Xirui Li

4244679380 | xiruili@g.ucla.edu | LinkedIn | GitHub

Education

University of California, Los Angeles (UCLA) Sep.2022-Jul.2024
Master of Electrical and Computer Engineering *Los Angeles, USA*
- GPA: 3.85

Technical University of Munich (TUM) Oct.2018-Jul.2022
Bachelor of Electrical Engineering and Information Technology *Munich, Germany*
- GPA: 3.84

Technical Skills

Programming Languages: Python, C/C++, Matlab, Java, VBA, RestAPI, MySQL, CUDA, NoSQL
Frameworks: Django, Pytorch, HTML5, CSS **System/Tools**: Linux, Docker, Git, Confluence, Jira

Research

Research Assistant Oct.2023-Pressent
University of California, Los Angeles *Los Angeles, California*
- Create an oversensitive benchmark to evaluate MLLMs' safety alignment balance.
- Propose an novel adversarial attack on LLMs by prompt decomposition and reconstruction.

Research Assistant Jul.2021-Dec.2021
Technical University of Munich *Munich, Germany*
- Investigated visual interpretations for DEtection TRansformers and human-in-the-loop workflows of DEtection TRansformers. [GitHub]
- Interpreted attention mechanism in Deformable DEtection TRansformer in Local Interpretable Model-Agnostic Explanations (LIME) architecture.
- Implemented the Deformable DEtection TRansformer in Caltech Pedestrian dataset.[GitHub]

Preprints

DrAttack: Prompt Decomposition and Reconstruction Makes Powerful LLMs Jailbreakers Feb.2024
- A novel jailbreaking method that leverages prompt decomposition and reconstruction to make LLMs generate harmful contents. DrAttack obtains a substantial gain of success rate on powerful LLMs over attackers (15% to 80% on GPT-4).
- Project details could be found in this [website].

MOSSBench: Is Your MLLM Oversensitive to Safe Queries? Jun.2024
- The first comprehensive MLLM benchmark to evaluate the over-sensitivity issue in MLLMs. Prevalent oversensitivity behaviors have been evaluated in SOTA MLLMs with Claude-3 of 76.33% oversensitivity rate.
- Project details could be found in this [website].

Work Experience

Software Engineer Intern Jul.2023–Sep.2023
Mathworks *Natick, MA*
- Developed HTML Verifier for HDL code generation reports for pattern inspection to improved use cases from 1 to 7 with optimized user experience.
- Performed unit test and system test on individual kernel HDL coder QE test constraints and achieved 100% code coverage for the constraints.
- Reduced coupling degree to 0 and improve robustness for kernel HDL coder 5 mostly-used test constraints calculation by refactoring for both Simulink and MATLAB HDL code generation workflow.

Software Engineer Intern Feb.2021-Jul.2021
BMW Group *Munich, Germany*
- Accelerated *Ticket Maker* script from 5 steps to 3 steps for automated Jira tickets generation by optimizing tickets generation logic and algorithm.
- Developed *Budget Viewer* script to generate ticket-related budgets visualization with customization filter based on VBA and Jira Rest-API, which reduce half-day work to 5 minutes.
- Optimized *Budget Viewer*, reducing reaction time by 96.67% (from 5 minutes to 10 seconds) and streamlined functional redundancy of *Ticket Maker* software.