BINGHAMTON UNIVERSITY SCHOOL OF COMPUTING SEMINAR SERIES

Abeer Matar A Almalky - PhD Student

LLWRA: Large Language Models Weight Replacement Attack

September 12, 2025
Noon-1
EB T1 or on Zoom

Abstract: The enormous size of large language models (LLMs) makes storing their weights in on-chip memory impractical, requiring off-chip memory that exposes them to memory fault injection attacks. To explore the vulnerability of LLMs against adversarial weight perturbation attacks, we adopt two representative attacks: Bit-Flip Attack (BFA) and Deep-TROJ, both of which exploit bit-flips to degrade accuracy or insert backdoors in vision models. Our experiments reveal that both attacks are significantly less effective on LLMs compared to vision applications. To overcome this limitation, we introduce a novel approach to compromise the performance of LLMs by exploiting a novel fault injection mechanism that introduces targeted bit-flips in page frame numbers of main memory. In the context of main memory, each weight block consists of a set of weights stored at a specific address. Thus, a single bit-flip in the page frame number can replace a target weight block with a new replacement weight block, disrupting the memory translation. However, the algorithmic challenge of creating a formal attack lies in the fact that random weight replacement faults fail to produce detrimental effects on model performance. In this work, we propose LLWRA, which effectively utilizes weight replacement fault injection to degrade the intelligence of state-of-the-art LLMs for the first time. Additionally, we present the ReBlock search algorithm, which efficiently identifies a set of vulnerable target and replacement weight blocks. We evaluate our approach, LLWRA, across three distinct attack objectives: untargeted classification, targeted classification, and untargeted causal modeling. Experimental results demonstrate that LLWRA requires fewer than five attack optimization rounds to reduce classification accuracy to a random guess level and fewer than nine iterations to reduce the causal model into a random generator, making our attack the most lethal weight manipulation attack against LLMs.

Bio: Abeer Almalky is a Computer Science Ph.D. student at Binghamton University, under the guidance of Prof. Adnan Siraj Rakin. She holds an M.S. in Computer Science from Southern Illinois University, and a B.S. in Computer Science and Information System from Umm AlQura University . Her research focuses on Deep Learning, Computer Vision, and the security of generative models.