上海交通大学自动化系
Department of Automation
Shanghai Jiao Tong University

系统控制与信息处理教育部重点实验室
Key Laboratory of System Control and Information Processing, Ministry of Education

学术报告会

时间： 2024年11月12日（周二） 14：00
地点： 腾讯会议：954-195-516
会议密码：241112

Prompt Injection Defense by Structured Queries and Secure Alignment

陈思哲　　博士研究生
加州大学伯克利分校 计算机系

摘要:

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 introduce two causes for prompt injections: (1) LLM input has no separation between prompt and data; (2) LLMs are trained to follow instructions anywhere in their input. For (1), we propose a secure front-end to format the prompt and filtered data with specially-reserved separation delimiters. For (2), we propose structured instruction tuning to SFT (supervised fine-tune) an LLM to follow the intended instruction in the presence of an injected instruction, and secure alignment to DPO (direct preference optimize) an LLM towards following the intended instruction and against following an injected instruction.

StruQ (secure front-end + structured instruction tuning) stops all existing (optimization-free) prompt injections to an attack success rate of <2%. SecAlign (secure front-end + secure alignment) reduces the optimization-based attack success rate by more than 2× from StruQ. Both StruQ and SecAlign preserve the model utility, and induce no training and inference overhead.

报告人简介:

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 with Prof. Xiaolin Huang. 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. https://sizhe-chen.github.io