

UF UNIVERSITY of FLORIDA

Low-Overhead Silent Error Detection for Large Language Model Inference at Scale

Abdullah Naveed, Mahina Sheikh, Guanpeng Li

Department of Electrical and Computer Engineering
University of Florida

SC26 • Technical Papers

Fast inference needs reliable execution

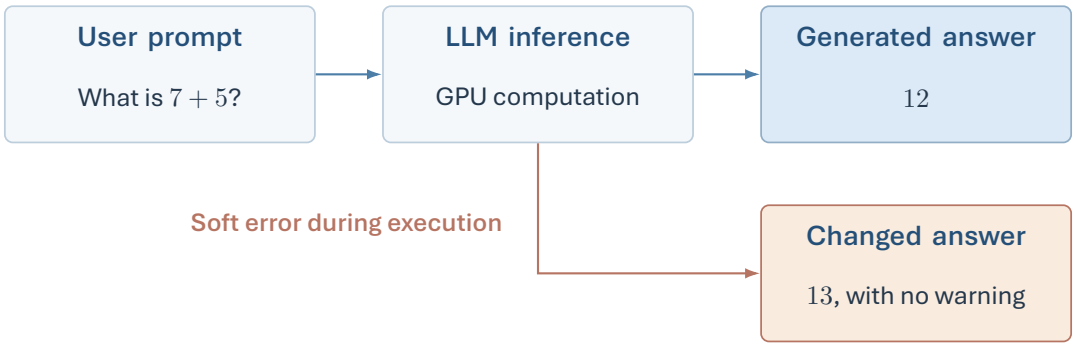
A soft error can change the answer even when inference completes.



We study changes caused by soft errors, not ordinary model mistakes.

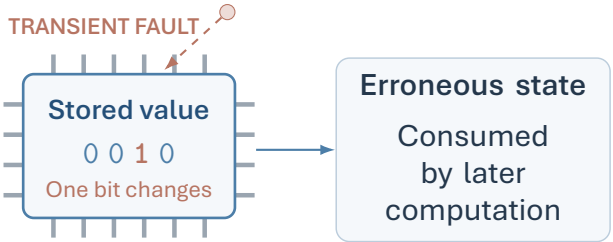
Fast inference needs reliable execution

A soft error can change the answer even when inference completes.

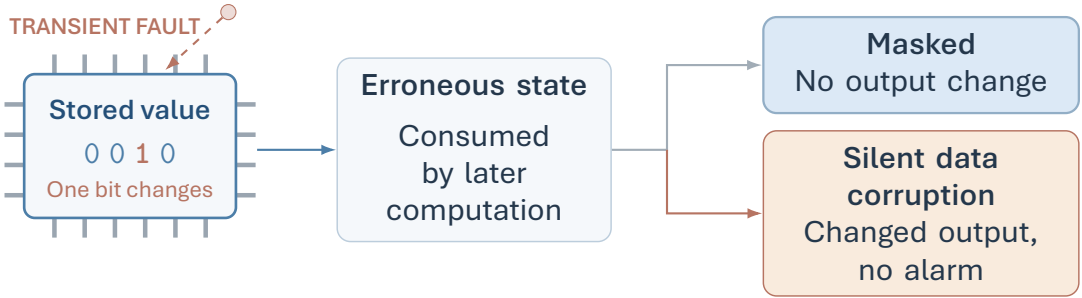


We study changes caused by soft errors, not ordinary model mistakes.

Soft errors can silently change an answer



Soft errors can silently change an answer



Silent corruption can change the output without triggering a warning.

Three goals for practical detection



Coverage

Catch corrupted outputs



Few false alarms

Avoid flagging benign runs



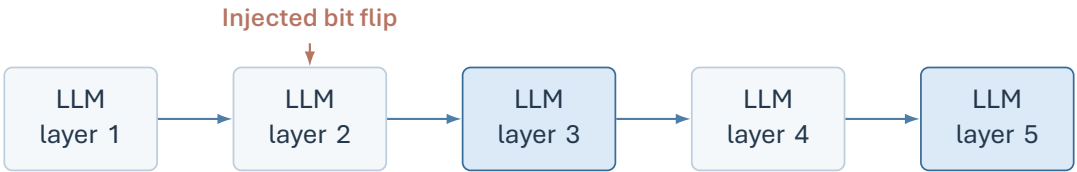
Low cost

Preserve inference latency

- Catch output-changing errors without flagging harmless changes.
- Keep detection lightweight enough for normal inference.

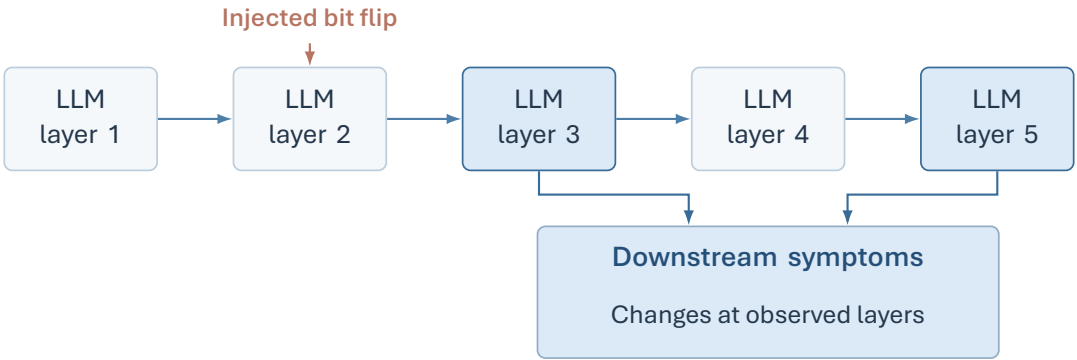
Where should we observe error propagation?

Activations are intermediate numerical values produced during inference.



Where should we observe error propagation?

Activations are intermediate numerical values produced during inference.

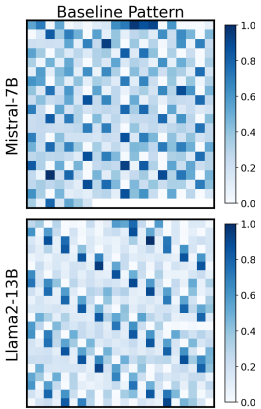


An error in one layer can leave detectable symptoms in later layers.

Corruption leaves structured symptoms

Mistral-7B / SQuAD2.0 (top); Llama2-13B / XTREME (bottom).

CLEAN REFERENCE



Measure each layer's activation spread.

$$R_{\ell,t}^{\text{run}} = x_{\ell,t}^{\max} - x_{\ell,t}^{\min}$$

$$S_{\ell}^{\text{base}} = \max_t \frac{R_{\ell,t}^{\text{run}}}{R_{\ell}^{\text{prof}} + \epsilon}$$

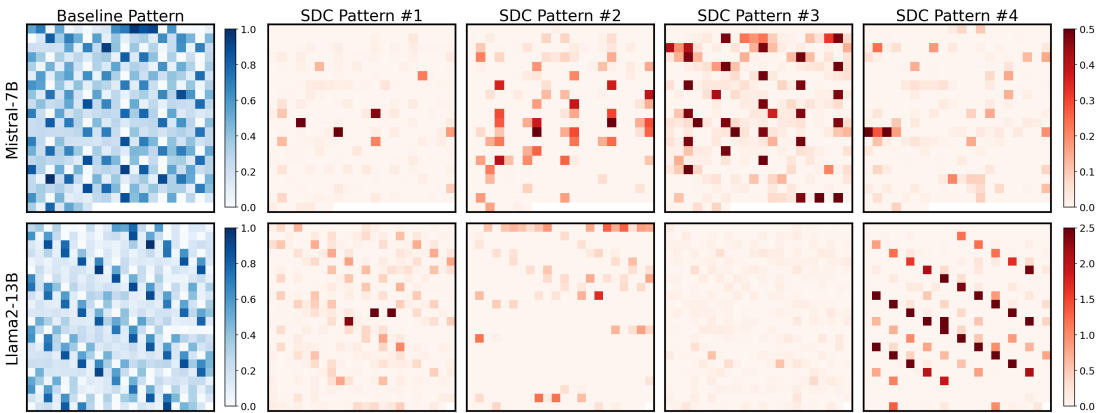
- Normalize by the clean profiled range; retain the maximum over decoding steps.

One cell per layer. Blue: clean score. Red: increase from baseline (separate color scales).

Corruption leaves structured symptoms

Mistral-7B / SQuAD2.0 (top); Llama2-13B / XTREME (bottom).

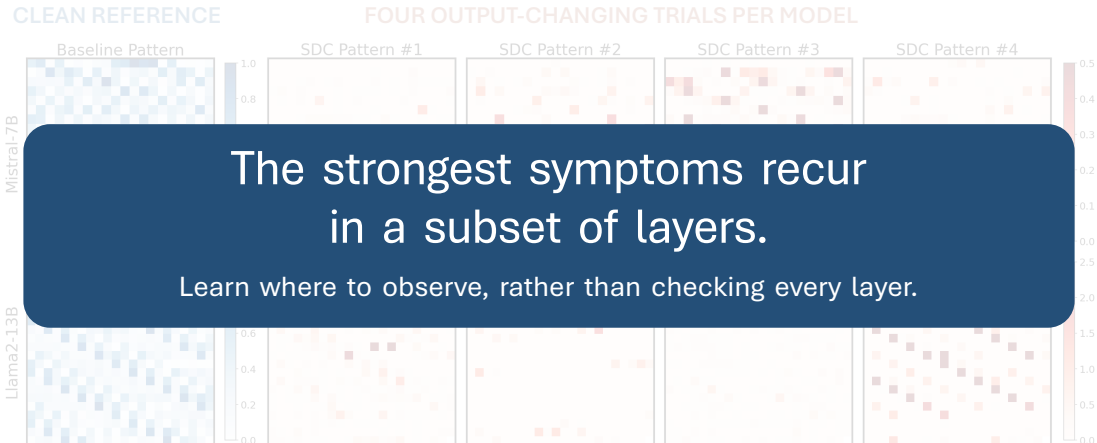
CLEAN REFERENCE **FOUR OUTPUT-CHANGING TRIALS PER MODEL**



One cell per layer. Blue: clean score. Red: increase from baseline (separate color scales).

Corruption leaves structured symptoms

Mistral-7B / SQuAD2.0 (top); Llama2-13B / XTREME (bottom).



One cell per layer. Blue: clean score. Red: increase from baseline (separate color scales).

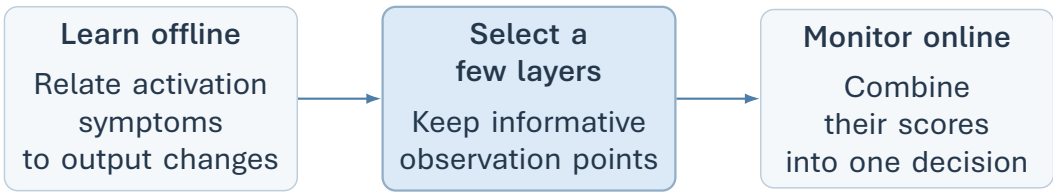
The strongest symptoms recur in a subset of layers.

Learn where to observe, rather than checking every layer.

TRACE: learn offline, monitor selectively

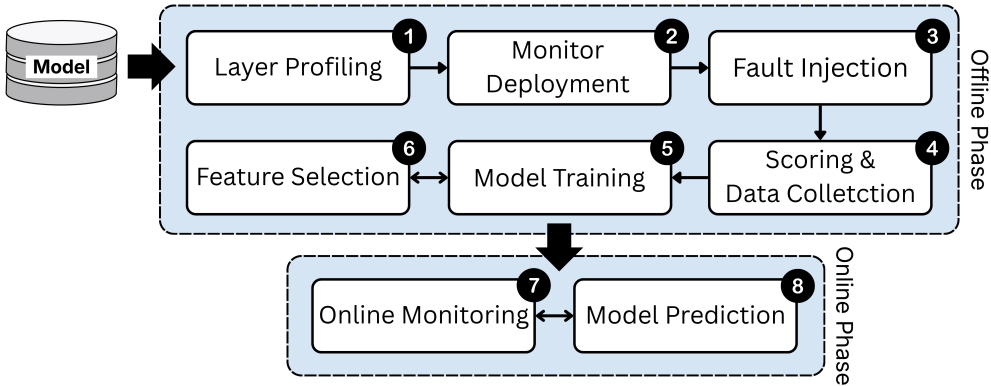
TRACE

Selective activation monitoring for LLM inference



Combine layer evidence instead of raising independent alarms.

Construct offline. Monitor online.



Train offline; monitor selected layers during inference.

One score per layer, over a generation

Clean profiling

500 clean inputs

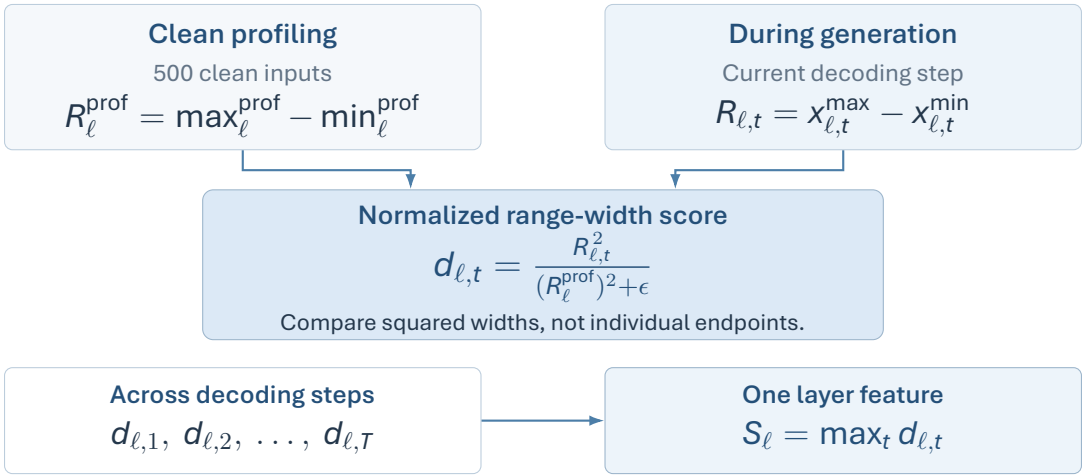
$$R_\ell^{\text{prof}} = \max_\ell^{\text{prof}} - \min_\ell^{\text{prof}}$$

During generation

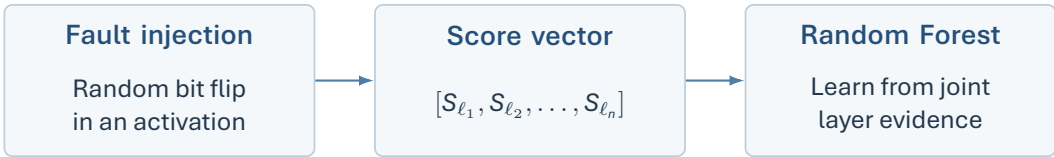
Current decoding step

$$R_{\ell,t} = x_{\ell,t}^{\max} - x_{\ell,t}^{\min}$$

One score per layer, over a generation



Learn from controlled fault injections



Compare the output with its fault-free reference:

Label: masked

Output matches reference

Label: SDC

Output differs from reference

Layer deviations are evidence; the joint classifier decides.

A few layers capture most of the signal

TOP-*K* FEATURE-SELECTION RESULTS (%)

K	Accuracy	Recall	Precision	F1
1	96.0	81.4	85.2	80.5
2	99.4	92.8	95.6	94.0
3	99.6	95.6	96.4	95.9
4	99.7	96.4	96.4	96.3
5	99.7	96.6	96.9	96.6

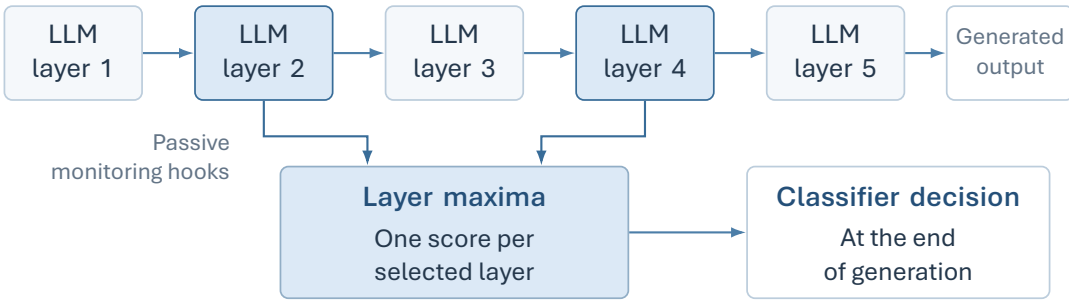
Rank, add, retrain.

- Rank scores by feature importance.
- Retrain as each layer is added.
- Stop when accuracy and recall meet user-defined thresholds.

Small selections retain strong detection quality; the final subset is fixed before independent testing.

Only selected monitors run during inference

The underlying LLM and decoding policy are unchanged.



The classifier combines selected layer scores into one decision.

Experimental setup

15	FP16	A100
Model-dataset pairs	Activation format	Primary GPU

WORKLOADS / 5 ASSIGNED MODELS EACH

GSM8K	SQuAD2.0	XTREME
Math reasoning	Question answering	Benchmark inputs

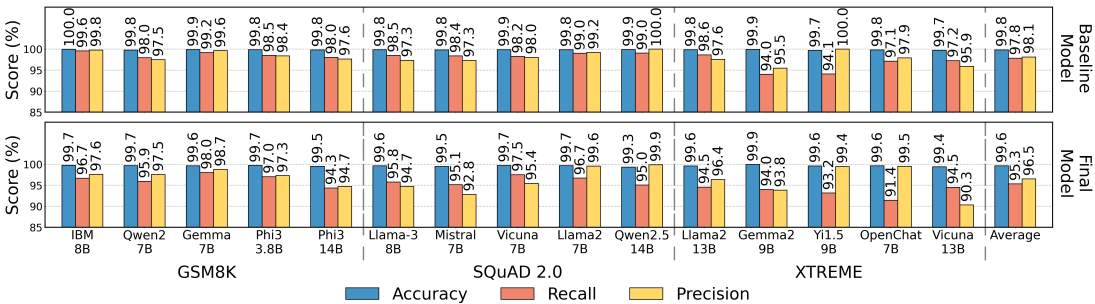
PER MODEL CONFIGURATION

500	4,000	1,000
Clean profiling inputs	Development FI trials	Independent test FI trials

Fault model: one random bit flip in an activation at a linear-layer invocation.

High coverage with a few monitored layers

Top: full-feature detector. Bottom: selected-layer detector.

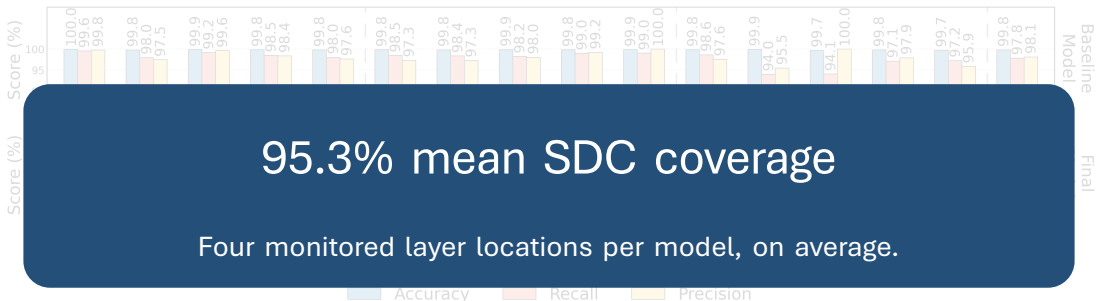


SELECTED DETECTOR / MEAN RESULTS

99.6% accuracy **95.3% recall** **96.5% precision**

High coverage with a few monitored layers

Top: full-feature detector. Bottom: selected-layer detector.



95.3% mean SDC coverage

Four monitored layer locations per model, on average.

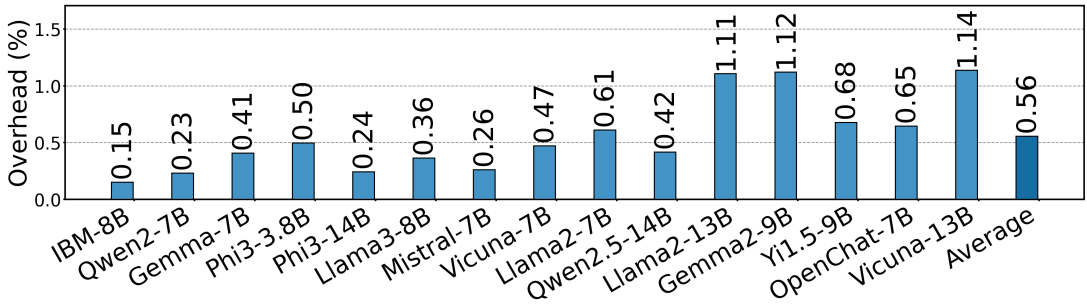
SELECTED DETECTOR / MEAN RESULTS

99.6% accuracy

95.3% recall

96.5% precision

Monitoring cost: 0.56% on average



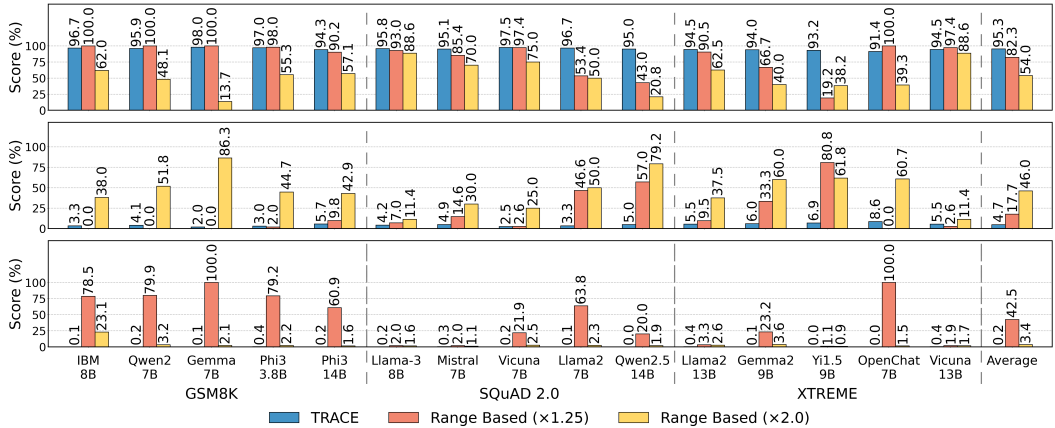
0.56%

mean latency overhead

0.15–1.14% range

Only selected layers are scored during normal inference.

Range-based detection: per-model comparison

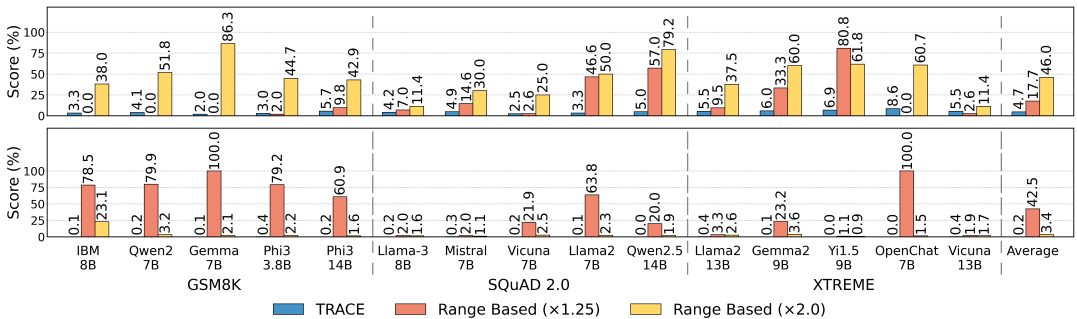


MEAN SDC DETECTION

TRACE	Range ($\times 1.25$)	Range ($\times 2.0$)
95.3%	**82.3%**	**54.0%**

Missed SDCs and false alarms

Range-based protection must balance missed corruption against unnecessary value clipping.

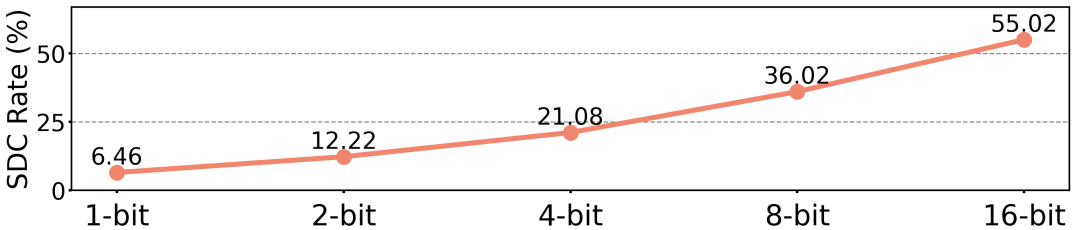


TRACE means: 4.7% missed SDCs; 0.2% reported false alarms.

Tight bounds can clip benign values; loose bounds can miss corruption.

More flipped sites cause more output changes

Multiple bits are flipped at independently selected sites within one inference.



Injected-run SDC frequency rises from 6.46% to 55.02%.

Detection remains high under multi-bit faults

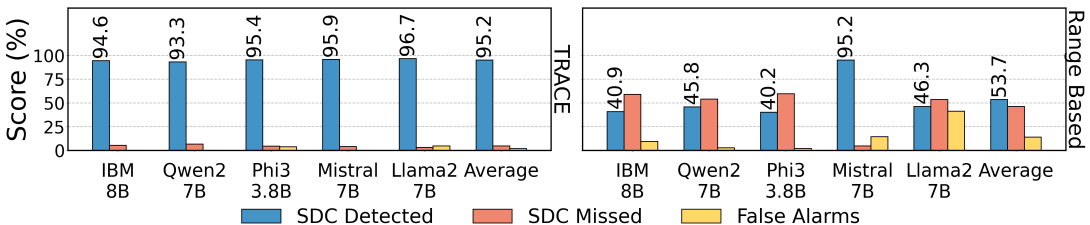
Columns indicate the number of flipped sites: 2, 4, 8, and 16.

	SDC detected (%)				False alarms (%)			
Model	**2**	**4**	**8**	**16**	**2**	**4**	**8**	**16**
IBM-8B	95.00	94.29	95.38	93.46	0.00	0.00	0.74	1.08
Qwen2-7B	96.30	98.08	95.65	96.58	0.00	0.68	1.85	0.00
Phi3-3.8B	90.48	95.24	97.06	93.22	0.00	0.63	0.00	3.66
Mistral-7B	98.46	93.75	96.67	96.00	0.00	0.00	0.71	2.00
Llama2-7B	94.59	94.00	97.30	93.58	0.00	0.00	0.79	1.10

90.48–98.46% detection; reported false alarms 0–3.66%.

Multi-dataset performance

Profile and train jointly on GSM8K, SQuAD2.0, and XTREME.



TRACE: 95.2% mean coverage with 0.40% mean monitoring overhead.

Range ($\times 1.25$): 53.7% mean coverage in the pooled-workload experiment.

Three takeaways

01

Errors leave observable symptoms

Output-changing faults produce structured activation deviations.

02

A few layers retain useful evidence

Offline selection reduces recurring observation work.

03

Selective monitoring is inexpensive

95.3% mean coverage at 0.56% mean online overhead.

Thank you

Abdullah Naveed
University of Florida