香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE
計算機科學系

TMLR Young Scientist SEMINAR

2024 SERIES

Trustworthy Machine Learning and Reasoning Group

Mr. Sizhe Chen

PhD student,
Computer Science,
University of California, Berkeley.

Date: 08 Nov 2024 (Friday)

Time: 09:00 – 10:00 (HKT)

Meeting: https://meeting.tencent.com/dm/VKUszYC7iTod

Prompt Injection Defenses by Structured Queries and Alignment Training

ABSTRACT

Recent advances in LLMs enable exciting LLM-integrated applications to perform text-based tasks. To accomplish these tasks, the LLM often uses external data sources such as user documents, web retrieval, results from API calls, etc. This opens up new avenues for attackers to manipulate the LLM via prompt injection. Adversarial prompts can be carefully crafted and injected into external data sources to override the user's intended instruction and instead execute a malicious instruction. We cover two proposed defenses against prompt injections.

We first introduce structured queries, a general approach to defend against prompt injection by separating prompts and data into two channels. We implement a system that supports structured queries (StruQ). This system is made of (1) a secure front-end that formats a prompt and user data into a special format, and (2) a specially trained LLM that can produce high-quality outputs from these inputs.

We also show that alignment can be a powerful tool to make LLMs much more robust against prompt injection. Our second method---SecAlign---first builds a preference dataset by simulating prompt injection attacks and constructing pairs of desirable and undesirable responses. Then, we apply existing alignment techniques to fine-tune the LLM to be robust against these simulated attacks.

BIOGRAPHY

Sizhe Chen is a Computer Science Ph.D. student at UC Berkeley with Prof. David Wagner, supported by the Meta-BAIR and Google-BAIR funding. Previously, Sizhe got his M.Eng. and B.Eng. from Shanghai Jiao Tong University. Sizhe’s research focuses on AI security in real-world applications, and he is currently working on prompt injection defenses for secure LLM systems. Sizhe has also studied transfer, query, and poisoning attacks against vision models.

ENQUIRY

Email: bhanml@comp.hkbu.edu.hk