WeightSentry: Real-Time Bit-Flip Protection for Deep Neural Networks on GPUs

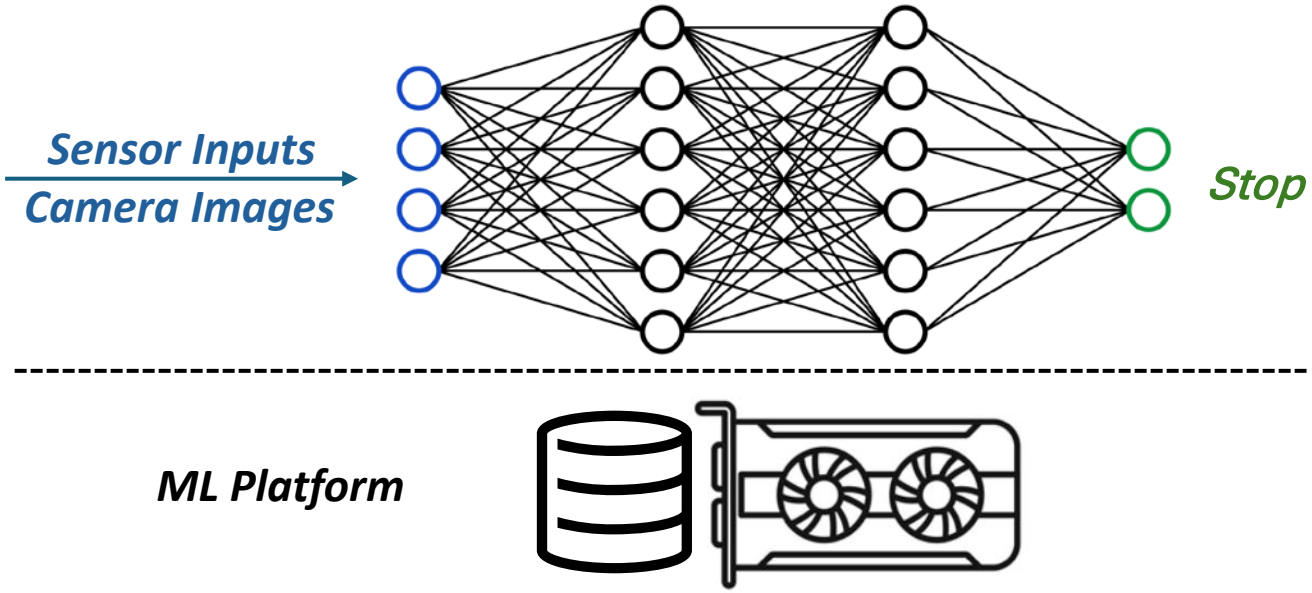
Mahmoud Abumandour[1], Srinija Ramichetty[2], Guru Venkataramani[2], Alaa R. Alameldeen[1]

[1] Simon Fraser University
Burnaby, BC, Canada

[2] George Washington University
Washington, DC, USA

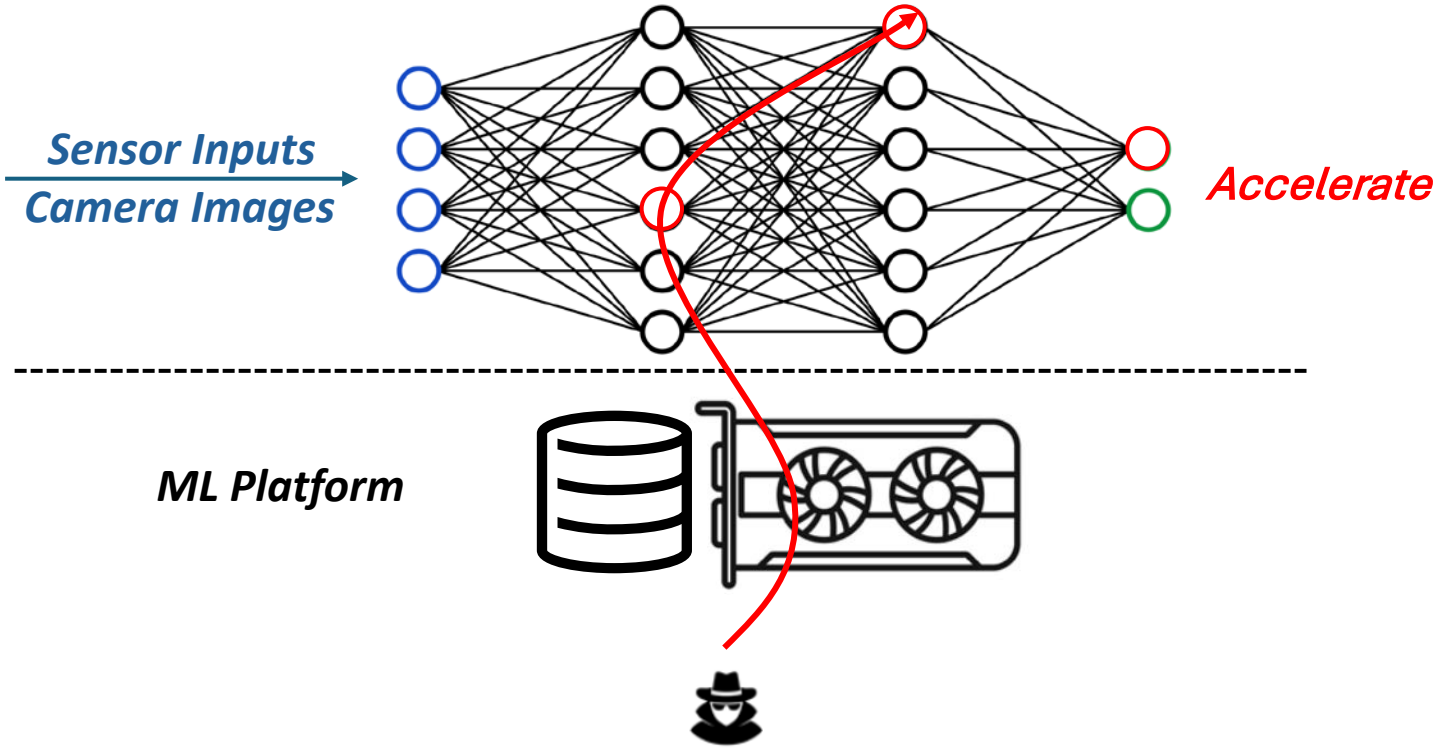
Motivation

- Deep Neural Networks (DNNs) power safety-critical systems
  - Disease diagnosis, autonomous vehicles, malware analysis



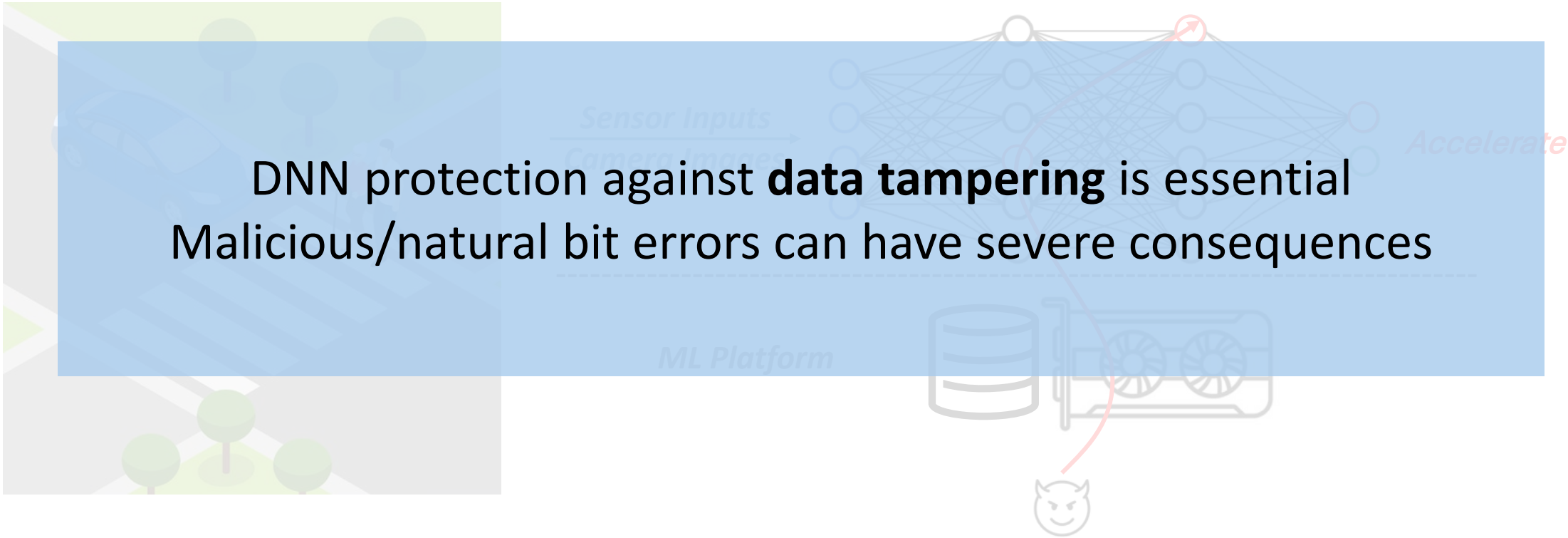
Motivation

- Bit-Flip Attacks (BFA) destroy DNNs using a small number of flips



Motivation

- Bit-Flip Attacks destroy DNNs using a small number of flips

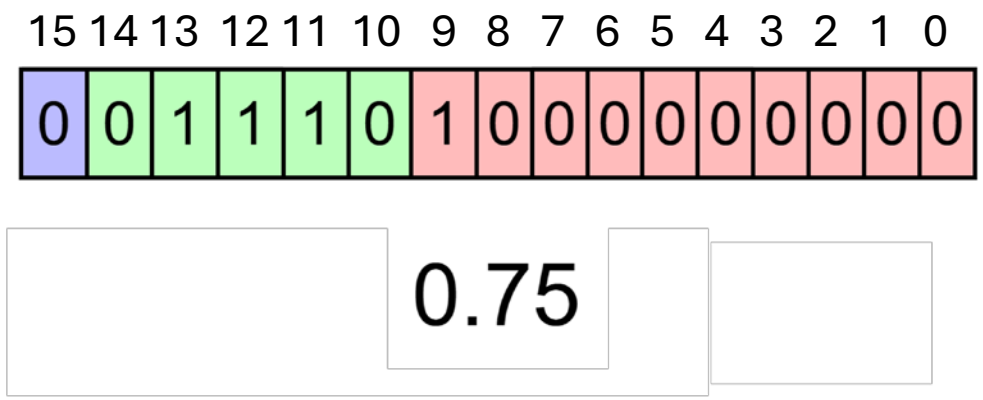


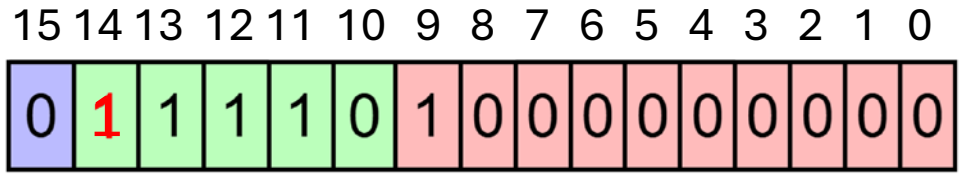
DNN protection against **data tampering** is essential
Malicious/natural bit errors can have severe consequences

Contributions

- We show that FP models are vulnerable to *gradual* BFA
- Propose WeightSentry, protects against strong BFA
- WeightSentry on LLMs & CNNs retains 100% of accuracy
- Suitable for GPU execution

BFA against DNNs: Blind

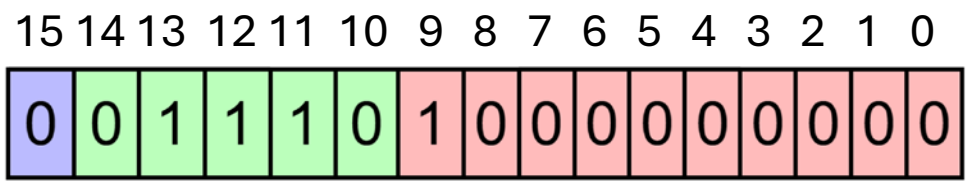




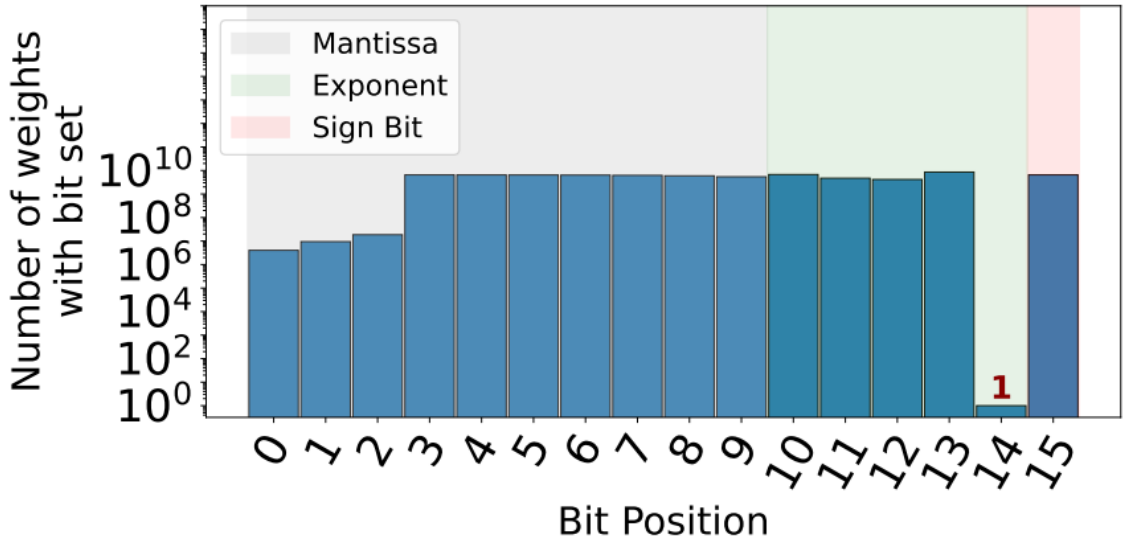
49152.0

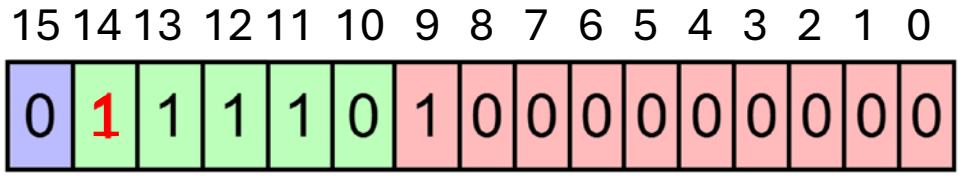
~20,000x
out of range for
Llama 3.2 13B

BFA against DNNs: Blind



0.75

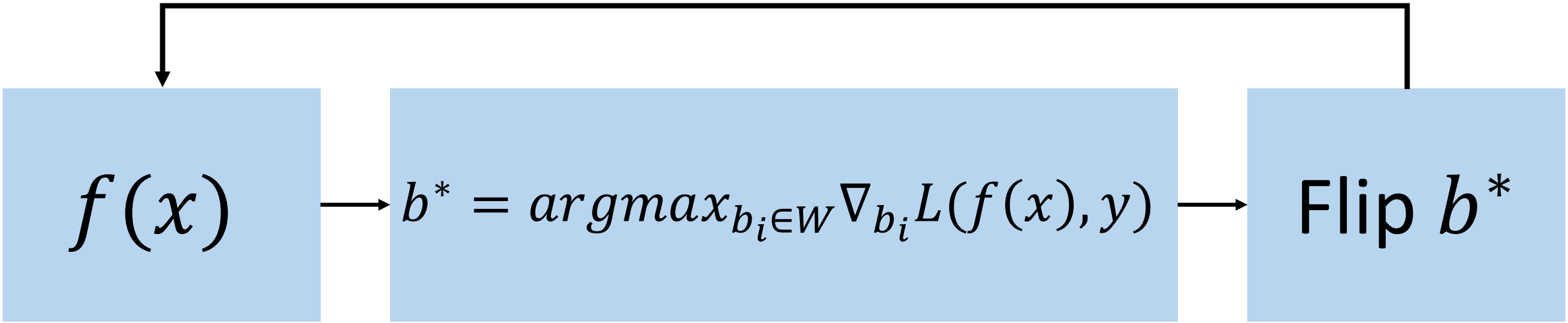




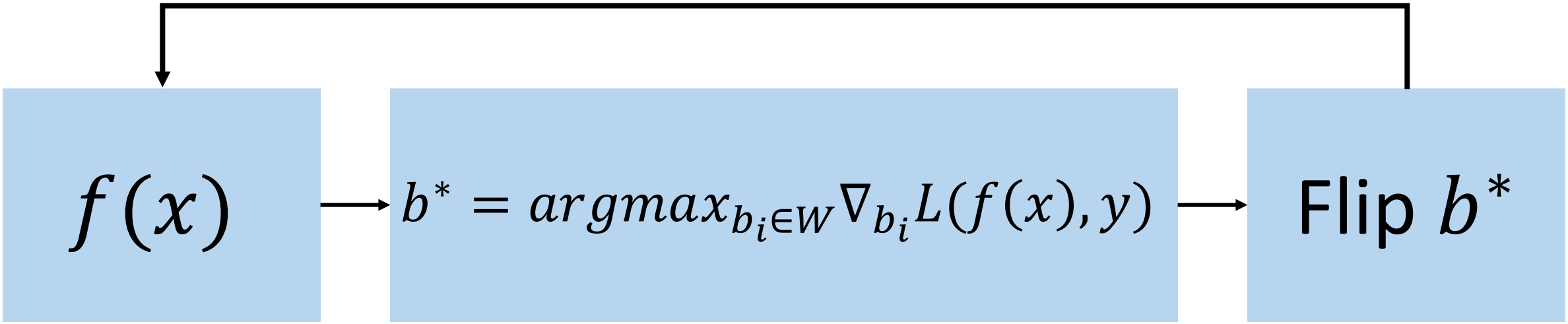
49152.0

~20,000x out of range for Llama 3.2 13B

BFA against DNNs: Gradient Ranking



BFA against DNNs: Gradient Ranking



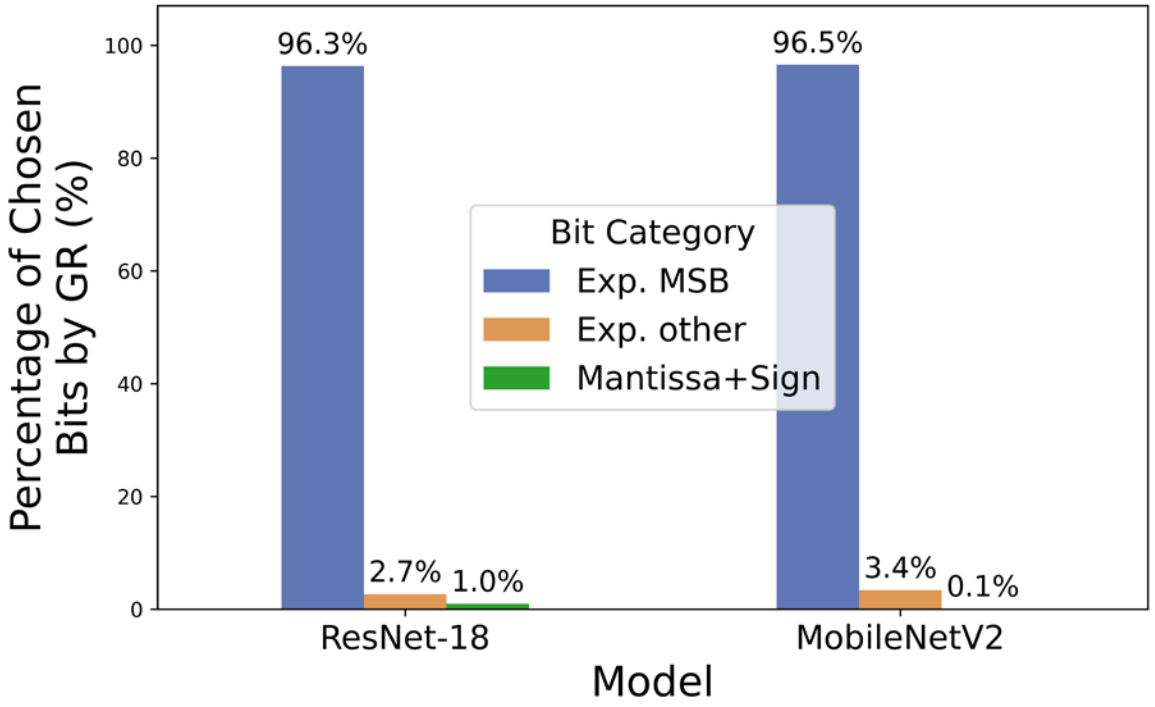
For FP, *almost always* the **exponent's MSB**

BFA against DNNs: Stealthy Gradient Ranking

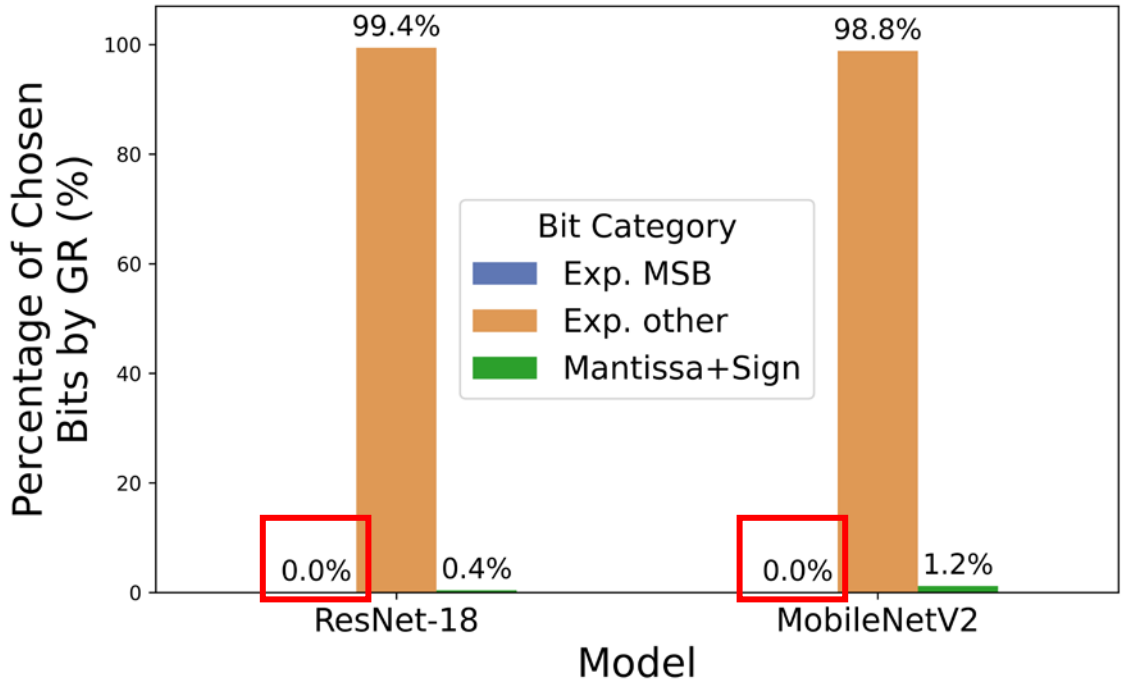
$$b^* = argmax_{b_i \in W} \Delta_{b_i} L(f(x), y); \Delta_{b_i} L < \tau$$

$$1 \leq \tau \leq 10^3$$

BFA against DNNs: GR Chosen Bits

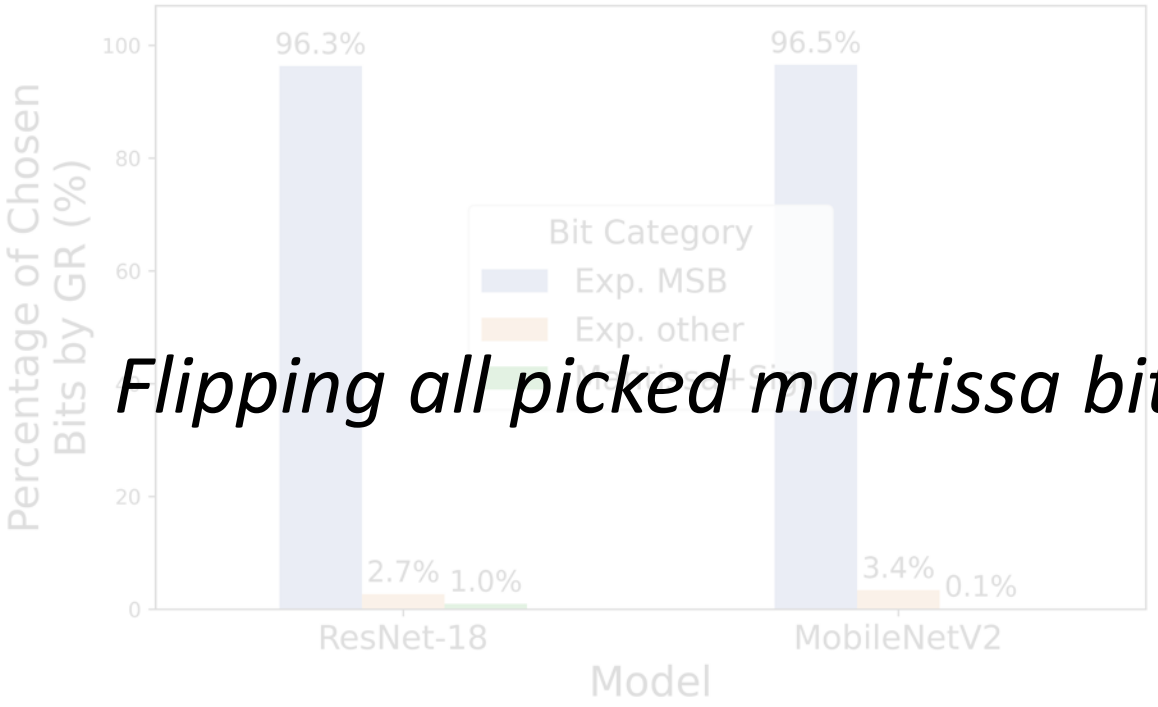


GR BFA

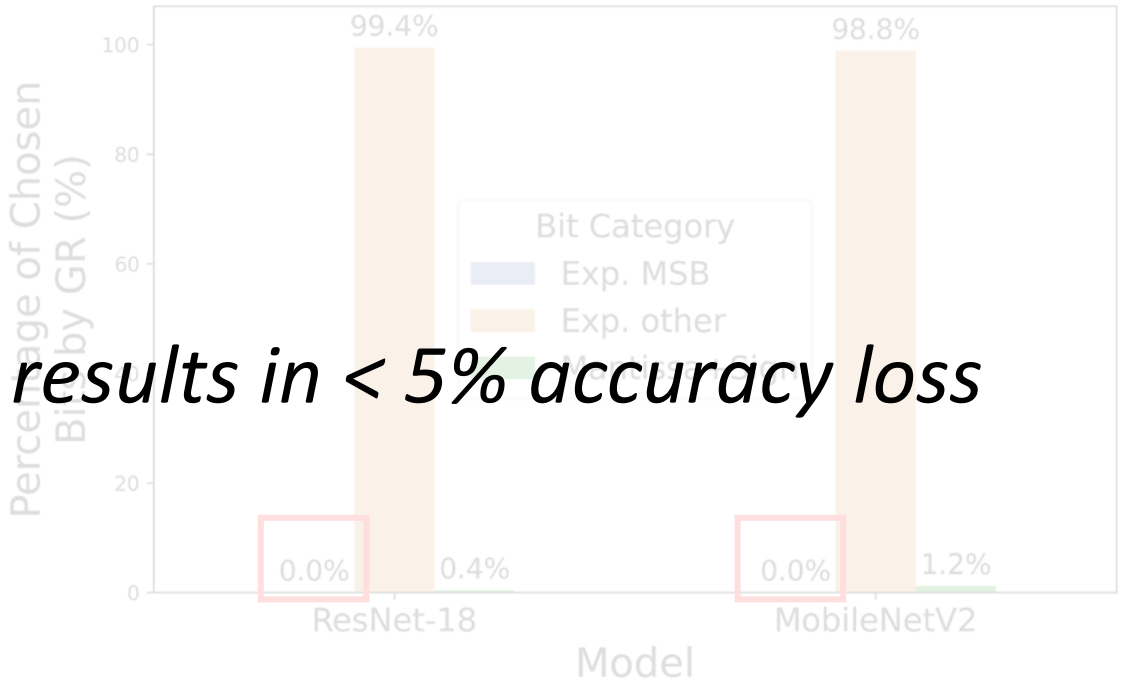


GR S-BFA

BFA against DNNs: GR Chosen Bits



GR BFA



GR S-BFA

Flipping all picked mantissa bits results in < 5% accuracy loss

BFA Defenses

- Existing BFA Defenses:
  - Modify model architecture
  - Require retraining
  - Tailored for quantized models
  - Work best on CPUs

- ***Desiderata**:*

BFA Defenses

- Existing BFA Defenses:
  - ~~Modify model architecture~~
  - ~~Require retraining~~
  - ~~Tailored for quantized models~~
  - ~~Work best on CPUs~~

- ***Desiderata**: None*

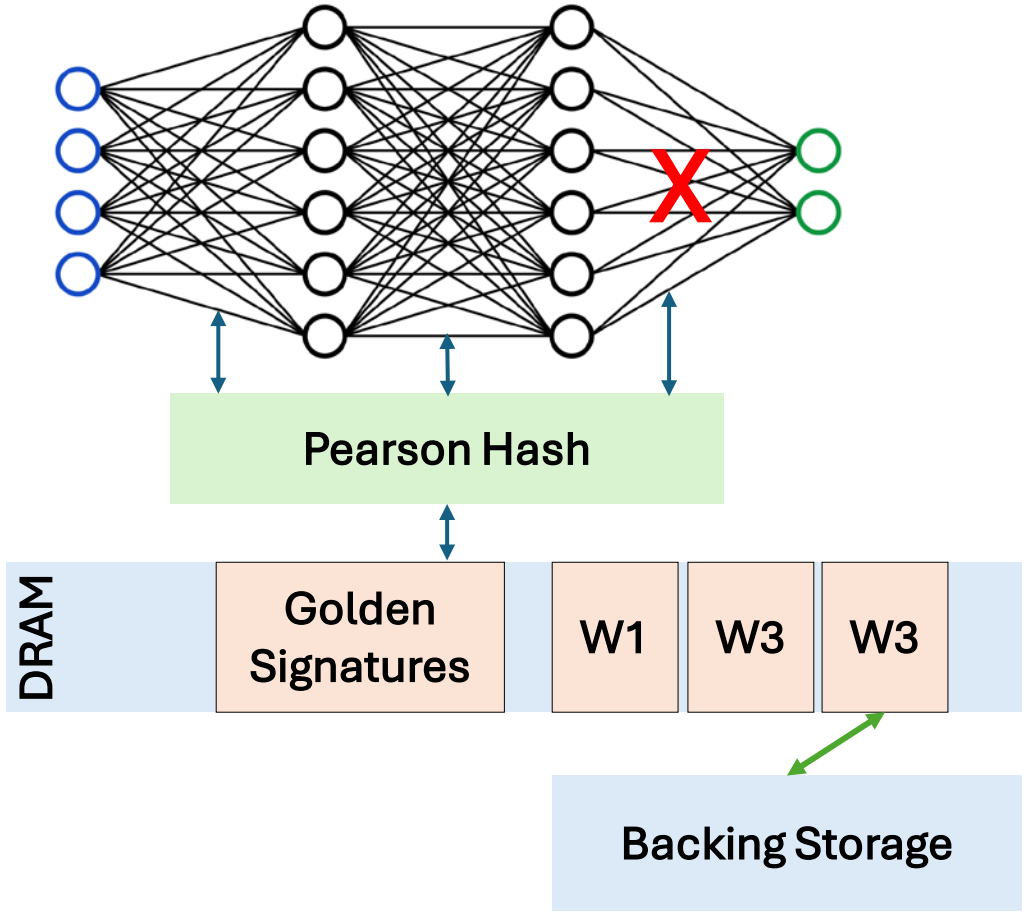
BFA Defenses

- Existing BFA Defenses:
  - ~~Modify model architecture~~
  - ~~Require retraining~~
  - ~~Tailored for quantized models~~
  - ~~Work best on CPUs~~

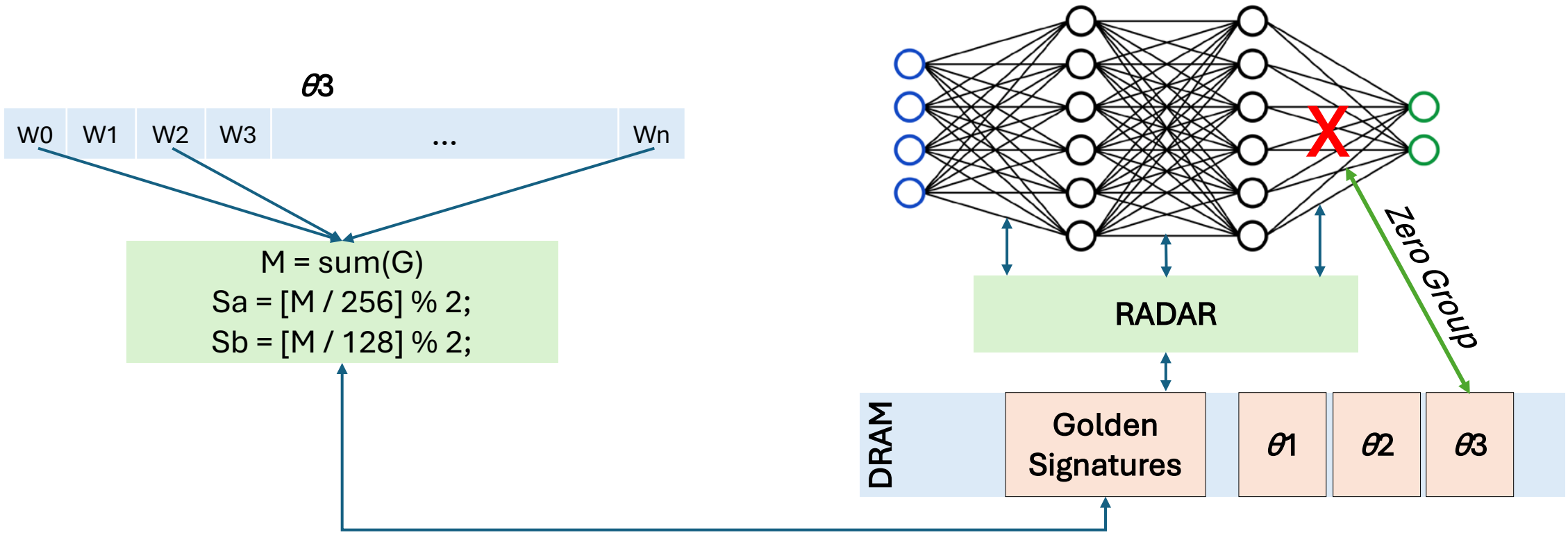
- ***Desiderata**:*
  - Minimal modifications to trained models
  - Work for FP
  - Scale to GPUs

Comparison Baselines: HASHTAG

$$h_{i+1} = T[h_i \oplus c]$$

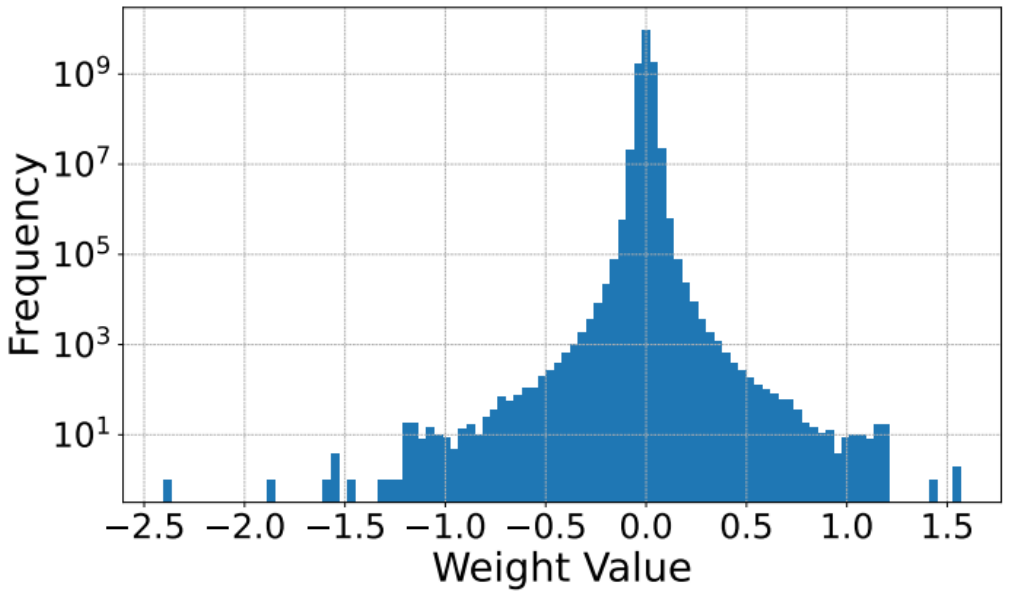


Comparison Baselines: RADAR

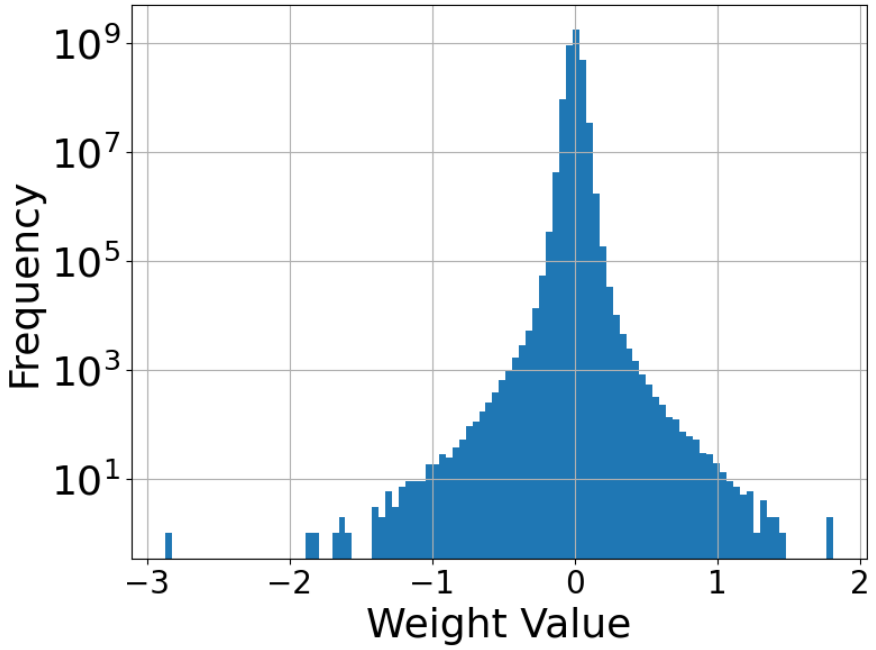


Key Insight

- Exponent bits do **most** of the damage
- FP model weights have tight statistical bounds

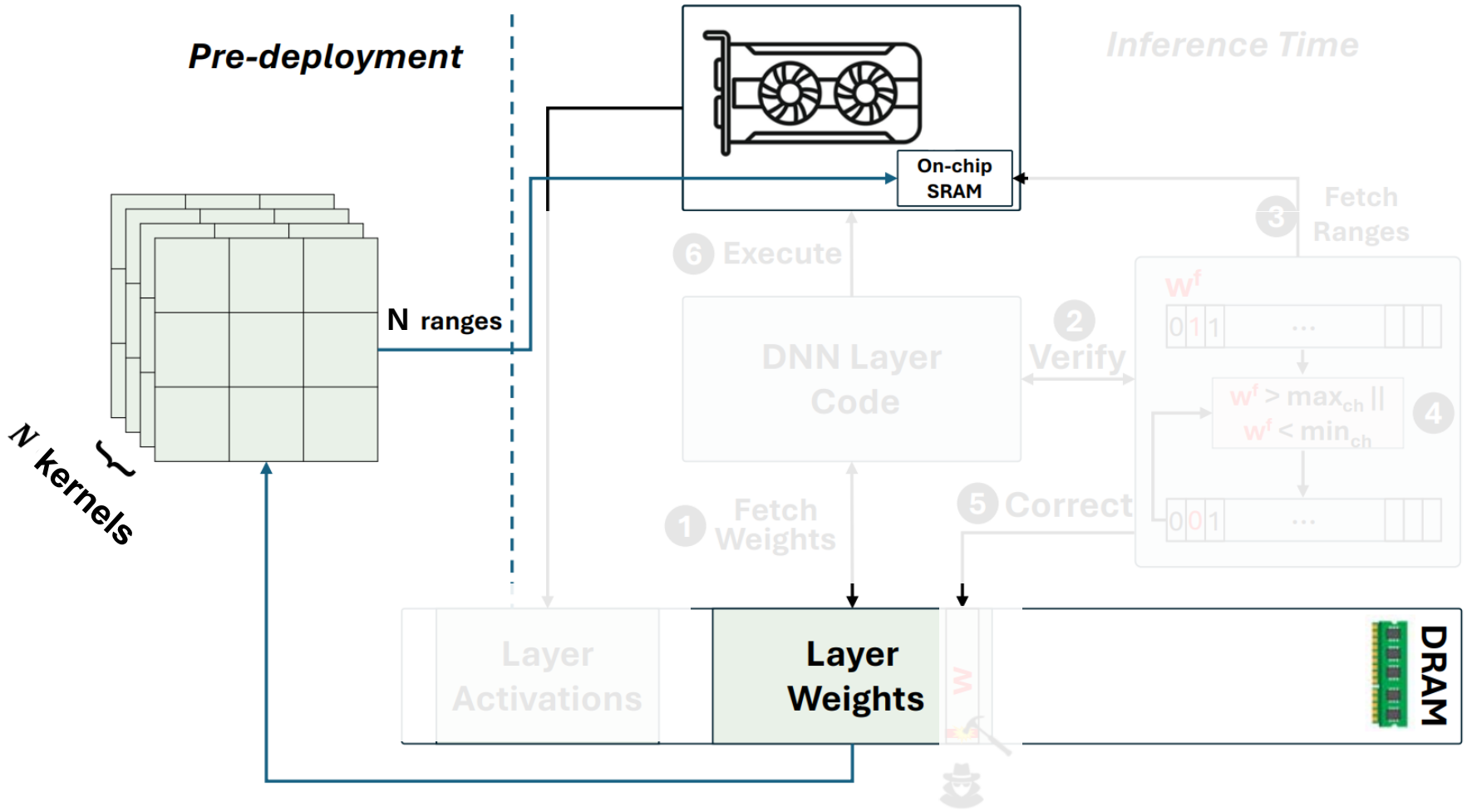


Llama 3.2 13B

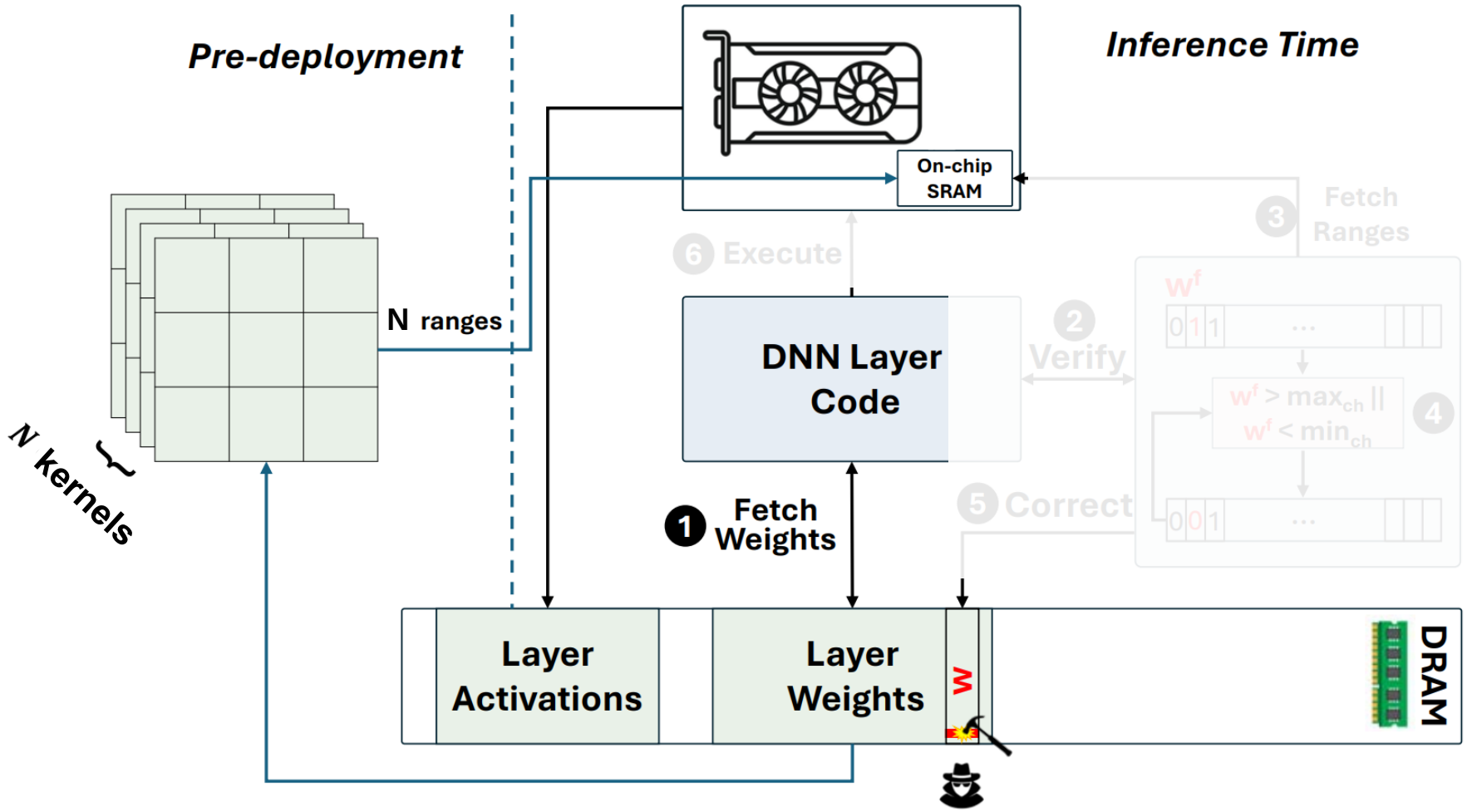


Phi-4-Mini

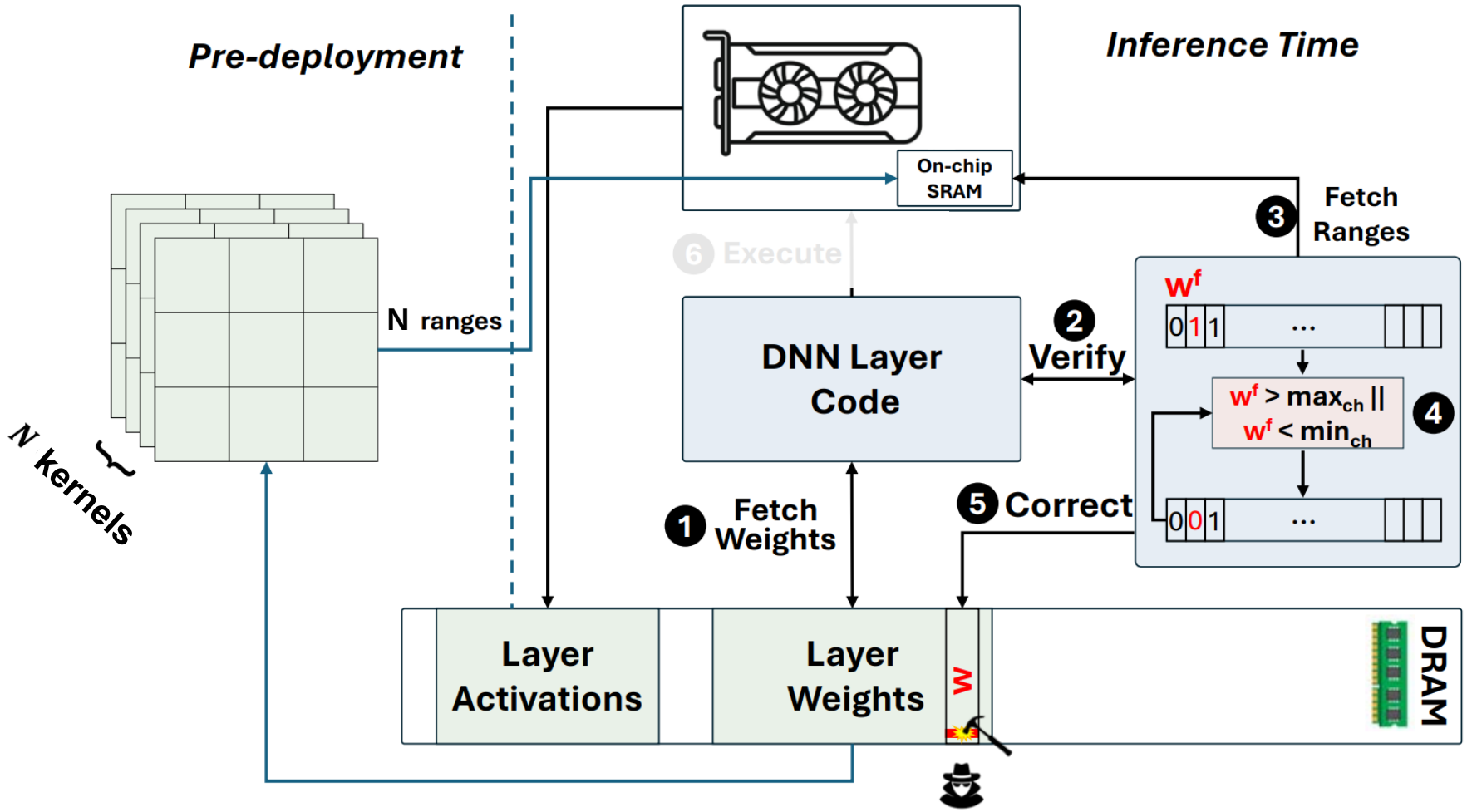
WeightSentry Design



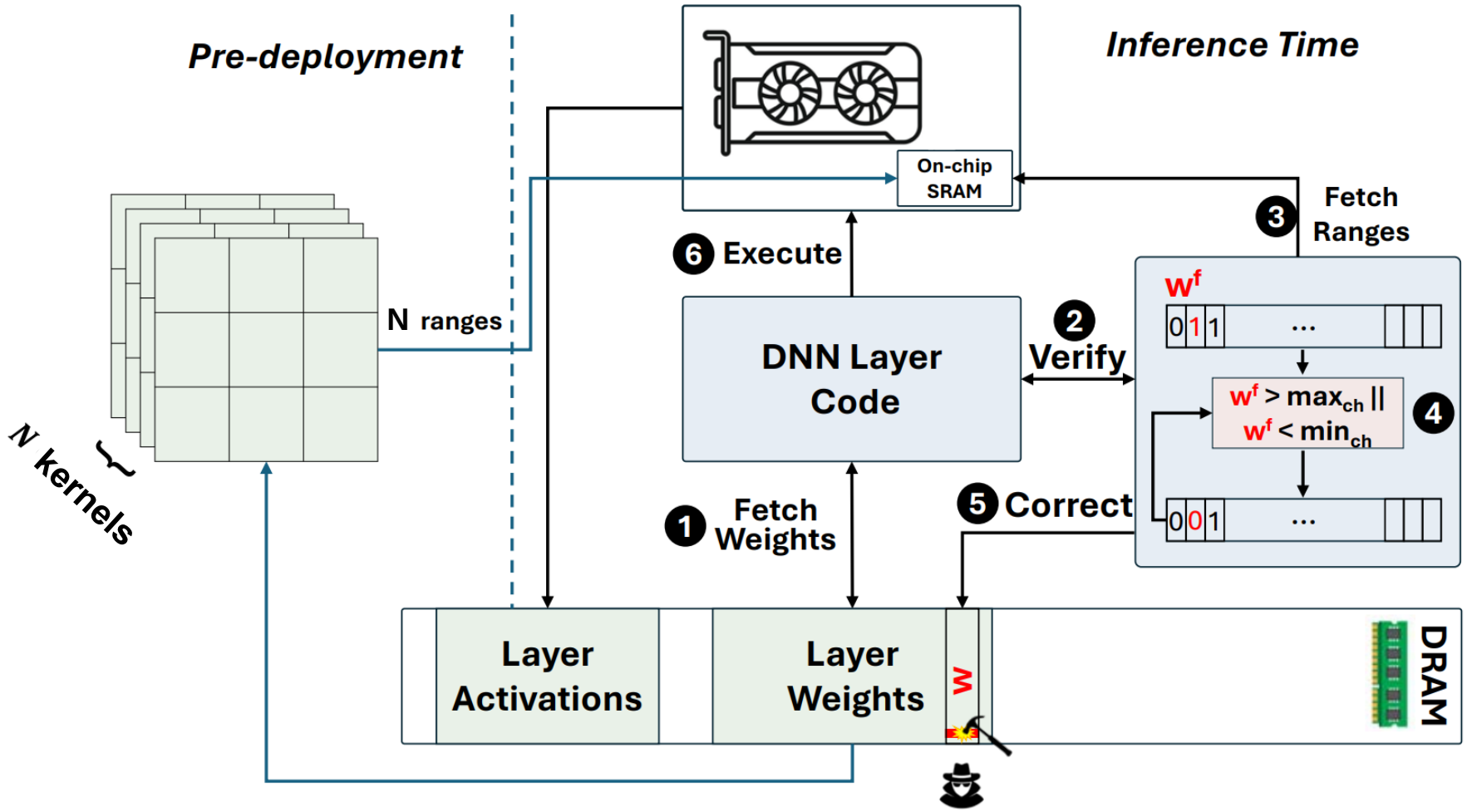
WeightSentry Design



WeightSentry Design



WeightSentry Design



Evaluation

- Hardware: NVIDIA RTX 4090 GPU
- CNNs: **ResNet-18/50**, **MobileNetV2**, **DenseNet121**
- LLMs: **LLaMA-3.2-3B**, **Qwen2.5-VL-3B**, **Phi-4-Mini**, **SmolLM3-3B**

Evaluation

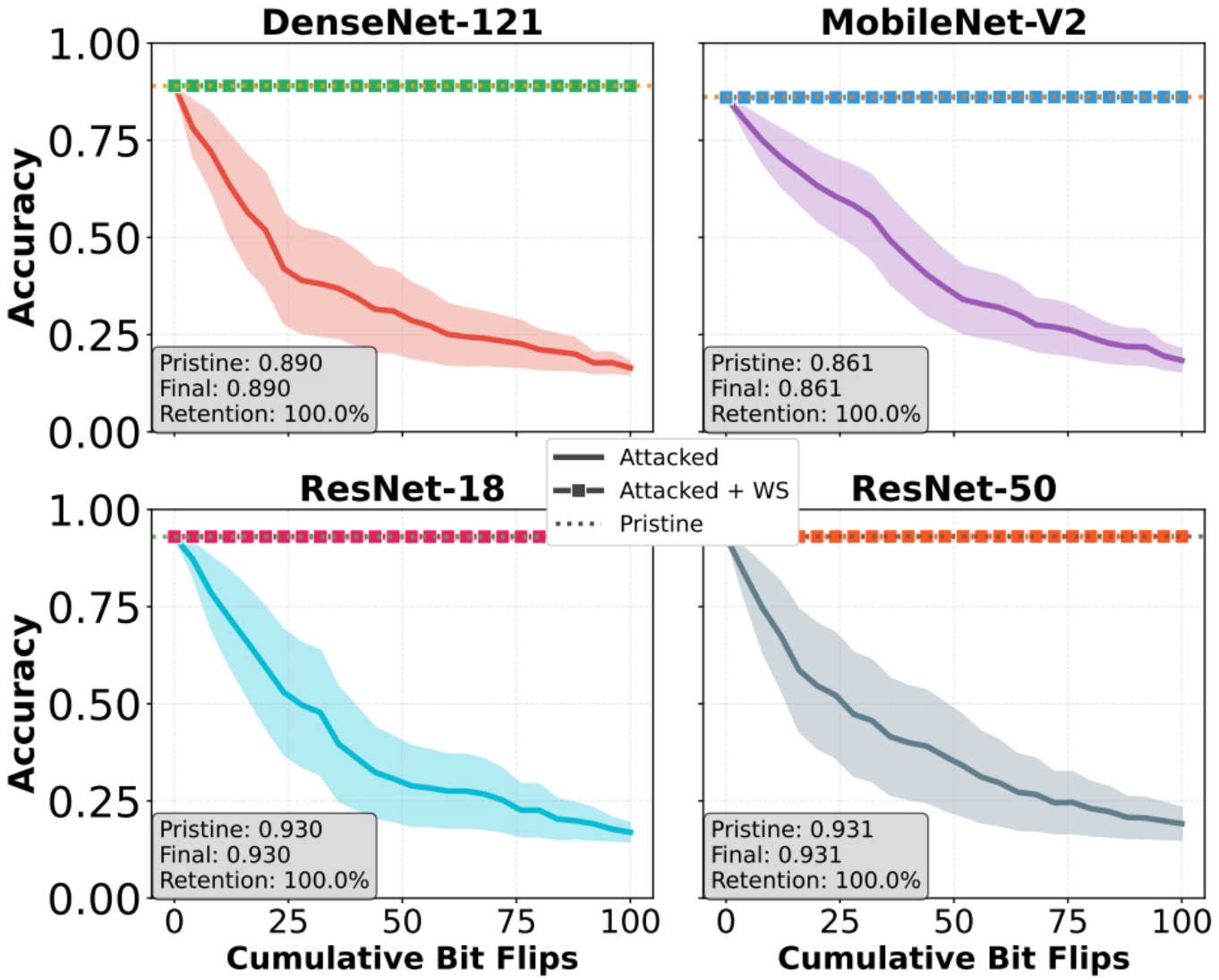
- Hardware: NVIDIA RTX 4090 GPU
- CNNs: **ResNet-18/50**, **MobileNetV2**, **DenseNet121**
- LLMs: **LLaMA-3.2-3B**, **Qwen2.5-VL-3B**, **Phi-4-Mini**, **SmolLM3-3B**

- White-box threat model:
  - Architecture, weights, framework
  - Hardware co-location & ability to flip *any* bit combination

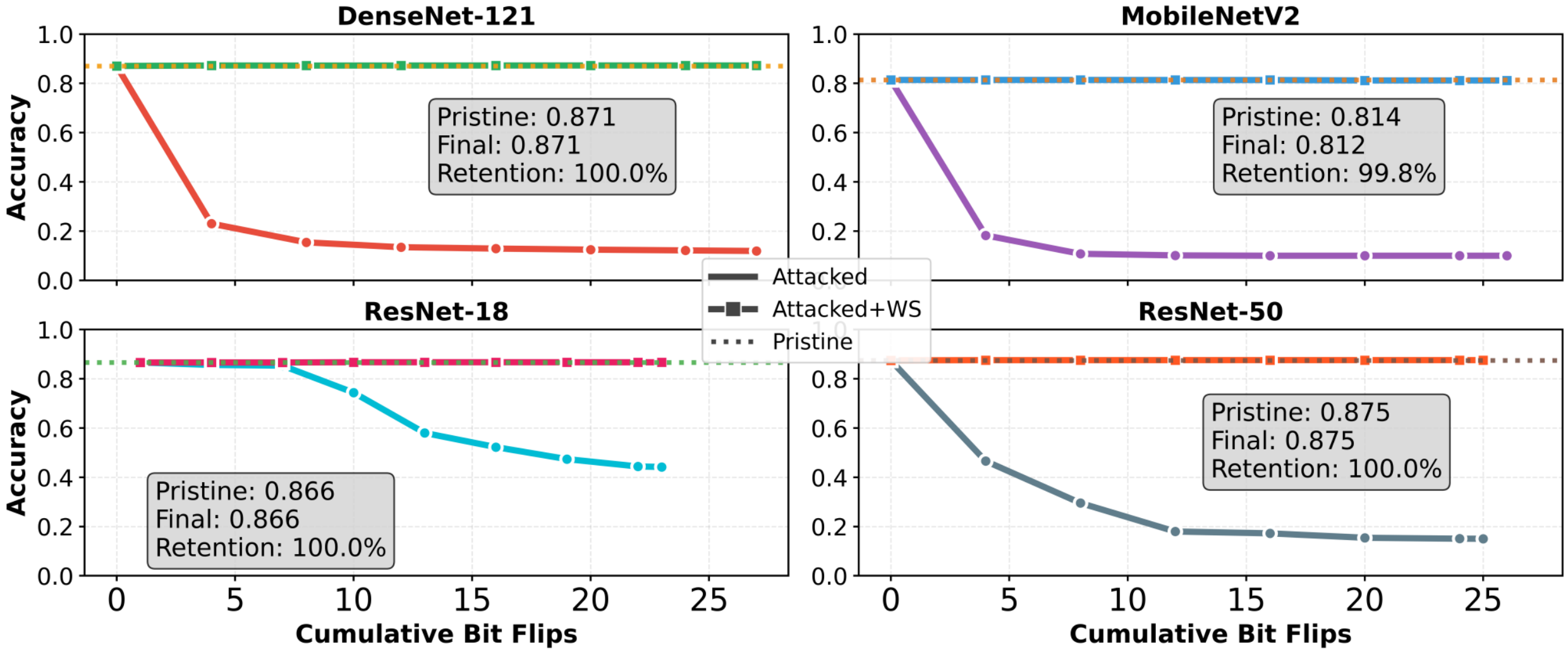
- Compare against two prior works: HASHTAG and RADAR

Random BFA on CNNs

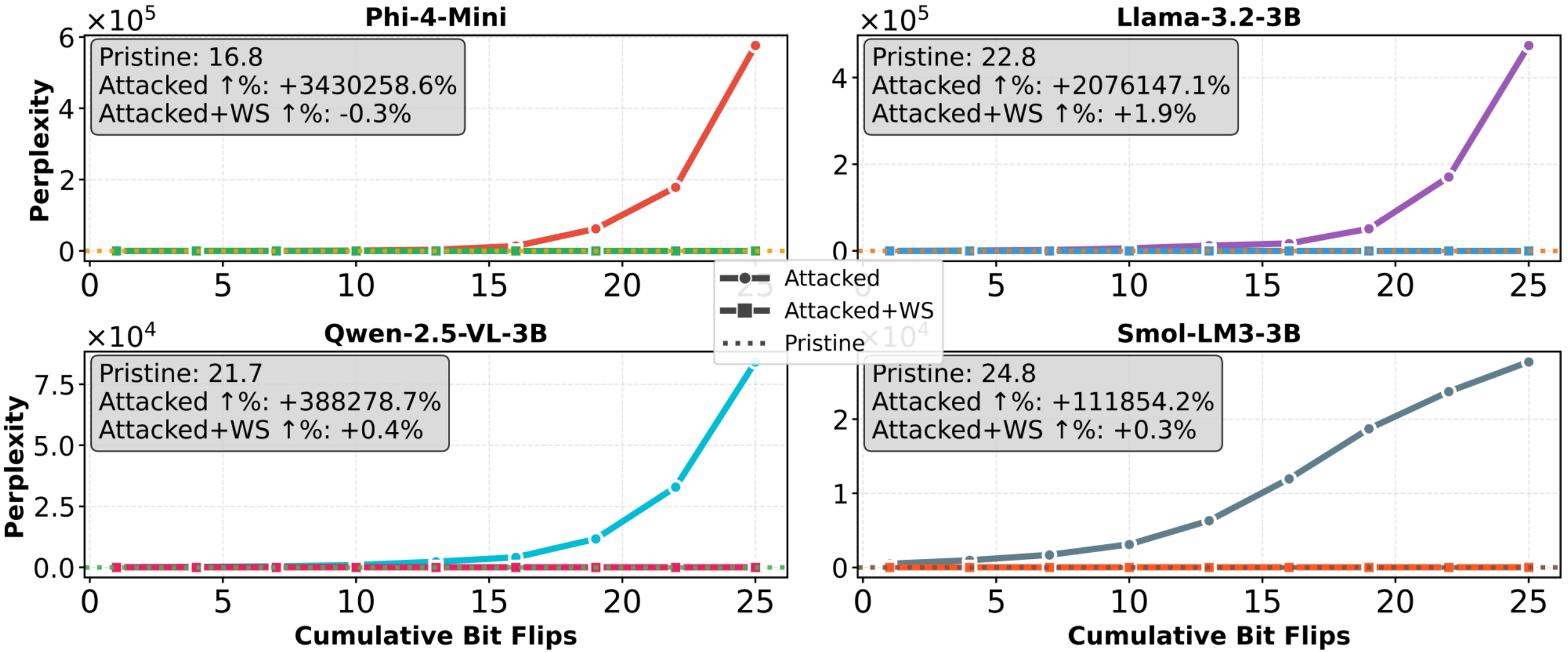
- Flips in the ***top two bytes***
- Average across 1000 runs



S-BFA on CNNs


DenseNet-121
Pristine: 0.871
Final: 0.871
Retention: 100.0%
MobileNetV2
Pristine: 0.814
Final: 0.812
Retention: 99.8%
ResNet-18
Pristine: 0.866
Final: 0.866
Retention: 100.0%
ResNet-50
Pristine: 0.875
Final: 0.875
Retention: 100.0%
Attacked
Attacked+WS
Pristine
Accuracy
Cumulative Bit Flips
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25

S-BFA on LLMs


Phi-4-Mini
×10^5
Pristine: 16.8
Attacked ↑%: +3430258.6%
Attacked+WS ↑%: -0.3%
Perplexity
Llama-3.2-3B
×10^5
Pristine: 22.8
Attacked ↑%: +2076147.1%
Attacked+WS ↑%: +1.9%
Attacked
Attacked+WS
Pristine
Qwen-2.5-VL-3B
×10^4
Pristine: 21.7
Attacked ↑%: +388278.7%
Attacked+WS ↑%: +0.4%
Perplexity
Smol-LM3-3B
×10^4
Pristine: 24.8
Attacked ↑%: +111854.2%
Attacked+WS ↑%: +0.3%
Cumulative Bit Flips
Cumulative Bit Flips

S-BFA on LLMs

"What is the meaning of life?"

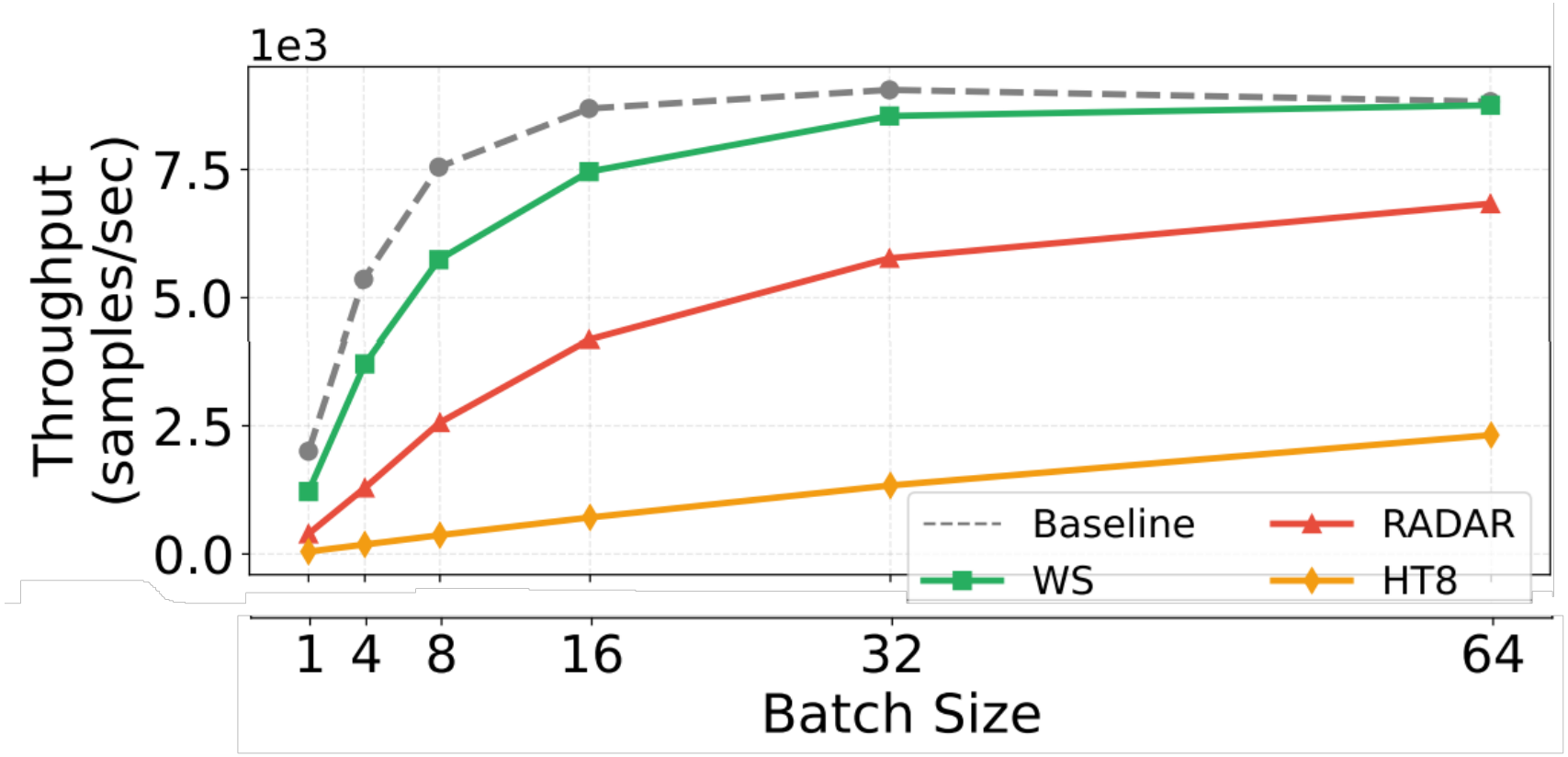
Model	Protected Output	Attacked Output
Llama-3.2-3B	It's a question that has puzzled philosophers and scientists for centuries...	The 2019 2019 2019 The 2019 The The 2019 2019...
Phi-4-Mini	It has puzzled humanity since time immemorial...	Drachenkönig wohnt auf einem hohen Hügdynamisch...
Gemma-3-1B	There is no single, universally accepted answer to the question "What is the meaning of life?"	= = ...
SmolLM3-3B	Is it to be happy? To love? To find meaning in our work? To leave a legacy?	What is the meaning of life? What is the meaning of life? What is the meaning of living?

Performance Impact

WeightSentry approaches baseline (ResNet18)

B=64:

- WS 1.5%
- RADAR 34.8%
- HASHTAG 280%



Storage Overhead

Model	HT	RD	WS
ResNet-50	40.41	22.43	**4.47**
ResNet-18	15.56	10.66	**2.64**
DenseNet121	91.36	6.63	**3.81**
MobileNetV2	37.65	**2.13**	5.55
Llama-3.2-3B	63.75	3063.92	**1.98**
Qwen2.5-3B	206.80	3580.69	**6.44**
Phi-4-Mini	48.69	3658.32	**1.52**
SmolLM3-3B	81.82	2932.64	**2.55**

Overheads in KB

Summary

WeightSentry:

- No architecture modifications/retraining
- Detects and corrects bit-flips in real-time
- Suitable for GPU inference (~1% overhead)
- 100% accuracy retention under strong attacks

Thank you. Questions?

mahmoud_abumandour@sfu.ca

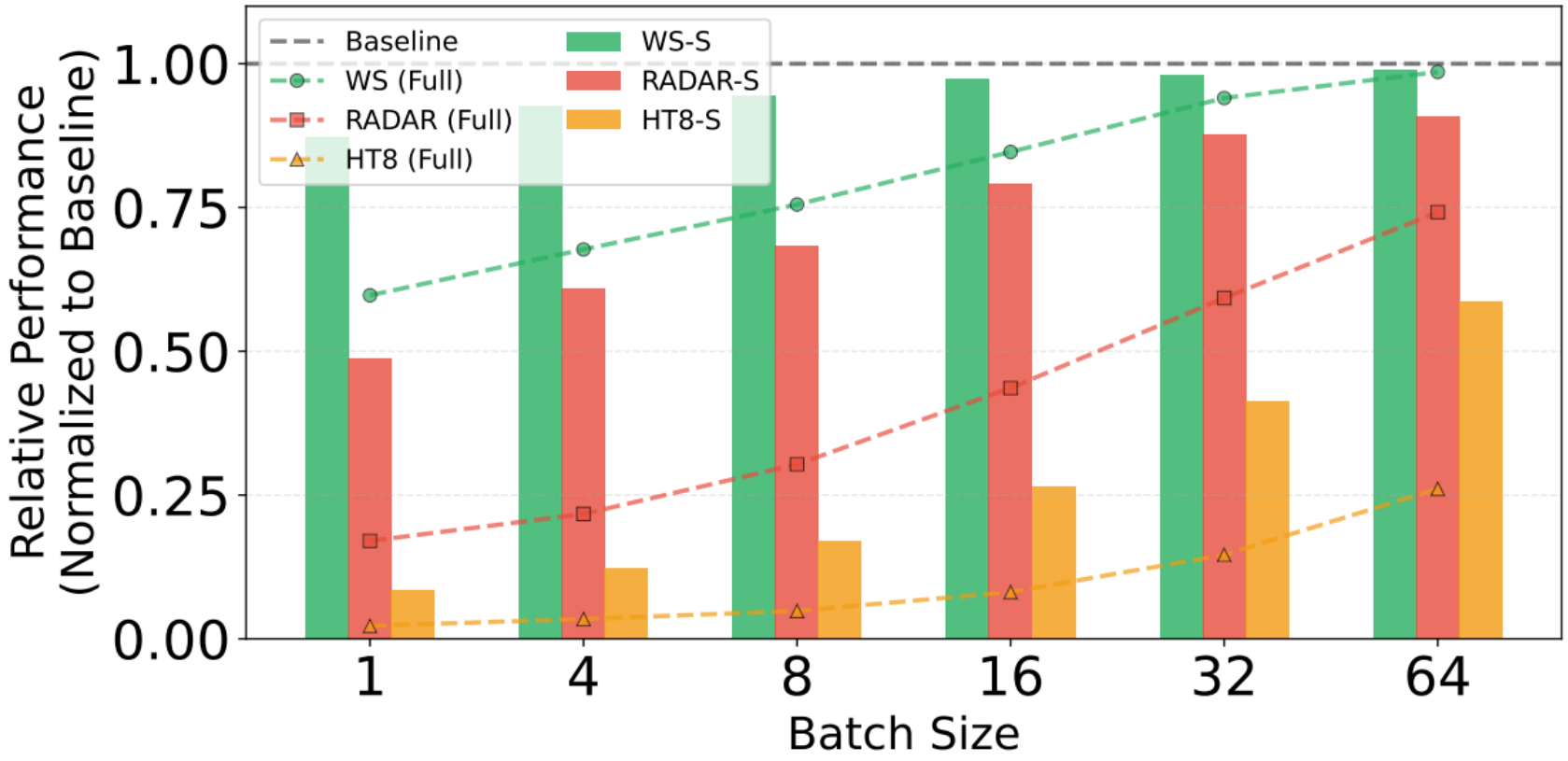
Backup slides

Performance Impact: Selective Protection

Protect top 3 layers only

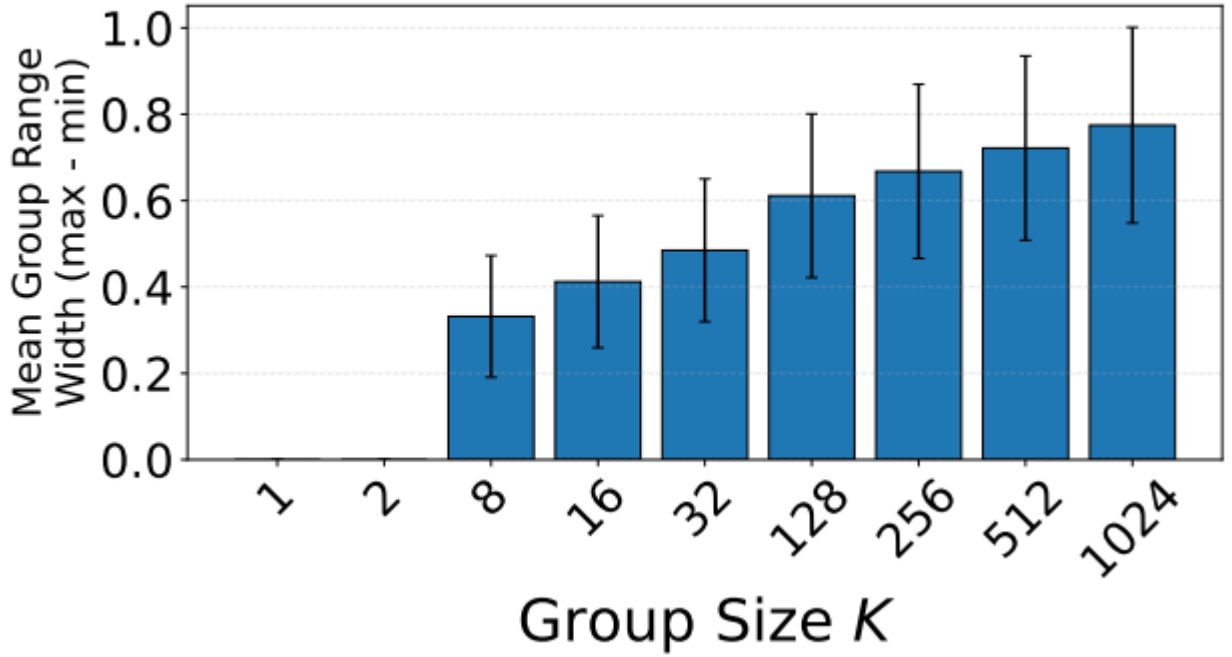
B=16:

- WS 2.8%
- RADAR 26%
- HASHTAG 275%



WeightSentry Design: Kernel Grouping

- Per-kernel ranges
- Conv 1x1 kernels are common
- We use a single range for K weights



Kernel Granularity

Kernel size tradeoff storage vs. security