Xirui Li

4244679380 | xirui@g.ucla.edu | xirui-li.github.io | xirui-li | xirui-li | xirui-li

Education

- **University of Maryland, College Park** *Sep 2025 - Now*
  PhD of Computer Science Washington DC, USA
- **University of California, Los Angeles** *Sep 2022 - Jul 2024*
  Master of Electrical and Computer Engineering Los Angeles, USA
- **Technical University of Munich** *Oct 2018 - Jul 2022*
  Bachelor of Electrical Engineering and Information Technology Munich, Germany

Work Experience

- **Mathworks [🌐]** *Jul 2023 – Sep 2023*
  Software Engineer Intern Natick, USA
  - Developed HTML Verifier for HDL code generation reports for pattern inspection that improves 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 zero and improve robustness for kernel HDL coder five mostly-used test constraints calculation by refactoring for both Simulink and MATLAB HDL code generation workflow.
- **BMW Group [🌐]** *Feb 2021 - Jul 2021*
  Software Engineer Intern 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.

First Author Projects and Publications

- **VisualThinker: R1-Zero's "Aha Moment" in Visual Reasoning [ArXiv] [🐙]** *Sep 2025 - Present*
  UMD, UCLA, TurningPoint AI | Zhou, Hengguang, Li, Xirui, et al. Submitted to ICLR 2026
  - Reproduced the first ever visual "aha moment" on a 2B non-SFT model by RLVR, demonstrating emergent reasoning capabilities in vision-language models.
  - Developed multimodal agentic reasoning pipeline integrating visual feedback with RLVR to enable systematic reasoning on visual tasks.
  - Created comprehensive gradient analysis toolkit to monitor memorization versus grokking behaviors during RLVR training, including metrics for effective rank, nuclear norms, and SVD-based statistics.
- **MOSSBench: Oversensitivity in Multimodal LLMs [ICLR 2025][🐙]** *Oct 2023 - Jun 2024*
  UCLA, TurningPoint AI | Li, Xirui, et al. ICLR 2025
  - Proposed the first benchmark to reveal and analyze the oversensitivity prevalence on vision-language models (VLMs) to safe queries.
  - Identified three key types of visual stimuli that trigger oversensitivity in multimodal LLMs: Exaggerated Risk, Negated Harm, and Counterintuitive Interpretation.
  - Revealed widespread oversensitivity across 20 SOTA MLLMs with refusal rates.
- **DrAttack: Decomposition-based Jailbreaking [EMNLP 2024][🐙]** *Oct 2023 - Aug 2024*
  UCLA, TurningPoint AI | Li, Xirui, et al. EMNLP 2024
  - Developed the first decomposition-based jailbreaking attacks on large language models (LLMs), achieving state-of-the-art attack success rate on GPT-4.
  - Implemented prompt decomposition and reconstruction techniques to bypass safety mechanisms in SOTA LLMs
  - Conducted evaluation across multiple commercial and open-source LLMs to demonstrate attack effectiveness.
  - Analyzed defense mechanisms and proposed improved safety alignment strategies based on attack insights.

Skills

- **Programming:** Python (Package: PyTorch, Tensorflow, PySpark, PyTest), Java, SQL, Shell Script, MATLAB, R
- **Languages:** English (Professional), German (Professional), Mandarin (Native)