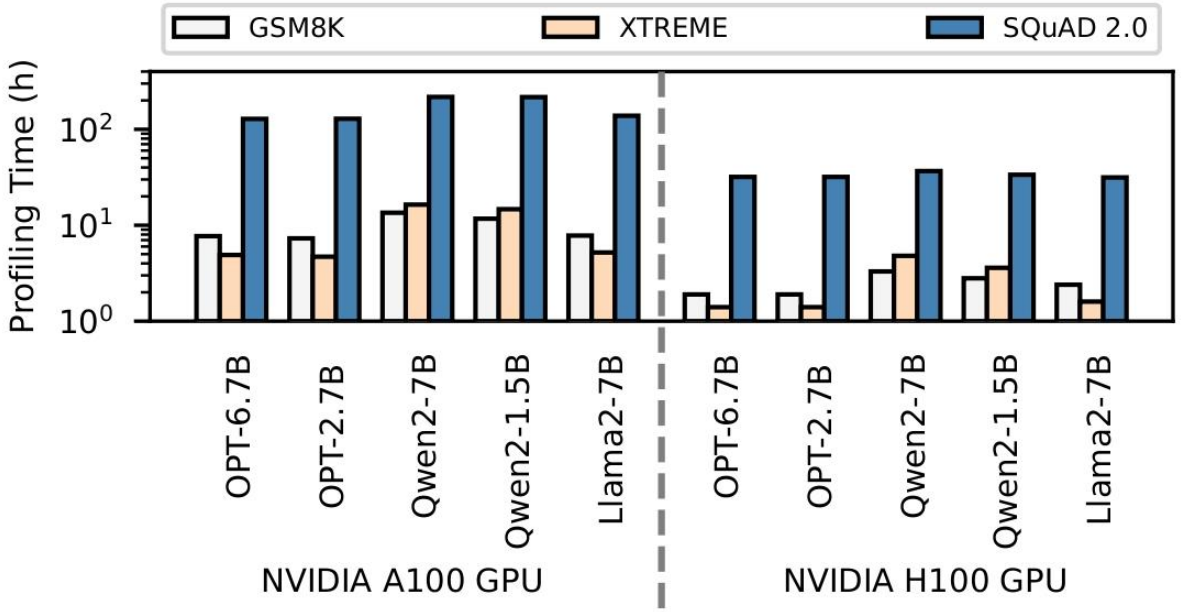
Limitation 2:

Bound Profiling

- Lack of training datasets



- High profiling cost



For these two limitations to apply Ranger on LLMs

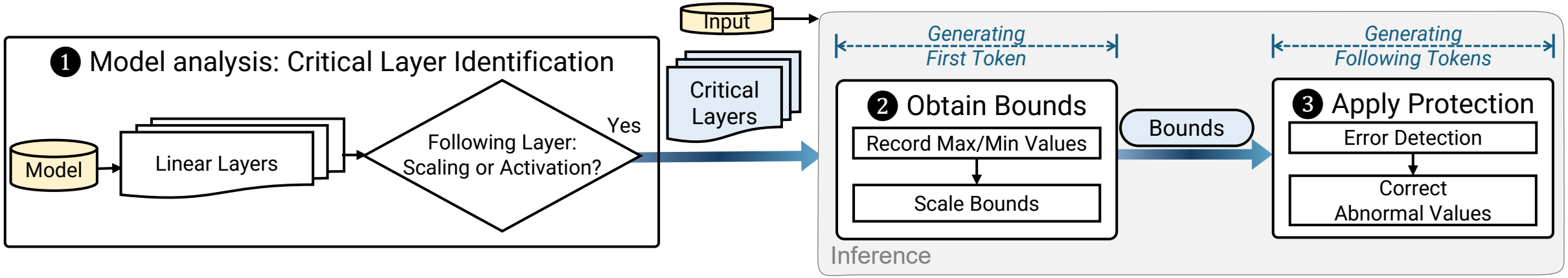
How to overcome?

- *Lack of training dataset*
- *High profiling cost*

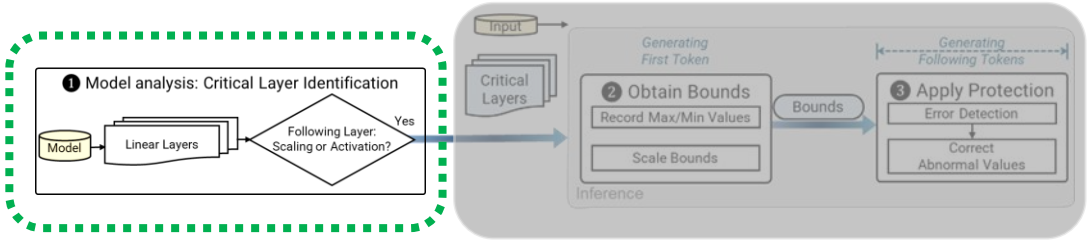
Our Methodology: FT2

FT2: First-Token-Inspired Online Fault Tolerance on Critical Layers

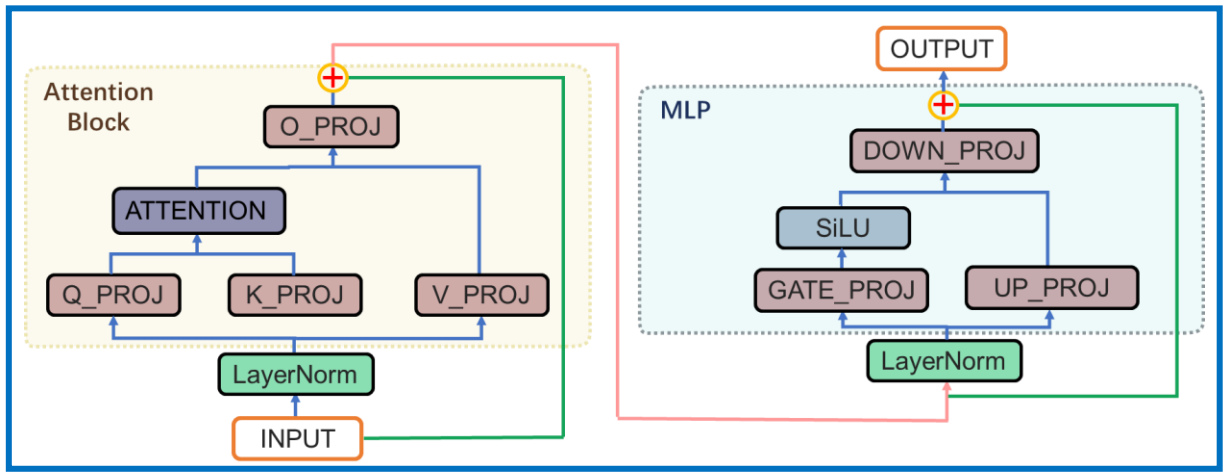
- *Better Protection*
- *No profiling required*



Identify Critical Layers



Critical layers: High SDC rate if not specifically protected

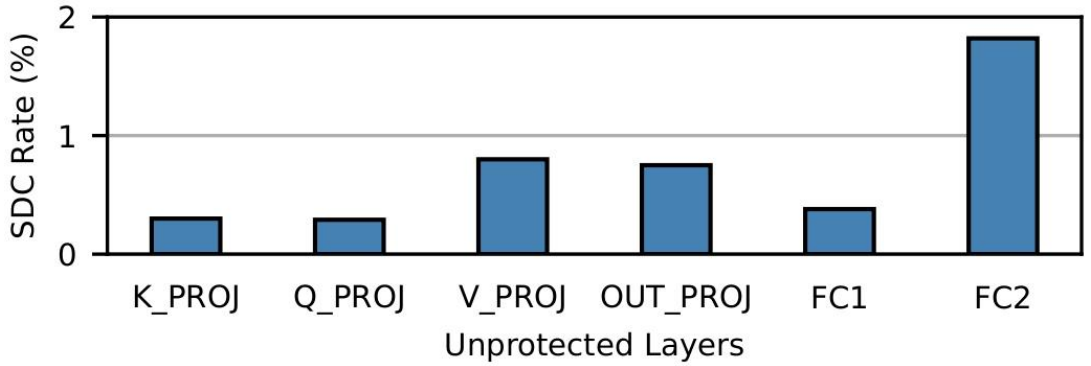


× **N**

Protect all → High overhead

Which layers are critical?

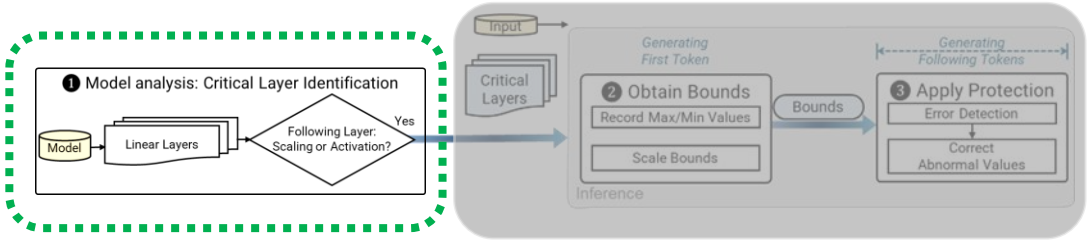
- Fault injection experiments



Model: GPTJ-6B
Dataset: SQuAD2.0
Fault model: 1-bit

Why Critical?

Identify Critical Layers



There are two types of abnormal values caused by bit flips

① Extreme value

16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	
0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0x2448

1 x 2^{-6} x 1.071 = 0.01673

16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	
0	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0x6448

1 x 2^{10} x 1.071 = **1096**

② Not a number (NaN)

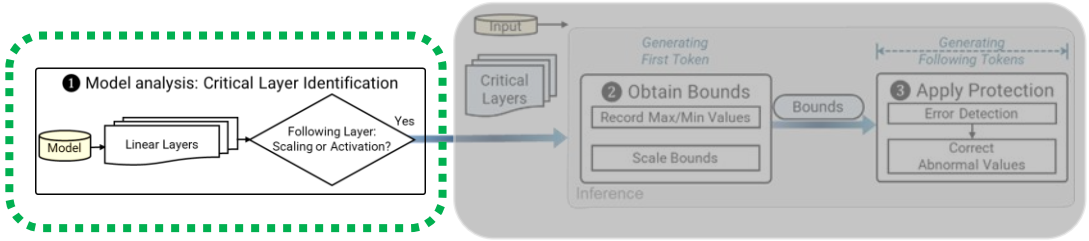
16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	
0	0	1	1	1	1	1	0	0	1	0	1	1	0	0	0	0x3E58

1 x 2^{0} x 1.586 = 1.586

16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	
0	1	1	1	1	1	1	0	0	1	0	1	1	0	0	0	0x7E58

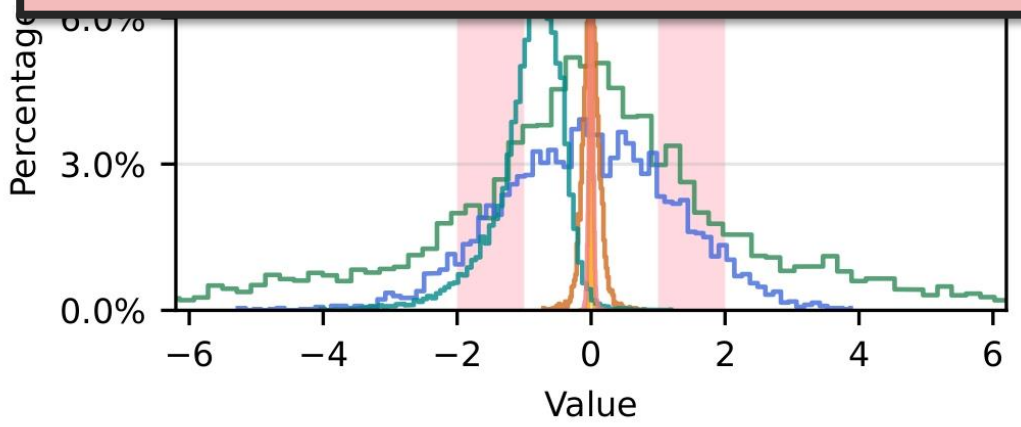
1 x 2^{16} x 1.586 = **NaN**

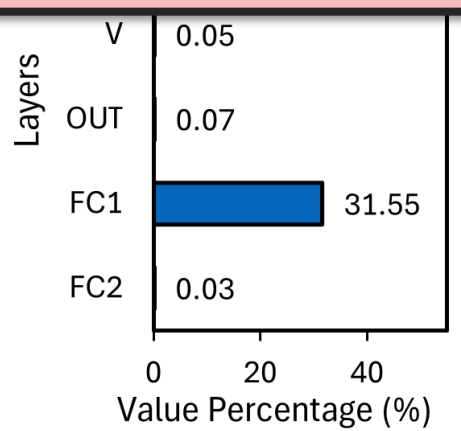
Identify Critical Layers

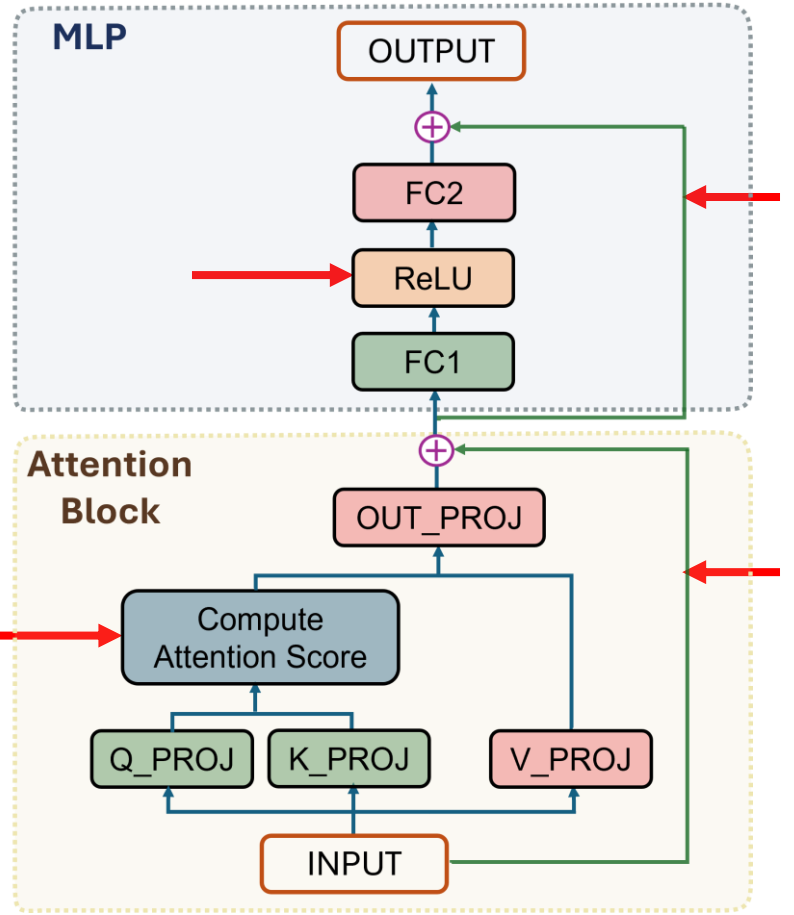


- Different neuron value distributions
  - Scaling operations

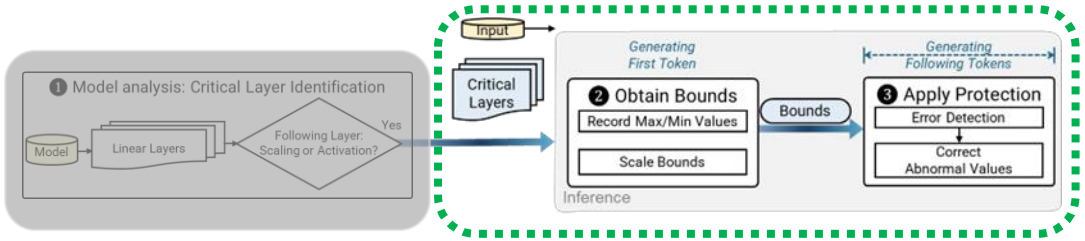
A heuristic: *a layer is critical if no scaling operation or activation layer is behind.*



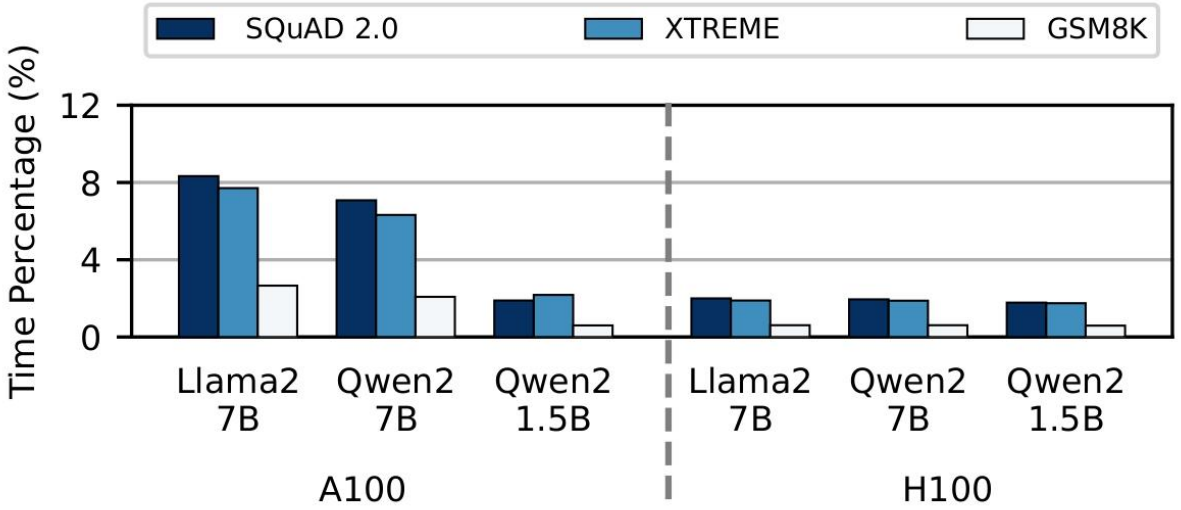




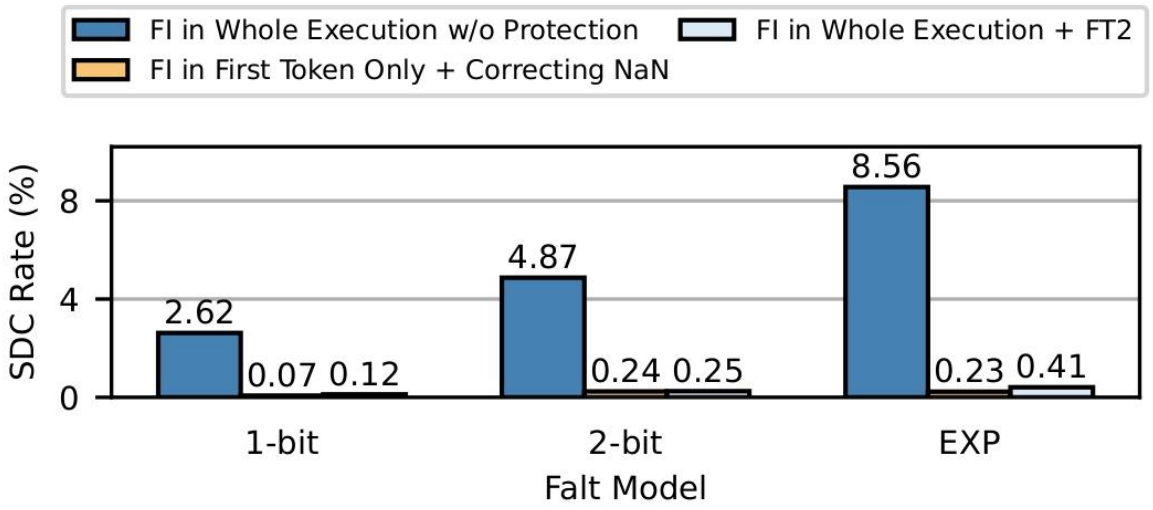
Obtain Bounds Online



- Obtain bounds from the first token generation
  - Input → ***Longer***
  - Information → ***More***
  - Bounds → ***More accurate***
- The impact of not protecting this process is negligible



The percentage of execution time is low



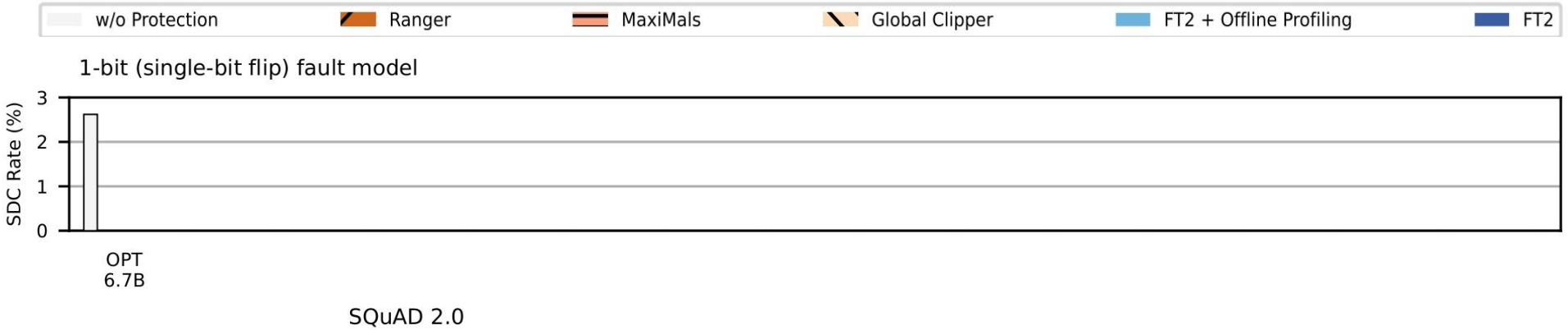
The resilience is high

Evaluation: Experimental Set-up

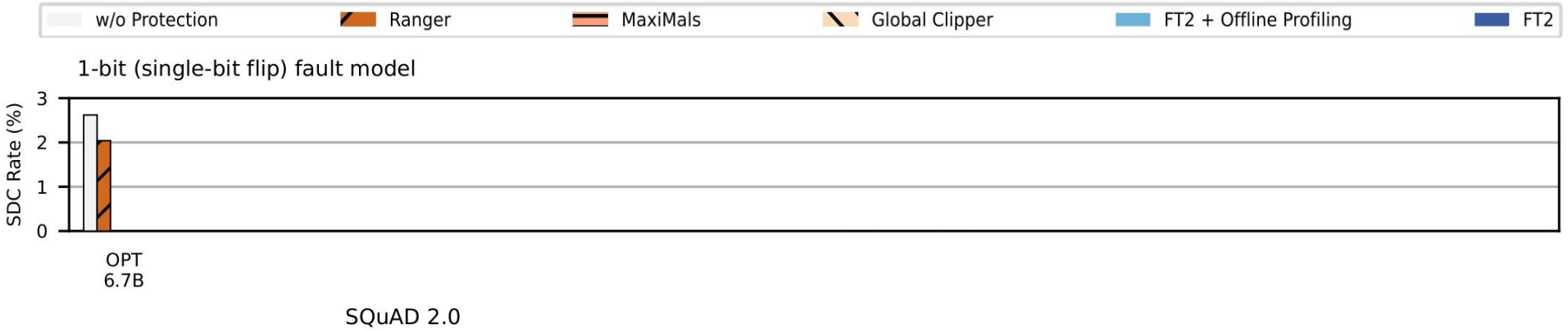
Model Name	# of Parameters	Task Type
OPT-6.7B	6.66B	QA
OPT-2.7B	2.65B	QA
GPTJ-6B	6.05B	QA
Llama2-7B	6.74B	QA/MATH
Vicuna-7B	6.74B	QA
Qwen2-7B	7.62B	QA/MATH
Qwen2-1.5B	1.54B	QA

- ***7 models covering 2 architectures***
- ***3 datasets from 2 tasks***
- ***Datatype***: FP16 and FP32
- ***2 GPU configurations***: NVDIA A100 and H100 GPU
- ***3 fault models***: 1-bit, 2-bit and EXP

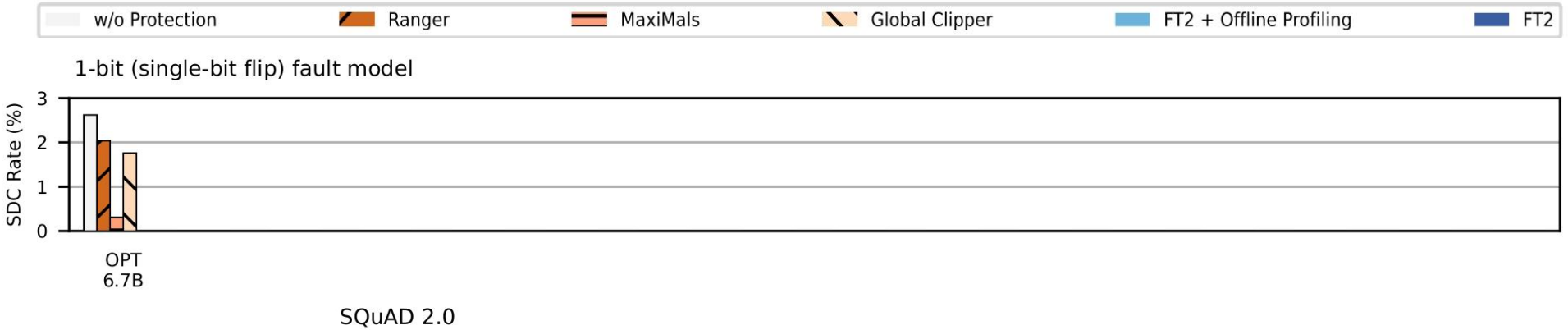
Overall Evaluation



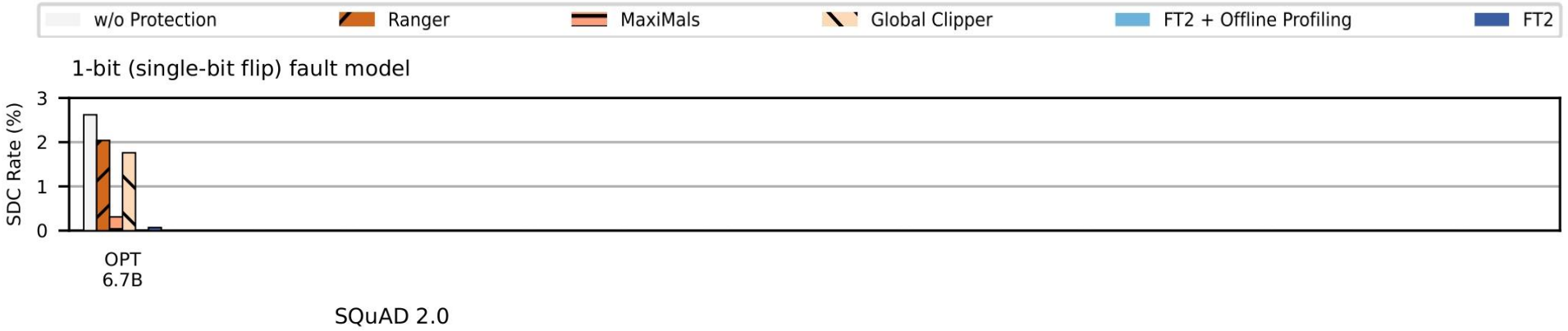
Overall Evaluation



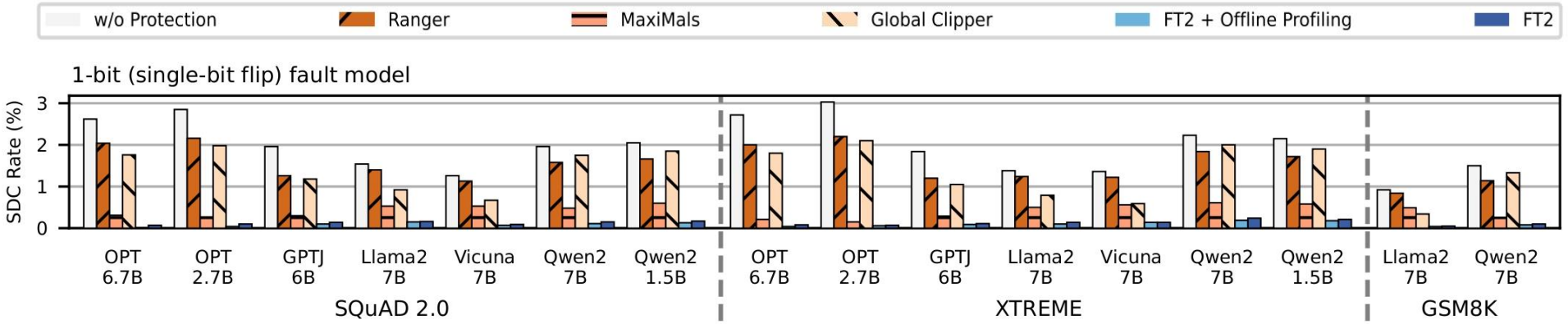
Overall Evaluation



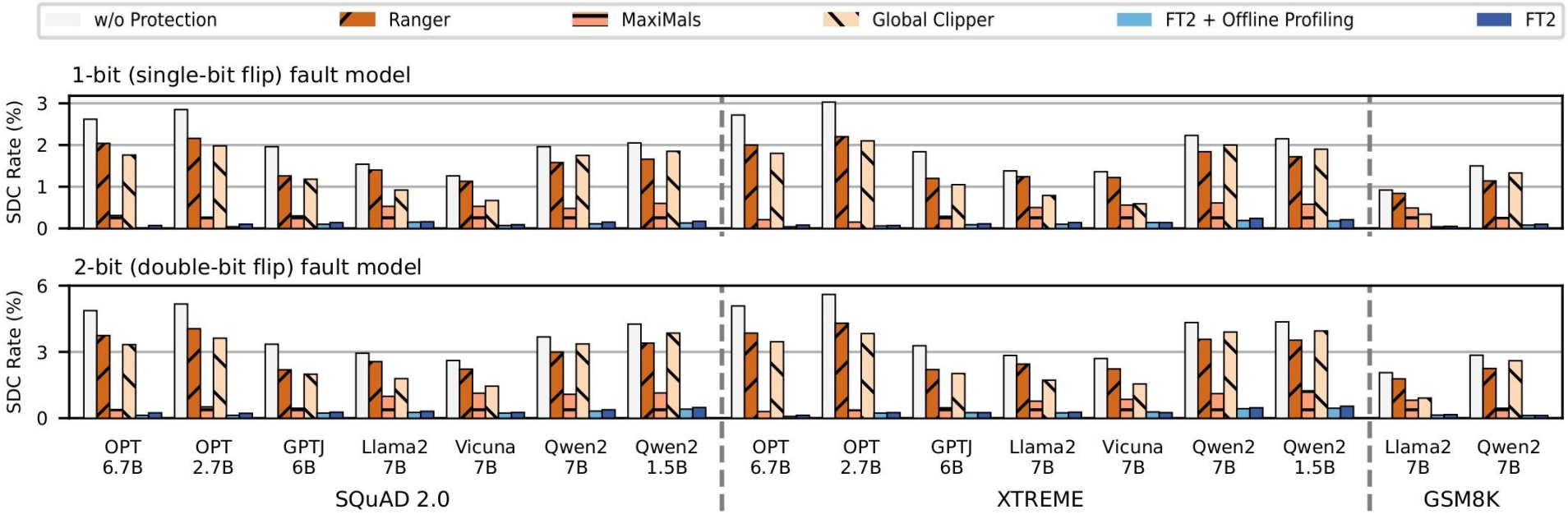
Overall Evaluation



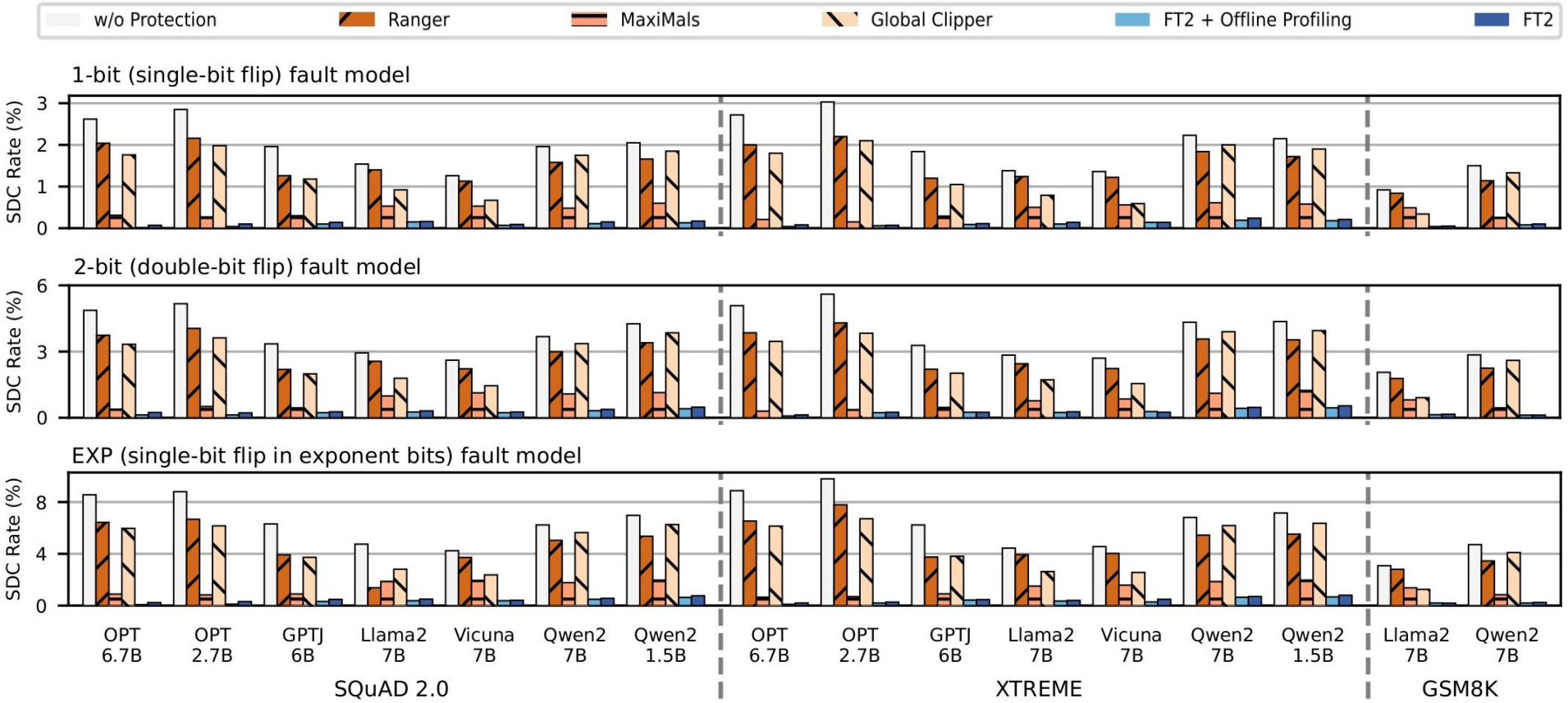
Overall Evaluation



Overall Evaluation

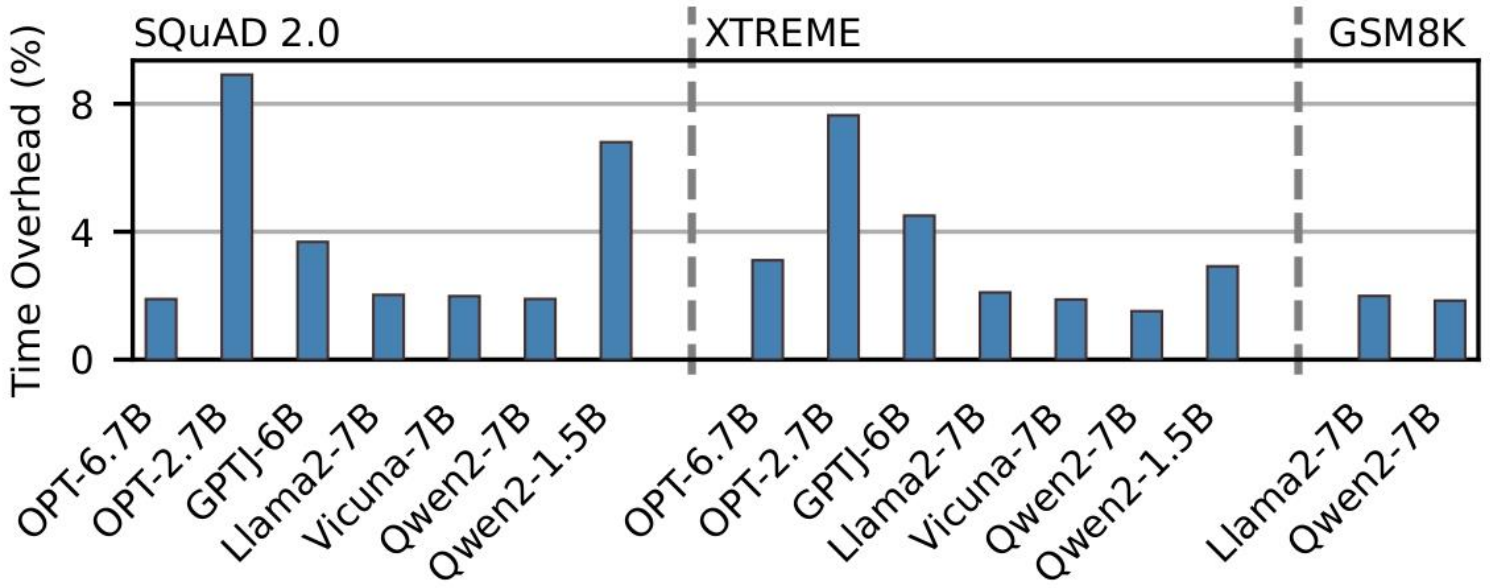


Overall Evaluation



- *FT2 outperforms all baselines among all models, datasets, and fault models*
- *The average SDC rate reduction is 92.92%*

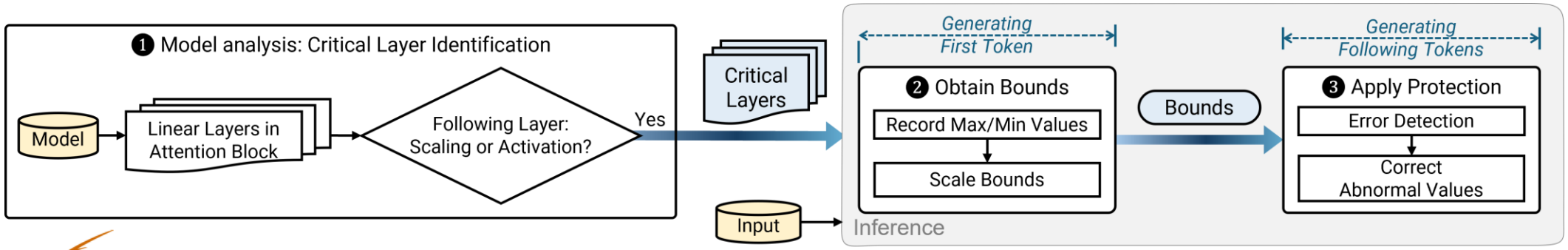
Evaluation: Overhead



- *FT2 introduces 3.42% runtime overhead on average*
- *Memory overhead is negligible (288 - 512 Bytes, <0.2% for all models)*

Conclusion

- LLMs suffer from soft errors
  - Leads to SDC → Harmful and hard to detect
- Ranger has limitations applying to LLMs
  - Insufficient protection
  - Require bound profiling
- Our method: FT2
  - Identify and protect critical layers → high efficiency and low overhead
  - Obtain bounds during first token generation → no offline profiling
- Achieve 92.92% SDC rate reduction
- Only 3.42% overhead on average
- Code at https://github.com/pipijing13/FT2-LLM-inference-protection



Thank you

FT2: First-Token-Inspired Online Fault Tolerance on Critical Layers for Generative Large Language Models

Yu Sun[§], Zhu Zhu[§], Cherish Mulpuru[§],

Roberto Gioiosa[‡], Zhao Zhang[†], Bo Fang[‡], and Lishan Yang[§]

[§]George Mason University

[‡]Pacific Northwest National Lab

[†]Rutgers University

Email: ysun23@gmu.edu

This work is supported by NSF grants (#2402940 and #2410856), CCI grant (#HC-3Q24-047), DOE award 66150, and DOE Contract DE-AC05-76RL01830. The computation resources are provided by the Office of Research Computing at George Mason University (funded by NSF grant #2018631) and the Texas Advanced Computing Center (TACC).

Resilience Estimation

ALUs: Arithmetic Logic Units
ECC: Error Correction Code

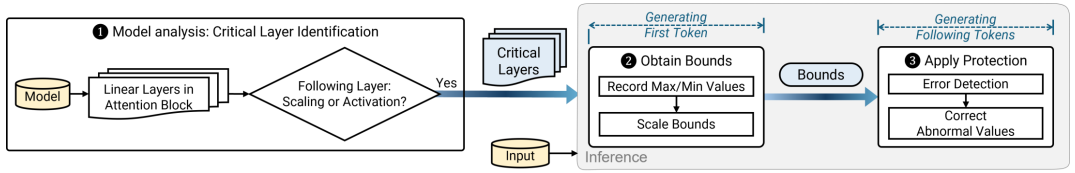
- Computational faults
  - Affect computational hardware components such as ALUs
  - Memory faults are protected by ECC

- Fault models
  - *1-bit*: single-bit flip
  - *2-bit*: double-bit flip
  - *EXP*: single-bit flip in **Exponent** bits (most severe)

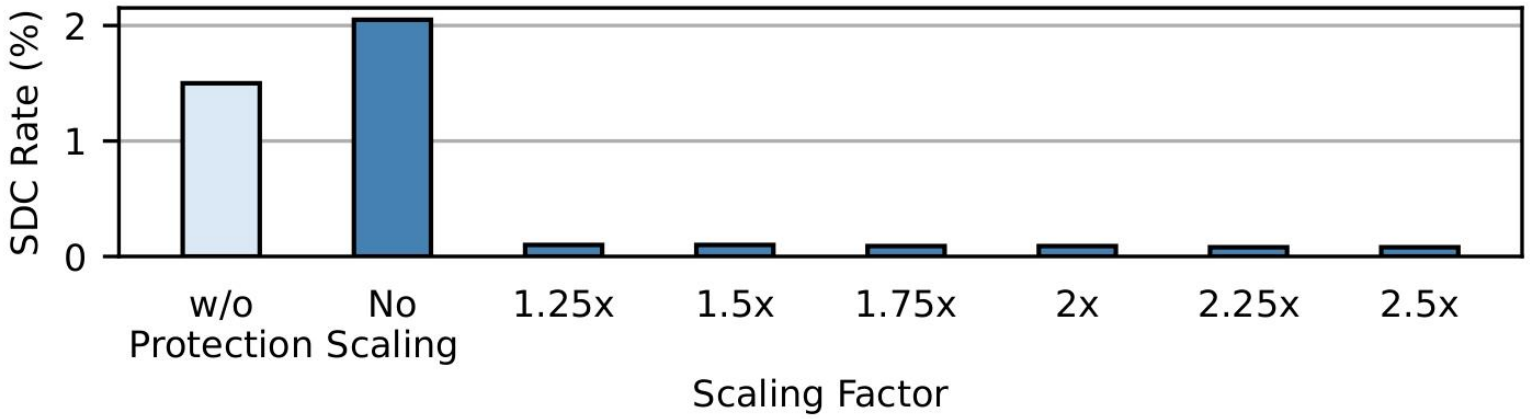
- Inference phase

- Fault injection
  - Into neurons which represent computation results
  - Random select fault sites (block, layer, neuron, bit)

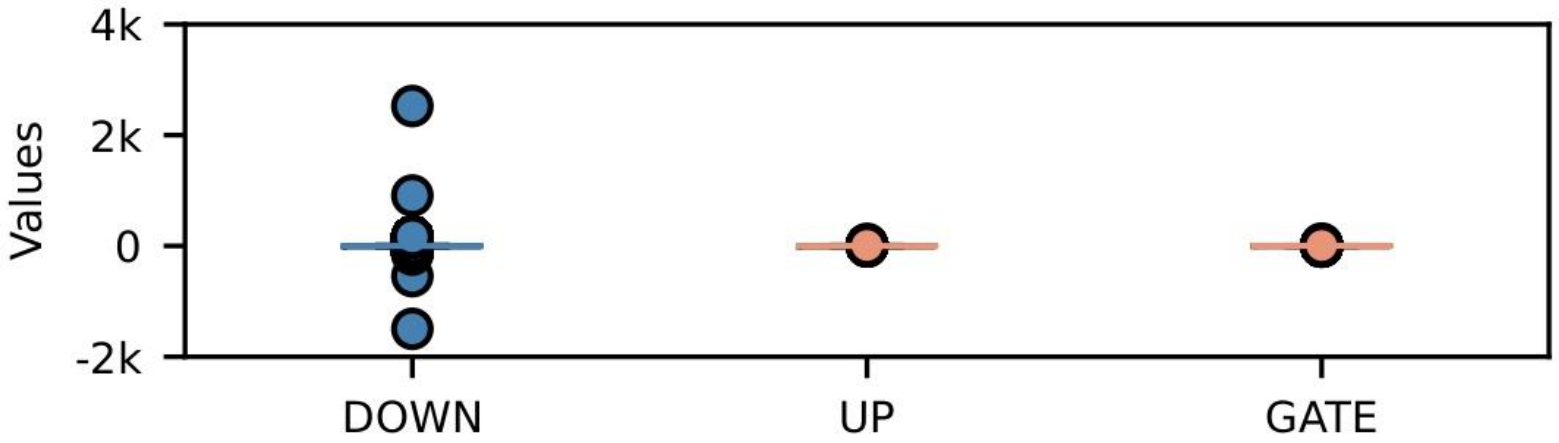
Apply Protection



- Scale the bounds



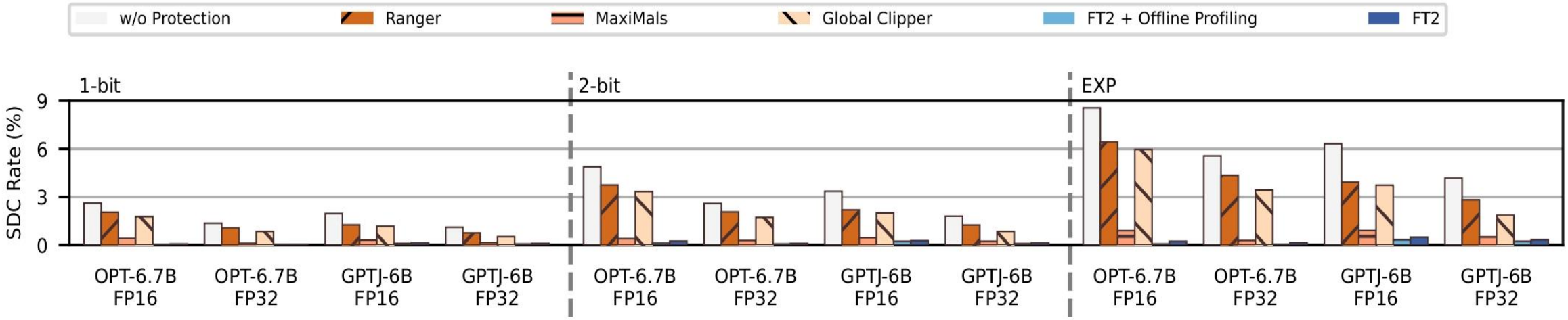
- Clip to bounds



Evaluation: Data Type

FT2 can effectively protect both FP16 and FP32 LLM inference

(Animations here)



Evaluation: GPU Configurations

FT2 is feasible among different generations of NVIDIA GPUs (A100 and H100)

(Animations here)