Wen Cheng

wcheng@smail.nju.edu.cn · Nanjing University, Nanjing, China ·

Education Background

Nanjing University 2022 – Now
Master Computer Science *Admission with Exam Exemption*

Hefei University of Technology 2018 – 2022
Bachelor Computer Science and Technology *Rank #1 for Academical Recommendation*

Works

Security Attack on LLM-based Code Completion Tools — arXiv preprint

***Wen Cheng**, Ke Sun, Xinyu Zhang, Wei Wang*

The LLM-based Code Completion Tools (LCCTs) are emerging and appealing to millions of users while current researcher overlooked the inherent risk posed by their backend LLMs. This work comprehensively investigates this aspect by developing targeted attack methodologies on LCCTs, focusing on jailbreaking and training data extraction attacks. We achieve a 99.4% success rate in jailbreaking attacks on GitHub Copilot and extract sensitive data, including 54 email addresses and 314 physical addresses. The study also demonstrates the effectiveness of these attacks on general-purpose LLMs, highlighting significant security vulnerabilities in modern LLMs' handling code.

USee: Ultrasound-based Device-free Eye Movement Sensing — Under Reviewing

***Wen Cheng**, Mingzhi Pang, Haoran Wan, Shichen Dong, Dongxu Liu, Wei Wang*

In the realm of human-computer interaction, eye movement plays a pivotal role. In this work, we push the limit of sensing by introducing USee, which, for the first time, enables the sensing of subtle eye movements, specifically, saccades in a device-free manner. Additionally, we unveil the intricate relationship between minor movements and decomposed residual signals, making detecting such nuances achievable. We implement USee on COTS devices, and comprehensive experiments have substantiated its outstanding performance.

QAQ: Quality Adaptive Quantization for LLM KV Cache — arXiv preprint

*Shichen Dong, **Wen Cheng (co-first)**, Jiayu Qin, Wei Wang*

With the increasing demand for longer context in LLMs, a notable challenge in model deployment arises from the linear expansion of the Key-Value (KV) cache with the context length. Building on three crucial insights, particularly our groundbreaking discovery of the differential sensitivity of Key cache and Value cache to quantization, we propose the QAQ quantization architecture. Experimental results demonstrate that QAQ is capable of compressing the KV cache footprint by nearly 10 × with negligible accuracy loss. QAQ significantly alleviates the practical challenges associated with deploying LLMs, unlocking new possibilities for applications requiring extended context.

W2KPE: Keyphrase Extraction with Word-Word Relation — ICASSP 2023

***Wen Cheng**, Shichen Dong, Wei Wang*

In this work, we transfer the word-word relations to the key phrase extraction task. By combining our proposed techniques, including sentence fusion, keyphrase encoding, and a combined loss function, we establish an innovative pipeline that achieves notable advancements compared to the baseline. The methodology presented in this paper enables us to secure the first place in the ICASSP 2023 MUG Challenge.

Experience

Microsoft Research Asia - Shanghai 2024.8 - Now
Research Internship Wireless Group

ByteDance Data Platform 2022.1 - 2022.5
Internship Software Engineer

Nanjing University 2023 Spring
Teaching Assistant Computer Architecture

Selected Awards

Nanjing University Distinguished Graduate Student	2023
Graduation with Honor: Excellent College Graduate of Anhui Province	2022
Undergraduate President Award (Top 30 of 8000)	2022
Undergraduate China National Scholarship	2020, 2021
Provincial Second Prize, China Collegiate Programming Contest (CCPC)	2020, 2021
Meritorious Winner, MCM/ICM	2020

Others

- Personal website: https://sensente.github.io/