Zirui Wen

wzrqczj@gmail.com | +1 201-932-5000 | LinkedIn | GitHub | Personal Page

Education

Stevens Institute of Technology, MS in Applied Artificial Intelligence, GPA 3.94/4.0 — Sep 2024 – May 2026
University of Birmingham, BS in Applied Mathematics with Information Computing Science — Sep 2020 – June 2024
Jinan University, BS in Information and Computing Science — Sep 2020 – June 2024

Experience

Research Assistant — Brain Imaging and Graph Learning Lab, Stevens Institute of Technology — Nov 2024 – Present

- **Architected a GraphRAG pipeline** integrating **Neo4j** (Cypher) with **FAISS** dense retrieval via **LangChain**. On a neurology FAQ benchmark, EM **+32%**; with RAGAS faithfulness/answer-relevance (mean**±95%** CI)
- **Designed and implemented a Python data analysis pipeline** for large-scale hypothesis testing on 10k+ de-identified patient records: **vectorized ETL** and **hypothesis testing** by **Shapiro–Wilk, Chi-square**.
- **Fine-tuned LLaMA-3 & Mistral-7B** with **LoRA** in the HF stack, exported 8-bit quantized artifacts for inference. Achieved **+20%** F1 on de-identified brain-science Q&A.
- **Built a custom RL environment** and trained **PPO** for multi-step seizure-trajectory prediction. Delivered **+17%** higher cumulative reward vs. clinician-derived baselines.

Data Team Assistant Intern — Siemens — Nov 2023 – March 2024

- **Built modular Python pipelines (Pandas/NumPy)** to automate multi-source data ingestion, cleansing, and transformation, reducing manual processing time by **60%** while boosting data-quality conformance to **99.8%**.
- **Analyzed 2+ years of transactional data** using **SQL** and **Excel** to compute KPIs and evaluate **A/B tests**, providing insights that informed product strategy and improved engagement by **12%**.

Publications

Shihao Yang, **Zirui Wen**, Wenxin Zhan, et al. Knowledge Graph Representation of the Mappings between Seizure Semiology and Epileptogenic Zones, Under review at Epilepsia, 2025.
Zirui Wen, Junjie Zhang, and Yuhao Zhang. "COVID-19 Infection Prediction using Physical Signs." International Conference on Cloud Computing, Performance Computing, and Deep Learning (CCPCDL 2022). Vol. 12287. SPIE, 2022

Projects

Recommendation service — Aug 2025

- **Built a Java Spring Boot (WebFlux) microservice** for real-time recommendation; integrated **Apache Kafka** for streaming and **Redis** for low-latency features; implemented model loading with **XGBoost4J** and safe hot reload.
- **Containerized and deployed with Docker and Kubernetes** using **Helm**; enabled **HPA** autoscaling, health probes, structured JSON logging; exported metrics via **Micrometer** to **Prometheus** and visualized in **Grafana**.
- **Implemented GitHub Actions CI/CD** with **Maven**; added unit and integration tests using **JUnit 5** and **Testcontainers**; built images with **Jib** and shipped versioned releases to a container registry.

BPlusTree Database Project — May 2025

- **Designed and implemented a mini-RDBMS in C++**, featuring an order-3 **B+Tree** storage engine, **buffer pool**, **secondary indexes**, and **WAL**; executed SQL-style point/range queries in **O(log n)** latency.
- **Developed an interactive SQL shell** that supports 15 SQL-style commands (SELECT, JOIN, LOAD); **built a Python ETL** to bulk-load Google Maps Saved Places CSVs, with reproducible **Makefile** builds and 90%+ unit-test coverage.

Kernel K-Means GPU Accelerator — May 2025

- **Refactored kernel-Kmeans to sparse linear algebra**, casting core steps as SpMM/SpMV, offloading to cuSPARSE/cuBLAS.
- **Tuned for memory throughput** and occupancy with **coalesced** global accesses, **shared-memory tiling**, and **register/SM balance** achieved **1000×** CPU and **2.6×** dense-GPU baselines on MNIST/CIFAR-10.

Automatic prompt optimization for medical prompts — Jan 2024

- **Built a text-gradient + momentum prompt optimizer** with Bayesian reverse validation; improved MedQA/PubMedQA accuracy by 20% vs **CoT** and shipped it as a **LangChain** service (**Flask API + Next.js**). Added a pragmatic RAG fallback using hybrid retrieval (**BM25** + **dense embeddings with RRF**) and a cross-encoder reranker to tighten answer relevance.
- **Implemented offline RAGAS metrics** (faithfulness, context precision/recall) with **LangSmith** datasets/traces, plus **OpenTelemetry** dashboards for latency/QPS/errors; exposed a low-latency gRPC scoring path for critical calls.

Technologies

Languages: Python, C++, Java, JavaScript, HTML/CSS, SQL
Deep Learning Frameworks: PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face, LangChain, vLLM, MLFlow, Optuna
Data Analysis: NumPy, Pandas, MySQL, PostgreSQL, MongoDB, Redis, Neo4j, FAISS, Feast, Milvus, Qdrant, Weaviate
Web & DevOps Tools: Flask, FastAPI, React, Next.js, Vue.js, gRPC, REST, WebSockets, Kafka, Docker, Kubernetes, Helm, Ray Serve, KServe, Triton, Git, Linux, CUDA, CI/CD, AWS, GCP, Terraform, GitHub Actions, Apache Airflow, Prometheus, Grafana, OpenTelemetry, Schema Registry, Azure, Google Cloud
Skills: Machine Learning, MLOps/LLMOps, Data Analysis , Data Structures & Algorithms, Microservices, Distributed Systems, Concurrency, System Design, Testing, Observability