Yifu **Wu**

Applied Scientist · Machine Learning Engineer · Research Scientist

The Greater Seattle Area, WA 98072

(330) 906-2730 | wuyifu2f@gmail.com | nnonno | yifu-wu-09609779 | Yifu Wu

"Bridging theory and practice in AI, distributed systems, and high-performance computing"

Professional Summary

Applied Scientist and Research Engineer with 14+ years of experience spanning NLP, distributed systems, HPC, cybersecurity, and AI infrastructure. Proven track record in both academic research (15+ peer-reviewed publications, 2 US patents) and industry applications (Amazon, healthcare AI). Expertise includes: LLM optimization achieving 50% latency reduction, distributed ML on embedded systems and spacecraft networks, HPC storage optimization on 32-node clusters, blockchain-based decentralized computing, and production-scale multilingual classifiers. Strong foundation in both theoretical research and practical system implementation across diverse domains including IoT, smart grids, space exploration, clinical NLP, and cloud infrastructure.

Professional Experience

Amazon

Bellevue, WA

Applied Scientist

Apr. 2025 - Present

- Optimized LLM inference infrastructure achieving 50% latency reduction through sglang and vLLM implementation with continuous batching and dynamic request scheduling on multi-GPU setups.
- Deployed production ML endpoints serving multilingual models (Sparse RoBERTa, Qwen3-0.6B) with optimized resource allocation for concurrent request handling on AWS infrastructure.
- Built scalable inference pipeline processing high-throughput production traffic, improving precision by 60% while maintaining operational stability across distributed systems.
- Developed guardrail models (PII, Policy Violation, Sensitivity) for Alexa's core LLM to ensure policy-compliant responses across multiple languages.
- Integrated advanced batching strategies for GPU workload optimization, enabling efficient concurrent request processing within memory constraints.

University of Colorado Anschutz Medical Campus

Aurora, CO

NLP Data Scientist

Aug. 2024 - Mar. 2025

- Designed and implemented LLM pretraining pipeline over large-scale healthcare knowledge bases with distributed training infrastructure.
- Built clinical diagnosis reasoning system using knowledge graph networks with LLM reasoning and response re-ranking for differential diagnosis.
- Contributed to LogosKG: hardware-optimized scalable knowledge graph retrieval system for efficient distributed inference.
- Built retrieval and reranking systems for large-scale medical document processing with optimized I/O patterns.
- Coauthored "Zero-shot Large Language Models for Long Clinical Text Summarization with Temporal Reasoning" (EMNLP 2025).

AI Newsletter Startup (Head of Data Science: Jay Wang, ex-Kuaishou)

Remote (Bay Area Time Zone)

Machine Learning Engineer (Remote Intern)

Jun. 2024 - Aug. 2024

- Designed AI agentic workflow for automated newsletter generation from web-scraped news across multiple topics (politics, technology, economics).
- Implemented DBSCAN-based clustering system for topic-based news organization and built vector database integration (Weaviate) for efficient retrieval.
- Engineered RAG system using LangChain with GPT-4o for multi-document summarization and newsletter generation with metadata-based filtering.
- Built LLM-powered content critique and refinement pipeline with automated email distribution using Python packages.

Purdue University

West Lafayette, IN

Research Assistant - HPC, Distributed Systems & AI Infrastructure

Dec. 2019 - May 2024

- Designed resilient decentralized machine learning system for NASA spacecraft networks in delay-tolerant environments using blockchain-based coordination.
- Implemented bandwidth allocation algorithms using ADMM (Alternating Direction Method of Multipliers) for heterogeneous network resource optimization in space missions.
- Built federated meta-learning framework for edge/fog computing with limited distributed compute resources, handling synchronization and fault tolerance.
- Built distributed machine learning system on embedded devices (4-16 Raspberry Pi and Nvidia Jetson Nano/TX2 nodes) using Torch-TensorRT framework.
- Solved synchronization and fault tolerance challenges in distributed training while maintaining computation precision using Integer-Vector Homomorphic Encryption scheme.
- Developed GPT-3.5 agentic coding workflow for autonomous robot control code generation, integrating LLM prompting with RL baselines (PPO, DDPG, HER).
- Built distributed control system for multi-human-multi-robot collaboration with real-time visualization and low-latency control protocols over network infrastructure.
- Implemented automated networking and monitoring system for distributed IoT sensor networks in greenhouse environments.
- Published multiple papers in IEEE and top-tier venues on distributed systems, cybersecurity, and AI applications.

Iowa State University

Ames, IA

Research Assistant - HPC Storage Systems

Aug. 2017 - Dec. 2017

- Optimized HPC distributed storage I/O throughput prediction using LSTM/RNN models on Lustre parallel filesystem across 32-node clusters (500GB-4TB per node).
- Predicted both throughput and IOPS metrics for large-scale distributed storage systems, enabling better workload scheduling and resource allocation.
- Used SLURM workload manager for submitting and managing distributed computing tasks across HPC clusters.
- Achieved significant performance improvements in distributed file system optimization (published: arXiv:2301.06622).

University of Akron

Akron, OH

Research Assistant - Networking Security & Smart Grid Systems

Aug. 2015 - Aug. 2017; Jan. 2018 - Jul. 2019

- Designed hardware-in-the-loop (HIL) testing infrastructure for large-scale renewable energy systems with distributed monitoring.
- Implemented attack detection and resilient communication middleware for smart grid systems using machine learning and software-defined networking.
- Developed blockchain-powered decentralized computing system for crowdsourced IoT applications (2 US patents granted).
- Published papers in IEEE Transactions on Emerging Topics in Computing, IEEE Internet of Things Journal, and IEEE conferences.

Tsinghua IT Training School

Xiangyang, Hubei, China

Instructor

Jan. 2015 - Jun. 2015

- Taught C programming fundamentals and web development (HTML/HTML5) to students in vocational training programs.
- Developed course materials and practical exercises for programming instruction.

Haite Measurement and Control Technology Co.

Xiangyang, Hubei, China

Software Engineer

Sep. 2011 - Dec. 2012

- Developed embedded system software using industrial fieldbuses: Modbus, Profibus-PA, and Profibus-DP protocols.
- Implemented measurement and control systems for industrial automation applications.

Education

Purdue University

West Lafayette, IN

Ph.D. in Computer and Information Technology

Aug. 2019 - May 2024

- Dissertation: "Data-driven Computing and Networking Solution for Securing Cyber-Physical Systems"
- Research areas: Distributed machine learning, HPC infrastructure, cybersecurity, resource allocation, AI for robotics
- Doctoral studies at University of Akron (Aug. 2015 - Aug. 2017; Jan. 2018 - Jul. 2019) and Iowa State University (Aug. 2017 - Dec. 2017) in Electrical Engineering and Computer Science, focusing on networking security, distributed ML, HPC storage optimization, and NLP
- Advisor: Dr. Juan Wei-Kocsis (formerly Dr. Juan Wei)

University of Limerick

Limerick, Ireland

M.E. in Electronic and Computer Engineering

Sep. 2013 - Jan. 2015

- Focus: VLSI circuit design and test, embedded systems, hardware-software integration

Harbin Institute of Technology

Weihai, China

B.S. in Automation

Sep. 2007 - Jul. 2011

- Focus: Industrial control systems and automation
- Multiple academic honors including Triple-A Outstanding Student, Third-grade People Scholarship, Excellent League Member

Publications

0.1 Journal Articles

2025 , Kruse, M.; Hu, S.; Derby, N.; Wu, Y.; et al., ”Zero-shot Large Language Models for Long Clinical Text Summarization with Temporal Reasoning,” EMNLP 2025

2025 , Cheng, H.; Wu, Y.; Khatwani, S.; Kruse, M.; Dligach, D.; Miller, T.; Afshar, M.; Gao, Y., ”LogosKG: Hardware-Optimized Scalable and Interpretable Knowledge Graph Retrieval,” ACL ARR (under review)

2020 , Y. Wu, et al., ”DDLPF: A Practical Decentralized Deep Learning Paradigm for Internet-of-Things Applications,” IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9740-9752

2020 , G.J. Mendis, Y. Wu, J. Wei, M. Sabounchi, R. Roche, ”A Blockchain-Powered Decentralized and Secure Computing Paradigm,” IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 4, pp. 2201-2222

0.2 Conference Papers

2025 , M. Soy, L Eu, Y. Wu, J. Zhang, D. Gan, J. Wei-Kocsis, B.C. Min, ”Remote Physical Control for Upgrading Heavy Construction Equipment,” ISARC Proceedings of the International Symposium on Automation and Robotics in Construction

2022 , Y. Wu, J. Wei-Kocsis, ”A Practical and Stealthy Adversarial Attack for Cyber-Physical Applications,” AAAI-22 Workshop on Adversarial Machine Learning and Beyond

2017 , Y. Wu, J. Wei, ”Towards Attack-Resilient Communications for Smart Grids with Software-Defined Networking,” IEEE Power & Energy Society General Meeting, pp. 1-5

2017 , Y. Wu, J. Wei, B.M. Hodge, ”A Distributed Middleware Architecture for Attack-Resilient Communications in Smart Grids,” IEEE International Conference on Communications (ICC), pp. 1-7

2017 , Y. Wu, Y. He, G.J. Mendis, J. Wei, ”A Privacy-Preserving Middleware Mechanism for Smart Grids,” IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)

2016 , Y. Wu, G.J. Mendis, Y. He, J. Wei, B.M. Hodge, ”An Attack-Resilient Middleware Architecture for Grid Integration of Distributed Energy Resources,” IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications

2016 , Q. Gao, Y. Wu, J. Wei, ”Social Community-Based Scheme for Preserving Privacy of Smart Meters,” Proceedings of the Workshop on Communications, Computation and Control for Resilient Smart Energy Systems

2019 , Y. Wu, J. Wei, ”A Domain Knowledge-Enabled Hybrid Semi-Supervision Learning Method,” IEEE Global Conference on Signal and Information Processing (GlobalSIP)

0.3 Book Chapters

2020 , Y. Wu, J. Wei, B.M. Hodge, ”Towards an Adaptive and Attack-Resilient Communication Infrastructures for Smart Grids,” Security of Cyber-Physical Systems: Vulnerability and Impact, pp. 293-323

0.4 Technical Reports & Preprints

2025 , K. Zhou, Y. He, C. Zhong, Y. Wu, ”Real-Time Cascade Mitigation in Power Systems Using Influence Graph Improved by Reinforcement Learning,” arXiv preprint arXiv:2506.08893

2024 , Y. Wu, ”Data-driven Computing and Networking Solution for Securing Cyber-Physical Systems,” Ph.D. Dissertation, Purdue University

2017 , B.S. Hodge, Y. Wu, J. Wei, ”A Distributed Middleware Architecture for Attack-Resilient Communications in Smart Grids,” National Renewable Energy Laboratory (NREL), Golden, CO (United States)

Patents

2024 **US Patent 12,034,770**, J. Kocsis, M.P.S. Fernando, Y. Wu, "3S-Chain: Smart, Secure, and Software-Defined Networking (SDN)-Powered Blockchain-Powered Networking and Monitoring System"

2021 **US Patent 11,063,759**, J. Kocsis, Y. Wu, G.J.M.I. Liyangahawatte, "Blockchain-Empowered Crowdsourced Computing System"

Selected Research Projects

Multi-Human-Multi-Remote-Robot Collaboration (NSF Funded)

Purdue University

Principal Researcher — *2022 - 2024*

- Designed GPT-3.5 agentic coding workflow for autonomous robot control code generation, integrating LLM prompting with RL baselines (PPO, DDPG, HER).
- Built distributed control system with visualization and synchronization for remote construction machines over network infrastructure.
- Implemented low-latency control protocols for real-time robot coordination across distributed network topology.
- Developed communication middleware for multi-party human-robot collaboration with fault tolerance and security mechanisms.

Decentralized ML for NASA Space Exploration (NASA Funded)

Purdue University

Graduate Researcher — *2020 - 2023*

- Designed resilient decentralized machine learning system for spacecraft in delay-tolerant networks using blockchain-based coordination.
- Implemented bandwidth allocation algorithms for heterogeneous network flows using ADMM optimization in resource-constrained distributed environment.
- Built federated meta-learning framework for edge computing with limited distributed compute across multiple autonomous nodes.
- Developed resource allocation strategies optimizing compute distribution and handling network partitions in space mission scenarios.
- Presented research at NASA Space Technology Day conferences (2019, 2020).

HPC Storage I/O Optimization (NSF Funded)

Iowa State University

Research Assistant — *2017*

- Built deep learning (LSTM/RNN) models predicting I/O throughput and IOPS for Lustre distributed filesystem across 4-32 node HPC clusters.
- Optimized parallel file system performance on clusters with 500GB-4TB storage per node, enabling better workload scheduling.
- Integrated predictions with SLURM workload manager for intelligent job scheduling based on predicted I/O patterns.
- Achieved significant performance improvements in distributed storage optimization (published: arXiv:2301.06622).

Federated Learning on Embedded Systems (DoE Funded)

Purdue University

Graduate Researcher — *2020 - 2022*

- Implemented distributed machine learning on 4-16 embedded devices (Raspberry Pi, Nvidia Jetson Nano/TX2) using Torch-TensorRT.
- Solved synchronization challenges and fault tolerance in distributed training while maintaining computation precision with Integer-Vector Homomorphic Encryption.
- Optimized network communication patterns for bandwidth-constrained embedded device clusters.
- Published framework achieving efficient federated learning on resource-constrained hardware (IEEE IoT Journal 2020).

Attack-Resilient Smart Grid Communications (DoE/NREL Funded)

University of Akron

Research Assistant — *2015 - 2019*

- Designed hardware-in-the-loop (HIL) testing infrastructure for large-scale renewable energy systems with distributed monitoring.
- Implemented attack detection and resilient communication middleware for smart grid systems using machine learning and software-defined networking.
- Developed blockchain-powered decentralized computing system for crowdsourced IoT applications (resulted in 2 US patents).
- Collaborated with National Renewable Energy Laboratory (NREL) on smart grid security research.
- Published multiple papers in IEEE conferences and transactions on cybersecurity and distributed systems.

IoT Greenhouse Monitoring System

Purdue University

Research Assistant — *2020 - 2022*

- Implemented automated networking and monitoring system for distributed IoT sensor networks in greenhouse environments.
- Developed data collection pipeline with real-time analytics for environmental monitoring.
- Built fault-tolerant distributed sensor network with automatic reconfiguration capabilities.

Technical Skills

NLP & LLMs	Multilingual NLP, LLM inference optimization (sglang, vLLM), text classification, encoder/decoder models, GPT-4o, prompt engineering, RAG systems, response re-ranking, clinical text summarization, knowledge g
Machine Learning & Deep Learning	PyTorch, TensorFlow, Torch-TensorRT, Hugging Face Transformers, LSTM/RNN, federated learning, meta-l learning (PPO, DDPG, HER), adversarial ML, continuous batching, distributed training
Distributed Systems & HPC	Distributed machine learning, Lustre filesystem, SLURM workload manager, 32-node HPC clusters, resour optimization, fault tolerance, synchronization, edge/fog computing, spacecraft networks
AI Infrastructure & Cloud	AWS (EC2, Lambda, S3), multi-GPU setups, inference acceleration, latency optimization, production ML sy deployment, Docker, Kubernetes, microservices
Blockchain & Decentralized Computing	Blockchain-based coordination, decentralized ML systems, crowdsourced computing, smart contracts, c
Cybersecurity & Networking	Attack detection, intrusion detection systems, software-defined networking (SDN), network security, priva computing, homomorphic encryption, secure multi-party computation
Embedded Systems & IoT	Raspberry Pi, Nvidia Jetson Nano/TX2, embedded ML, IoT sensor networks, Modbus, Profibus-PA/DP, har testing, industrial automation
Data Science & Analytics	DBSCAN clustering, LangChain, vector databases (Weaviate), data pipelines, ETL workflows, evaluation m
Programming & Development	Python, C/C++, CUDA, Java, JavaScript, HTML/HTML5, Git, RESTful APIs, SQL, NoSQL, parallel programmi
Research & Scientific Computing	Experiment design, statistical analysis, algorithm development, optimization, numerical methods, simula peer review

Honors & Awards

2020 **Purdue & Midwest NSF I-Corps**, For Project "Crowdsourced AI"
2019 **Dean's Travel Grant for NASA's Space Technology Day**, Purdue University, College of Technology
2019 **CIT Graduate Student Travel Grant**, For NASA's Space Technology Day, Purdue University
2019 **UAkron NSF I-Corps**, For Project "3S-Chain"
2018 **UAkron NSF I-Corps**, For Project "Crowdsourced AI"
2016 **National Renewable Energy Laboratory Student Travel Grant**, For Smart Grid Research Collaboration
2008, 2009 **Annual Third-grade People Scholarship**, Harbin Institute of Technology
2008 **Annual Triple-A Outstanding Student**, Harbin Institute of Technology
2008 **Annual Excellent League Member**, Harbin Institute of Technology

Teaching Experience

Purdue University, College of Technology

West Lafayette, IN

Teaching Assistant

Aug. 2019 - Dec. 2019

- Course: Deep Learning for Cyber-Physical Systems
- Assisted in teaching deep learning concepts and applications to cyber-physical systems
- Held office hours, graded assignments, and provided technical support to students
- Developed supplementary materials and hands-on exercises for distributed systems topics

University of Akron, Department of Electrical & Computer Engineering

Akron, OH

Teaching Assistant

Aug. 2015 - Aug. 2017; Jan. 2018 - Jul. 2019

- Assisted with multiple graduate and undergraduate courses in electrical engineering and computer science
- Conducted lab sessions, graded assignments and exams, held office hours
- Mentored students on course projects related to embedded systems, networking, and machine learning
- Developed teaching materials and programming assignments

Tsinghua IT Training School

Xiangyang, Hubei, China

Instructor

Jan. 2015 - Jun. 2015

- Taught C programming fundamentals and web development (HTML/HTML5)
- Developed comprehensive course materials including lectures, labs, and programming exercises
- Provided one-on-one mentoring and career guidance to vocational training students
- Assessed student progress and adapted teaching methods to ensure learning outcomes