COLLEGE STATION, TX, US • YUXUANZ0526@GMAIL.COM • +1 (412)-979-8762 • HTTPS://MICHAELZHANG05.GITHUB.IO

YUXUAN ZHANG

PROFESSIONAL SUMMARY

Ph.D. researcher specializing in Agentic AI Security, Trustworthy AI, AI for Security, and Network Security. Conducts cutting-edge research on developing defense-in-depth frameworks for LLM security, specifically focusing on prompt injection defense, semantic reasoning, and in-network machine learning. Skilled in adversarial red teaming, alignment safety, and high-performance systems programming to enable the secure deployment of scalable AI in real-world environments. Strong foundation in systems programming, neural network optimization, and GPU-accelerated computing. Dedicated to advancing secure and efficient AI deployment in real-world environments.

TECHNICAL SKILLS

- **AI Security & Trustworthy AI:** Prompt injection defense, adversarial machine learning, jailbreaking mitigation, semantic intent invariance, robust model evaluation, backdooring detection, and AI-driven workflow security.
- **Large Language Models (LLMs) & Generative AI:** LLM fine-tuning, Reinforcement Learning from Human Feedback (RLHF), Prompt Engineering, retrieval-augmented generation (RAG), LLM-based policy generation.
- **LLM Red Teaming & Alignment:** Manual and automated red teaming, prompt leaking, jailbreak benchmarking, and safety alignment techniques like PPO (Proximal Policy Optimization) and DPO (Direct Preference Optimization).
- **Systems & Programmable Networking:** Switch-native neural networks, programmable data planes, in-network machine learning, hardware-efficient model abstraction, and GPU-accelerated computing (CUDA).
- **Network Security & Distributed Systems:** Asynchronous Byzantine Fault Tolerance (BFT), verifiable information dispersal, digital signatures, consensus protocols, interactive URL triage, and network traffic analytics.
- **Machine Learning Frameworks & HPC:** PyTorch, TensorFlow, Scikit-learn, large-scale distributed model training, High-Performance Computing (HPC) clusters, and SLURM workload management.
- **Programming & Software Development:** Python (NumPy, pandas), C, C++, Golang, Flask, Verilog (Hardware Description), Git, Bash/Shell scripting, MATLAB , MySQL, and Linux/WSL environments.
- **Data Science & Signal Processing:** Real-time analytics, A/B testing evaluation, signal enhancement, multi-modal deep learning, time-series data analysis, and quantitative performance reporting.

EDUCATION

PH.D. IN COMPUTER SCIENCE (PI: DR. GUOFEI GU) — **Sep 2024 - May 2028 (Expected)**
Texas A&M Unversity — ***College Station, TX, USA***

M.SC. IN COMPUTER SCIENCE (PI: LONGFEI SHANGGUAN) — **Sep 2022 - May 2024**
University of Pittsburth — ***Pittsburgh, PA, USA***

B.E. IN COMPUTER SCIENCE AND TECHNOLOGY — **Sep 2018 - Jun 2022**
Huazhong University of Science and Technology — ***Wuhan, China***

EMPLOYMENT HISTORY

GRADUATE RESEARCH ASSISTANT — **Sep 2024 - Present**
Texas A&M Unversity, Supervisor: Dr. Guofei Gu — ***College Station, TX, USA***

- Developed novel defense framework for prompt injection detection by identifying semantic intent invariance, significantly hardening LLM-based applications against adversarial manipulation.
- Designed a decoupled adjudication agent system for interactive URL triage, ensuring secure execution in autonomous AI-driven workflows by isolating and auditing high-risk external interactions.
- Designed and executed comprehensive red-teaming benchmarks to expose critical vulnerabilities in current AI defenses, specifically measuring and mitigating failures in LLM reasoning and networking policy generation.
- Improved LLM reliability for intent-based networking by developing robust agentic framework to generate and validate complex networking policies, bridging the gap between natural language intent and secure system configuration.
- Architected system utilizing switch-native neural networks and hardware-efficient model abstractions to enable high-precision, scalable machine learning inference directly on programmable data plane for network security defense.
- Collaborate with multi-disciplinary research team to advance secure and efficient AI deployment, contributing to multiple manuscripts currently under review at top-tier security and networking conferences.

GRADUATE TEACHING ASSISTANT **Sep 2025 - Present**
Texas A&M Unversity, Department of Computer Science ***College Station, TX, USA***

- CSCE 310: Database Systems
  - ◊ Facilitated lectures and provided technical mentorship for 100+ students on Database System concepts, including schema design, query optimization, and the implementation of complex SQL queries.
  - ◊ Assisted in the design and implementation of comprehensive database systems, developing both front-end interfaces and back-end architectures to support course objectives.

RESEARCH ASSISTANT **Jan 2023 - May 2024**
University of Pittsburth, Supervisor: Dr. Longfei Shangguan ***Pittsburgh, PA, USA***

- Proposed an innovative passive mobile sensing methodology for roadside parking availability information retrieval.
- Designed a sophisticated end-to-end Deep Neural Network (DNN) that fuses asynchronous audio signals and image data to achieve high-accuracy, drive-by sensing for real-time roadside parking detection.
- Implemented feature-level fusion strategies within the DNN to reconcile disparate data streams, enabling the system to maintain detection accuracy even in challenging conditions such as low visibility or high acoustic clutter.
- Developed a robust data preprocessing pipeline to denoise and enhance passive mobile sensing data, significantly improving the signal-to-noise ratio in high-interference urban environments.

DATA ANALYST INTERN **Jul 2021 - Sep 2021**
Tencent (TCEHY), Online Video Division ***Beijing, China***

- Architected and deployed a real-time analytics system that streamlined the evaluation of large-scale A/B tests, reducing reporting latency and enabling rapid, data-driven feature iterations for the Online Video Division.
- Designed and executed rigorous experimental frameworks, utilizing hypothesis testing and multivariate regression to quantify the impact of UI/UX changes on key user engagement metrics.
- Collaborated with product managers and engineering teams to define success metrics and optimize recommendation algorithms, contributing to measurable improvements in user retention.
- Recognized by the “Head of Data & Data Science at Tencent News and Tencent Online Video Business Unit” as top 10% of interns for exceptional performance and contribution.

PUBLICATIONS/MANUSCRIPTS

[1] PromptSleuth: Detecting Prompt Injection via Semantic Intent Invariance
Yuxuan Zhang*, Mengxiao Wang, Guofei Gu,* Under Review

[2] NetInspector: Measuring and Improving LLM Capabilities for Reliable Intent-Based Networking Policy Generation
Yuxuan Zhang*, Hongxing Hu, Guofei Gu,* Under Review

[3] TraceScope: Interactive URL Triage via Decoupled Checklist Adjudication
Haolin Zhang, William Reber, ***Yuxuan Zhang****, Guofei Gu, Jeff Huang,* Under Review

[4] Small Planes, Big Models: Scaling Robust Neural Networks on Programmable Data Planes
Huancheng Zhou, ***Yuxuan Zhang****, Guofei Gu,* Under Review

[5] JUMBO: Fully Asynchronous BFT Consensus Made Truly Scalable
Hao Cheng, Yuan Lu, Zhengliang Lu, Qiang Tang, ***Yuxuan Zhang****, Zhenfeng Zhang,* Co-primary authors, IEEE Transactions on Dependable and Secure Computing

[6] Drive-by sensing for on-street parking spot detection
Yuxuan Zhang*, Longfei Shangguan,* Master's Thesis, University of Pittsburgh

PROFESSIONAL SERVICE

- Program Committee Member: The Web Conference (WWW), 2026
- Journal Reviewer: Transactions on Dependable and Secure Computing (TDSC)
- External Reviewer: ACM Conference on Computer and Communications Security (CCS), 2025, 2026
- External Reviewer: USENIX Security Symposium, 2025, 2026
- External Reviewer: Network and Distributed System Security (NDSS) Symposium, 2026