Yuanda Wang

517-721-9800 | yuandawang.msu@gmail.com | yuandaw.github.io | LinkedIn | Google Scholar

Education

Michigan State University East Lansing, MI, USA
Doctor of Philosophy in Computer Science *Jan. 2020 - May 2025 (Expected)*

- Advisor: Dr. Qiben Yan
- Research area: Security and Privacy, Large Language Models , Speech AI, Adversarial Machine Learning.

North China Electric Power University Beijing, China
M.S. in Electrical Engineering *2016 - 2019*

Xi'an Jiaotong University Xi'an, China
B.S. in Electrical Engineering *2012 - 2016*

Work Experience

ByteDance Inc. San Jose, CA, USA
AI Security Research Scientist Intern *Feb. 2025 – Present*

- Fine-tune LLM to develop universal solutions for network security, such as DDoS and Bot traffic detection and defense.
- Launch fine-tuning and tokenizer adaptation on foundation models to fit downstream tasks and use distillation to reduce the model size and computational load.

Samsung Research America Mountain View, CA, USA
Research Scientist Intern *Sep. 2022 – Dec. 2022*

- Investigate machine learning models including speech and speaker recognition for voice assistants, specifically focusing on their robustness and safety.
- Train, debug, and test speech AI models applied on Bixby to enhance its performance and reliability.

Skills

Programming Languages: Python, C++, C, JavaScript, Matlab, SQL.
Machine Learning Frameworks: PyTorch, TensorFlow, Keras, CUDA.
Data Analysis Frameworks: Numpy, Pandas, Jupyter, Digital Signal Processing (DSP).
Machine Learning Skills: Deep Learning Model Design, Speech Synthesis, Speech/Speaker Recognition, Trustworthy AI, Adversarial Machine Learning, Real-time Machine Learning.
Large Language Model (LLM): LLM Fine-tuning, Prompt Engineering, LLM Safety Analysis, LLM-based AI agent.
Operating Systems: Ubuntu, MacOS, Windows.
Cloud Platforms: Google Cloud, AWS, Microsoft Azure.

Highlighted Research Projects

The Dark Side of Human Feedback | LLM Safety

- Uncovers how human feedback can exploit vulnerabilities within LLM training pipelines.
- Demonstrates an attack that successfully poisons LLMs, including GPT and Llama, causing toxic outputs.

ClearMask | Speech AI & Adversarial Machine Learning

- ClearMask is a noise-free defense mechanism that protects speech audio against malicious voice deepfake attacks.
- Prevent over 99% of voice deepfake attacks in a zero-knowledge setup while maintaining audio quality.

ClearAI | Speech AI & Healthcare

- ClearAI is an AI-driven speech enhancement tool to improve the speech quality of Parkinson's disease patients.
- Increases the word recognition rate of hypophonic speech by over 50% in noisy environments.

ToxicChat | LLM & Chatbot Safety

- Proposes a new attack that induces toxic chatbot outputs through multi-turn conversations.
- By fine-tuning a chatbot for attack, ToxicChat achieves over a 60% toxicity activation rate.

VSMask | Speech AI & Adversarial Machine Learning

- VSMask is a real-time defense against voice deepfake attacks for instant communication applications.
- Achieves a 100% protection success rate in a white-box setup without adding latency.

GhostTalk | Mobile Security & Side-channel Attack

- The first attack to inject inaudible voice commands via charging cables to manipulate voice assistants.
- Achieves a 100% attack success rate on nine different COTS phones, including iPhones and Android devices.

Publications

Conference Papers (11)

- *The Dark Side of Human Feedback: Poisoning Large Language Models via User Inputs*
  Bocheng Chen, Hanqing Guo, Guangjing Wang, **Yuanda Wang**, Qiben Yan.
  Under Review
- *ClearMask: Noise-Free and Naturalness-Preserving Protection against Voice Deepfake Attacks*
  Yuanda Wang, Bocheng Chen, Hanqing Guo, Guangjing Wang, Weikang Ding, Qiben Yan.
  Under Review
- *ClearAI: AI-Driven Speech Enhancement for Hypophonic Speech*
  Yuanda Wang, Qiben Yan, Thea Knowles, Daryn Cushnie-Sparrow.
  IEEE International Conference on E-health Networking, Application & Services (**HealthCom**), 2024.
- *WavePurifier: Purifying Audio Adversarial Examples via Hierarchical Diffusion Models*
  Hanqing Guo, Guangjing Wang, Bocheng Chen, **Yuanda Wang**, Xiao Zhang, Xun Chen, Qiben Yan, Li Xiao.
  International Conference on Mobile Computing and Networking (**MobiCom**), 2024. (Acceptance rate: 20.8%)
- *Protecting Activity Sensing Data Privacy Using Hierarchical Information Dissociation*
  Guangjing Wang, Hanqing Guo, **Yuanda Wang**, Bocheng Chen, Ce Zhou, Qiben Yan.
  IEEE Conference on Communications and Network Security (**CNS**), 2024.
- *Understanding Multi-Turn Toxic Behaviors in Open-Domain Chatbots*
  Bocheng Chen, Guangjing Wang, Hanqing Guo, **Yuanda Wang**, Qiben Yan.
  The 26th International Symposium on Research in Attacks, Intrusions and Defenses (**RAID**), 2023.
- *PhantomSound: Black-Box, Query-Efficient Audio Adversarial Attack via Split-Second Phoneme Injection*
  Hanqing Guo, Guangjing Wang, **Yuanda Wang**, Bocheng Chen, Qiben Yan.
  The 26th International Symposium on Research in Attacks, Intrusions and Defenses (**RAID**), 2023.
- *VSMask: Defending Against Voice Synthesis Attack via Real-Time Predictive Perturbation*
  Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, Qiben Yan.
  The 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks (**WiSec**), 2023.
- *SpecPatch: Human-In-The-Loop Adversarial Audio Spectrogram Patch Attack on Speech Recognition*
  Hanqing Guo, **Yuanda Wang**, Nikolay Ivanov, Li Xiao, Qiben Yan.
  The ACM Conference on Computer and Communications Security (**CCS**), 2022. (Acceptance rate: 22.0%)
  Best Paper Honorable Mention
- *GhostTalk: Interactive Attack on Smartphone Voice System Through Power Line*
  Yuanda Wang, Hanqing Guo, Qiben Yan.
  The Network and Distributed System Security Symposium (**NDSS**), 2022. (Acceptance rate: 16.2%)
- *SDR Receiver Using Commodity WiFi via Physical-layer Signal Reconstruction*
  Woojae Jeong, Jinhwan Jung, **Yuanda Wang**, Shuai Wang, Seokwon Yang, Qiben Yan, Yung Yi, Song Min Kim.
  International Conference on Mobile Computing and Networking (**MobiCom**), 2020. (Acceptance rate: 16.1%)

Journal Papers (3)

- *Beyond Boundaries: A Comprehensive Survey of Transferable Attacks on AI Systems*
  Guangjing Wang, Ce Zhou, **Yuanda Wang**, Bocheng Chen, Hanqing Guo, Qiben Yan.
  Under Review
- *A Practical Survey on Emerging Threats from AI-driven Voice Attacks: How Vulnerable are Commercial Voice Control Systems?*
  Yuanda Wang, Qiben Yan, Nick Ivanov, Xun Chen.
  Under Review
- *URadio: Wideband Ultrasound Communication System for Smart Home Applications*
  Qiben Yan, Qi Xia, **Yuanda Wang**, Pan Zhou, Huacheng Zeng.
  IEEE Internet of Things Journal, January 2022.

Awards

Dissertation Completion Fellowship, Michigan State University, 2024.
Best Paper Honorable Mention Award, ACM CCS, 2022.
Student Travel Grant Award, IEEE CNS, 2020.