Luze Sun

University of Pennsylvania, Philadelphia, USA

(267)366-4117 | 13818105854slz@gmail.com | luze-sun-45624a1b7

Education

Master of Science in Engineering, Systems Engineering *Philadelphia, PA*

University of Pennsylvania, School of Engineering and Applied Science *May 2025*

GPA: 3.95/4.0; Selected Courses: Conversations and conversational Bots, Cryptography, Applied Machine Learning, Ethical Algorithm Design

Bachelor of Science in Computer Science *Bristol, UK*

University of Bristol, Faculty of Engineering *June, 2023*

GPA: First Class Honours(74/100); Selected Courses: Image Processing and Computer Vision, Human-Computer Interaction, Computer Systems, Cryptology, Computational Neuroscience, Computer Graphics

Research Experience

Co-researcher. Collaborator: Davis Brown, Advisor: Eric Wong *Philadelphia, PA*

Benchmarking Misuse Mitigation Against Covert Adversaries *Sep 2024 - May 2025*

- Decomposed dangerous questions into less dangerous/more jailbreakable/easier to answer so that a less capable model can answer
- Developed Benchmarks for Stateful Defenses (BSD), a pipeline generating challenging questions consistently refused by frontier models and unsolvable by weaker models, to measure misuse uplift
- Introduced stateful detection mechanisms capable of identifying covert adversarial queries, showing substantial improvements over standard single-query detectors
- Achieved sota results, increasing misuse uplift effectiveness on models like GPT-4o and Claude-3.5 Sonnet, highlighting vulnerabilities in current safety measures

Research Assistant. Advisor: Florian Tramèr *Zürich, Switzerland*

ETH Summer Research Fellowship *Jul 2024 - Apr 2025*

- Investigated the utility degradation (**Jailbreak Tax**) of large language models (LLMs) following adversarial jailbreak attacks, demonstrating substantial performance drops using carefully constructed benchmarks.
- Developed novel benchmark suites including EvilMath and UnicornMath, transforming benign questions into harmful contexts to rigorously measure model robustness post-jailbreak.
- Implemented comprehensive evaluation scripts in Python to systematically assess jailbreak effectiveness and quantify the utility loss across various attack methods (e.g., GCG, PAIR, TAP)
- Ensured fine-tuned models maintained their intended functionality and safety measures post-jailbreak

Co-researcher. Collaborator: Xuan Jiang *Berkeley, CA*

LPSim (Large (Scale) Parallel (Computing) regional traffic Simulation) *Nov 2023 – Sep 2024*

- Implemented a discrete time-driven simulation platform with a highly parallelized GPU implementation using cuda
- Allocated graph information across multiple GPUs and manages the spatio-temporal data efficiently
- Ran benchmark test in AWS with docker and validated the significant advantages in using multiple GPUs over a single GPU setup, including scalability and efficiency in handling large-scale traffic simulations
- Enhanced LPSim by introducing multi-mode including personal and public transportation, making the simulation more realistic

Human Computer Interaction Research. Supervisor: Elaine Czech *Bristol, UK*

Applying Technology in a Hybrid Fashion to Create Dementia-Inclusive Community Spaces *Oct 2022 – May 2023*

- Led a collaborative effort with patients, caregivers, and experts, resulting in the development of a dementia-friendly technology framework, improving community space accessibility for over 100 dementia patients
- Co-created a design framework reflecting the needs and motivations of individuals with dementia and factors amplifying these motivations
- Formulated explicit design recommendations for improving dementia-friendly technology solutions for community space access

Selected Projects

Application of SLAM and Path Planning for Indoor Navigation using ROS *Philadelphia, PA*

ESE 6500 Learning in Robotics with DR. Pratik Chaudhari *Jan 2024 - May 2024*

- Optimized ORB SLAM2 for dense point cloud and optimized grid maps, integrated A* and TEB algorithms for path planning
- Conducted simulations with adaptive Monte Carlo localization in Gazebo and implemented the improved SLAM, localization, and path planning algorithms on a Mecanum-wheeled mobile robot

AI Workflow *Bristol, UK*

Backend developer and Leader, Advised by IBM Master Inventor John McNamara *Apr 2022 - Mar 2023*

- Developed a tool for integrating and automating actions using third-party applications, leveraging IBM Watson for sentiment analysis on social media content
- Implemented the application on Minikube and AWS/IBM cloud infrastructure, integrating with a database and using React for workflow storage and execution

Game of Life *Bristol, UK*

COMS 20001 Computer System with DR. Sion Hannuna *Sep 2021 - Dec 2021*

- Implemented Conway's Game of Life using Golang, featuring multi-threaded processing on a local machine while successfully managing concurrency issues
- Expanded the project into a distributed system using AWS nodes, enabling collaborative calculation and communication of game states across a network, with additional optimizations for improved efficiency

Honors & Awards

2025	**Master's TOP GPA Award,** UPenn Electric and System Engineering Department	*Philadelphia, PA*
2024	**ETH Summer Student Research Fellowship - CHF 4000,** ETH Department of Computer Science	*Zürich, Switzerland*
2022	**Oracle Summer research internship program - £ 5000,** Oracle	*Bristol, UK*

Publications

1 Kristina Nikolić, **Luze Sun**, Jie Zhang, Florian Tramèr *The Jailbreak Tax: How Useful Are Your Jailbreak Outputs?* International Conference on Machine Learning (ICML), 2025. arXiv:2504.10694

2 Davis Brown, Mahdi Sabbaghi, **Luze Sun**, Alexander Robey, George Pappas, Eric Wong, Hamed Hassani *Benchmarking Misuse Mitigation Against Covert Adversaries* Under submission to NeurIPS 2025.

Internship Experience

Software Engineering Intern *Bristol, UK*

Oracle *Jun 2022 - Sep 2022*

- Spearheaded the successful implementation of FAASC/CASE, an innovative tool that generated real-time data from **Oracle Cloud Infrastructure**, allowing for efficient server performance monitoring and parallel request processing with automatic updates every 5 minutes.
- Engineered and implemented **a robust server-side retry program**, ensuring uninterrupted requests; reduced downtime by 75%, resulting in improved user experience and increased customer satisfaction.

Junior Intern in Software Development *Shanghai, China*

Dell *Jul 2021 - Sep 2021*

- Collaborated with the DEEP group to validate the deployment of **Clusternet in Kubernetes**, improving cluster connectivity efficiency by 25% and simplifying the integration process across multiple projects.
- Developed **a full-stack application for VxRail node performance monitoring**, delivering minute-by-minute updates and boosting monitoring efficiency by 30%, thereby streamlining decision-making for system administrators.

Teaching Assistant Experience

University of Bristol *Bristol, UK*

Teaching Assistant *Sep 2021 - May 2023*

- Teaching assistant for senior-level courses including Computer Sytems (Dr Sion L Hannuna), Interaction and Society (Dr. Paul Marshall) and Cryptology (Dr Francois Dupressoir)
- Held weekly office hours and lab course to help students overcome understanding and implementation challenges. Help teachers design and mark the exam questions for 21 pages.

University of Pennsylvania *Philadelphia, PA*

Teaching Assistant *Aug 2024 - May 2025(Expected)*

- Teaching assistant for graduate-level courses including Statistics for Data Science (Prof Hamed Hassin) and Learning in Robotics (Prof Pratik Chaudhari)
- Held weekly office hours, recitation and online Q&A session to help students overcome course materials, example questions, mock exams and homework. Help teachers design and mark the exam questions for 15 pages.