Yuanda Wang

wangy208@msu.com | yuandaw.github.io | LinkedIn | Google Scholar

Summary

Final-year Computer Science Ph.D. specializing in **Machine Learning, Large Language Models, and Multimodal AI**, with **6 years** of hands-on experience in machine learning research, and **1+ year** of full-time industry experience applying cutting-edge research to production systems.

Education

Michigan State University — East Lansing, Michigan, USA
Doctor of Philosophy in Computer Science — *Graduating in Spring 2026*

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

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

Work Experience

Adobe Research — San Jose, CA, USA
Research Scientist Intern — *Feb. 2026 – Now*

Dolby Laboratories — Sunnyvale, CA, USA
Multi-modal AI Research Intern — *Sep. 2025 – Dec. 2025*

TikTok Inc. — San Jose, CA, USA
AI Security Research Scientist Intern — *Feb. 2025 – Jul. 2025*

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

Projects

LLM-Based Bot Detection for Cloud Security | Tiktok
- Designed and shipped an **LLM-powered bot-detection** capability for TikTok Cloud Security Service, applying instruction-tuned LLMs to classify automated abuse at WAF/edge scale.
- Built an AI/ML data pipeline: transformed raw traffic into session-level behavior representations, engineered model-friendly features, and curated high-quality labels with hard negatives to improve generalization to unseen bot strategies.
- Fine-tuned a Qwen-based model via **SFT + RL**, added score calibration/thresholding for production precision, and established offline replay + adversarial evaluation; achieved **>95% detection accuracy** against state-of-the-art bot scripts with **zero offline false positives** and filed a patent.

AI Agent for LLM Security Literature Intelligence | Tiktok
- Architected and built an internal literature-intelligence system for TikTok's LLM Security Research team, addressing the scale of daily arXiv publications and enabling efficient frontier-technology tracking.
- Developed a custom **agentic system** on top of **Seed** LLM to orchestrate topic-based paper retrieval, reading, and structured summarization.
- Designed the end-to-end pipeline and storage layer: metadata/impact enrichment (authors, affiliations, citation signals when available) and a searchable internal database to support triage and trend analysis.

Robust Music Watermark for Copyright Protection | Dolby Lab.
- Designed a high-performance audio watermarking system for Dolby Laboratories as part of an AI music generation pipeline, enabling copyright protection and content fingerprinting; introduced a **cross-attention-based architecture** to improve watermark detection and decoding accuracy.
- Hardened the watermark against audio tampering (e.g., music source separation/MSS) via MSS-targeted robustness training; improved **detection accuracy by 7%** and **decoding accuracy by 4%** under MSS processing while preserving audio quality.

ToxicChat: Black-Box Multi-Turn Red-Teaming for Commercial LLMs | LLM Safety Research in MSU
- Designed a black-box multi-turn red-teaming system using a fine-tuned GPT-2 agent to adaptively craft dialogue attacks against commercial LLMs for toxic-output elicitation.
- Built an automated attack-and-evaluation pipeline (trajectory replay, toxicity scoring, success metrics) and achieved up to a **67% toxicity activation rate** to inform guardrail hardening.

Skills

Programming Languages: Python, C++, C, JavaScript, SQL.
LLM Skills: LLM post-training (SFT, reasoning), Prompt Engineering, Retrieval-Augmented Generation (RAG), Multi-modal LLM, AI agent.
Machine Learning: Designing and optimizing deep learning systems (CNN, RNN, Transformer, VAE) for real-world applications across supervised and self-supervised learning paradigms.
ML Frameworks & Toolkits: PyTorch, TensorFlow, CUDA, NumPy, JAX, Pandas, Jupyter.
Cloud Platforms: AWS, Google Cloud Platform, Microsoft Azure.

Publications

Conference Papers (12)

- *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
- *AUDIO WATERMARK: Dynamic and Harmless Watermark for Black-box Voice Dataset Copyright Protection*
  Hanqing Guo, Junfeng Guo, Bocheng Chen, **Yuanda Wang**, Xun Chen, Heng Huang, Qiben Yan, Li Xiao
  The 34th USENIX Security Symposium (**USENIX Security**) 2025
- *ClearMask: Noise-Free and Naturalness-Preserving Protection against Voice Deepfake Attacks*
  Yuanda Wang, Bocheng Chen, Hanqing Guo, Guangjing Wang, Weikang Ding, Qiben Yan.
  The 20th ACM ASIA Conference on Computer and Communications Security (**AsiaCCS**) 2025
- *ClearAI: AI-Driven Speech Enhancement for Hypophonic Speech*
  Yuanda Wang, Qiben Yan, Thea Knowles, Daryn Cushnie-Sparrow.
  IEEE International Conference on E-health Networking, Applications & 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.