Ganyu Wang

Toronto, ON, Canada | wangganyu0@gmail.com | (+1) 2265045633 | linkedin.com/in/ganyu-wang

Summary

Machine Learning Researcher with a PhD in Computer Science and 4+ years of experience designing, implementing, and deploying large-scale machine learning systems in production. Specialized in distributed learning, optimization, and data-efficient training methods, with hands-on experience building backend and data infrastructure for cloud-based ML platforms supporting LLM and GenAI workloads. Published in top-tier machine learning conferences, including NeurIPS, ICLR, and ICML. Proven ability to translate research ideas into reliable, scalable systems, maintain complex codebases, and collaborate cross-functionally to deliver production-ready ML solutions. Proficient in Python, PyTorch, and Spark-based data pipelines.

Professional Experiences

Senior Researcher - AI Big Data System **Oct. 2025 - present**

Huawei Technologies Canada Co., Ltd.

- Designed and optimized LLM-powered semantic operators integrated into **Spark-based big data engines**, enabling semantic analytics over large-scale unstructured data.
- Owned performance optimization of LLM inference pipelines, improving **GPU utilization**, prefix-cache hit rate, and end-to-end throughput through prefix-aware inference, batching strategies, and system-level profiling.
- Built and evaluated distributed clustering and semantic aggregation pipelines, combining classical algorithms (e.g., K-means) with LLM-based reasoning under **strict latency and cost constraints**.
- Applied **low-level performance analysis** (async-profiler, hardware counters, JVM/native profiling) to optimize JNI, Arrow-based data exchange, and native execution components.
- Collaborated with cross-functional teams on research-driven system design, translating experimental results into practical optimizations for large-scale AI data platforms.

Machine Learning Researcher **Sept. 2021 - July 2025**

Western University

- Developed **scalable Distributed ML, Federated Learning framework for LLM training**, collaborative learning across institutions without sharing sensitive raw data.
- Applied black-box prompt tuning to cloud-based large language models (e.g., GPT-3.5 Turbo), reducing inference cost and enhancing adaptability in **GenAI applications**.
- Published peer-reviewed papers in **top-tier ML conferences (ICML[1], NeurIPS[2], ICLR[3], MLJ[4], KDD[5])** as **first author and project leader**, advancing the field of distributed ML systems and LLMs.
- **Integrated cutting-edge machine learning methods** into real-world ML systems, including Online Learning, Differential Privacy, and Zeroth-Order Optimization, using PyTorch, TensorFlow, and OpenAI APIs.
- **Deployed models** with AWS, Kubernetes, and serverless cloud infrastructure to ensure scalability and efficiency in production environments.

Lecturer – Data Mining **Jan. 2022 – May 2022**

Wilfrid Laurier University

- Designed and taught a hands-on undergraduate course covering real-world applications of **data mining**, including environmental data, health analytics, and social data mining.

Open-Source Projects

Efficient Distributed Prompt Learning for GenAI/LLM **Dec. 2023 - May 2025**

- Published as **the first** author in **top-tier ML conference ICML-2025** [1].
- Developed FedOne, a novel federated learning framework for **black-box prompt tuning for cloud-based LLMs** (e.g., GPT-3.5 Turbo), significantly reduced API query costs when training a discrete prompt.
- Conducted **comprehensive experiments on mainstream LLMs and standard benchmarks** to demonstrate the effectiveness of the proposed framework; released code as an open-source project.
- **Led research team**: planning and execution, overseeing architecture and experiment design, milestone tracking, and progress monitoring. **Managed version and branching strategies** on GitHub to ensure efficient collaboration.
- Conducted the *first theoretical analysis* of query efficiency in Federated black-box prompt learning, identifying the relationship between client activation strategies and cloud-based LLM service query costs.
- Demonstrated **significant cost savings** and improved generative performance in resource-constrained environments.

Privacy-Preserving and Communication-Efficient Vertical Federated Learning **Apr. 2022 - Jan. 2024**

- Published as the **first author** in the **top-tier ML conference (NeruIPS-2023)**[2] and **journal (MLJ)** [4].
- Designed a **large-scale distributed ML system**, enabling cross-organization collaboration (e.g., cities, companies) that significantly improves efficiency while preserving privacy, addressing critical challenges in distributed ML systems.
- Introduced *theoretical advancements* with novel analyses of optimization techniques and innovative implicit differential privacy guarantees, establishing new benchmarks in the field.
- Practically achieved a *substantial reduction in communication costs* through strategic algorithmic optimizations, paving the way for scalable AI solutions in resource-constrained large-scale distributed ML environments.

Efficient Online Learning Paradigm in Vertical Federated Learning **Jan. 2023 - Oct. 2024**

- Published as the **first author** in **top-tier ML conference, ICLR-2025** (top **5.2%** review score) [3].
- Proposed a novel distributed asynchronous online learning framework to address **streaming data and irregular updates**, which is often encountered in climate sensors, urban monitoring, or citizen science platforms.
- Improved robustness of distributed ML systems to **real-time data streaming, dynamic data arrival** in collaborative environments.

AI/ML Expertise and Technical Skills

ML Tools: PyTorch, TensorFlow, Scikit-Learn, JAX, HuggingFace, LangChain, OpenAI API.
ML expertise: Distributed system application, Large Language Model (LLM), Federated learning, Parallel computation, Optimization, Differential privacy.
Programming Languages: Python, C/C++, Java, JavaScript, Scala, Rust, SQL
Developments: Apache Spark, AWS, Kubernetes, Docker, FastAPI, DynamoDB, MongoDB, Sealos Cloud, Git.

Education

Ph.D. in Computer Science **Sept. 2021 - July 2025**
Western University

M.Sc in Computer Science (Thesis-based) **Sept. 2019 - July 2021**
Ontario Tech University

B.Sc in Computer Science (with Honor Bachelor's Degree) **Sept. 2015 - July 2019**
University of Electronic Science and Technology of China
*Yingcai Honors College (**Top 5%** of undergraduates)* *Overall GPA: **3.84/4.00** (87.02/100)*

Selected Peer-Reviewed Publications

[1] **Wang, Ganyu**, Jinjie Fang, Maxwell Juncheng Yin, Xi Chen, Boyu Wang, Bin Gu, and Charles Ling. Fedone: Query-efficient federated learning for black-box discrete prompt learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[2] **Wang, Ganyu**, Bin Gu, Qingsong Zhang, Xiang Li, Boyu Wang, and Charles X Ling. A unified solution for privacy and communication efficiency in vertical federated learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

[3] **Wang, Ganyu**, Boyu Wang, Bin Gu, and Charles X. Ling. Event-driven online vertical federated learning. In *International Conference on Learning Representations (ICLR)*, 2025.

[4] **Wang, Ganyu**, Qingsong Zhang, Xiang Li, Boyu Wang, Bin Gu, and Charles X Ling. Secure and fast asynchronous vertical federated learning via cascaded hybrid optimization. *Machine Learning*, 113(9):6413–6451, 2024.

[5] Ke Zhang, **Wang, Ganyu**, Han Li, Yulong Wang, Hong Chen, and Bin Gu. Asynchronous vertical federated learning for kernelized auc maximization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4244–4255, 2024.