AUTO-BOTS: Defense Framework Against Decepticon Attack Architecture

CALIFORNIA STATE UNIVERSITY BAKERSFIELD

Dr. Amin Malek, Tom Regpala, John Ramos

Project Overview

This project focuses on developing a defense against the Large Language Model adversarial attack Decepticon as written in the 2023 paper Decepticon: Attacking Secrets of Transformers through understanding their attack vectors. We focus on manipulating the GPU kernels of each LLM in order for the attack to be less effective

Goals/Objectives

- Recreate the Decepticon attack as described in the Decepticon: Attacking Secrets of Transformers paper
- Conduct a literature review on the development of transformers and adversarial machine learning
- Develop a comprehensive defense against the Decepticon attack by manipulating GPU kernels

Benefits and Application

Prevent IP Stealing: Through scrambling the GPU Kernel fingerprints, Decepticon will not be able to discover the correct pretrained model for the LLM.

Security for private information: Through this defense framework we are able to defend against white box adversarial attacks.

Research and Development Platform

Serves as a testbed for research into defense framework that involve GPU kernel manipulation against novel attacks such as the Decepticon architecture.

Figures: BERT Base vs. Defended

Fig. 1: Base Fingerprint:
Each colored dot represents a kernel being executed on the GPU. We capture kernel duration (Y) across 200ms of execution time (X).

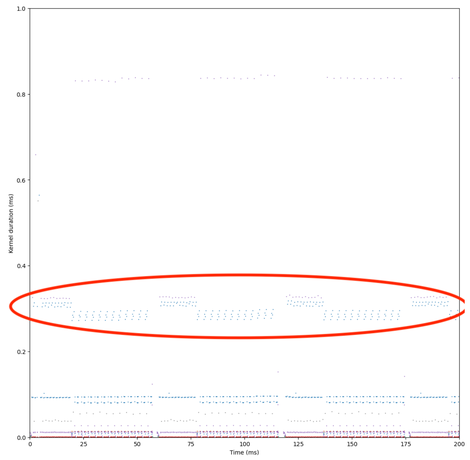


Fig. 2: Constant Fingerprint:
We force the GPU execution for each model to be uniform, making fingerprints indistinguishable.

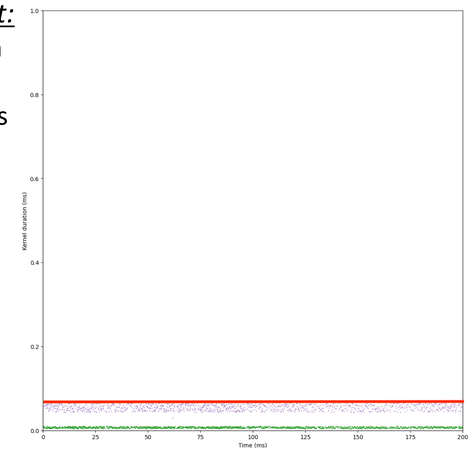
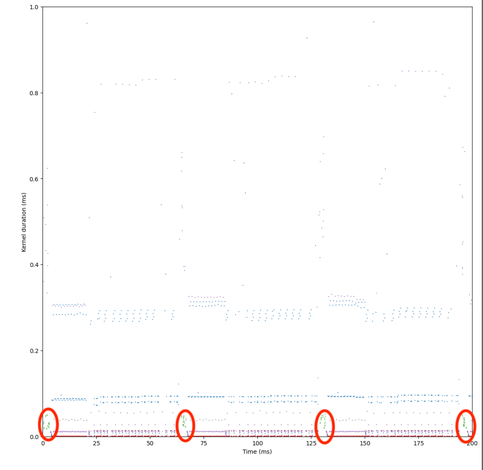


Fig. 3: Jitter Fingerprint:
Bursts and delays of kernels perturb the GPU execution schedule to create noise on the timing graph.



Defense Framework

We create two defense mechanisms:

- **Constant GPU Kernel Fingerprints:** Maintaining a stream of short and busy workloads to create a background noise where the real executions are hidden.
- **Jitter GPU Kernel Fingerprints**: Perturb GPU kernel scheduling to create smears and bursts of noise across execution runs.

Results and Conclusion

BASE CASE: Confusion Matrix

Top-1 accuracy: 0.95 — Top-3 accuracy 1.00

True \ Predictions	bert cased	bert uncased	distilbert	google electra	gpt2	roberta	t5 small	xlnet
bert cased	22	8	0	0	0	0	0	0
bert uncased	1	27	0	0	0	0	0	0
distilbert	0	0	30	0	0	0	0	0
google electra	0	0	0	30	0	0	0	0
gpt2	0	0	0	0	30	0	0	0
roberta	2	0	0	0	0	28	0	0
t5 small	0	0	0	0	0	0	30	0
xlnet	0	0	0	0	0	0	0	30

CONSTANT: Confusion Matrix

Top-1 accuracy: 0.097 — Top-3 accuracy 0.403

True \ Predictions	bert cased	bert uncased	distilbert	google electra	gpt2	roberta	t5 small	xlnet
bert cased	0	0	9	0	0	0	0	0
bert uncased	0	0	8	0	0	0	1	0
distilbert	1	0	7	0	0	0	1	0
google electra	0	0	9	0	0	0	0	0
gpt2	0	0	9	0	0	0	0	0
roberta	1	0	8	0	0	0	0	0
t5 small	1	0	8	0	0	0	0	0
xlnet	1	0	8	0	0	0	0	0

JITTER : Confusion Matrix

Top-1 accuracy: 0.611 — Top-3 accuracy 0.750

True \ Predictions	bert cased	bert uncased	distilbert	google electra	gpt2	roberta	t5 small	xlnet
bert cased	0	0	2	0	0	7	0	0
bert uncased	0	0	9	0	0	0	0	0
distilbert	0	0	0	0	0	9	0	0
google electra	0	0	0	8	0	1	0	0
gpt2	0	0	0	0	9	0	0	0
roberta	0	0	0	0	0	9	0	0
t5 small	0	0	0	0	0	0	9	0
xlnet	0	0	0	0	0	0	0	9

RESULTS & CONCLUSION

Defense	Median(ms)	Top-1	Top-3
Base	305.79	95%	100%
Constant	2,026.72	9.70%	40.30%
Jitter	406.71	61.10%	75.00%

Acknowledgements

The authors gratefully acknowledge financial support from the 2025 Summer Research Apprenticeship for Doctoral Programs, which made this research possible.

"An Introductory Survey of Transformers and Adversarial Machine Learning" Under review at Cyber Security and Applications