(Jeremy) 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 – Present
University of Birmingham, BS in Applied Mathematics with Information Computing Science Sep 2020 – Jun 2024
Jinan University, BS in Information and Computing Science Sep 2020 – Jun 2024

Experience

Research Assistant —Intelligent System (IntelliSys) Lab, Stevens Institute of Technology Nov 2025 – Present
Forgetting Score Guided Continual Post Training for LLMs

- Implemented a **training time forgetting monitoring metric** that combines loss non-quadraticity, Hessian spectrum curvature scale, and parameter update misalignment to track **catastrophic forgetting** during continual post-training.

Research Assistant — Brain Imaging and Graph Learning Lab, Stevens Institute of Technology Nov 2024 – Present
Clinical Epilepsy QA (KG+Dense RAG, LoRA Fine-tuning)

- **Built a pipeline** to extract relationships between seizure symptoms and cortical regions from **10k+** de-identified clinical records and served them as a Neo4j **knowledge graph** for medical question answering grounding.
- Implemented **GraphRAG** using Neo4j Cypher queries and **FAISS dense retrieval**, then **QLoRA** fine-tuned **LLaMA-3** and **Mistral-7B**, achieving a **20 %** improvement in factual consistency on internal benchmarks compared to the baseline.

Seizure Trajectory Reinforcement Learning Modeling

- Modeled seizure progression as a discrete-state decision **MDP** process over semiology transitions and designed a reward that prefers early and stable predictions under partial observations.
- Trained with a **PPO-style** objective and achieved a **40 %** improvement in predictive return with improved cross-patient generalization compared to baseline MC methods.

Data Team Assistant Intern — Siemens Nov 2023 – Mar 2024

- Built reusable Excel templates for weekly data updates, including schema standardization, KPI definitions, and automated summary tables; wrote basic **SQL** queries to generate analysis-ready datasets.
- Implemented data quality validation and reconciliation across multiple tables and source exports using SQL checks and spot-audit workflows to ensure consistent reporting.

Publications

Zirui Wen, Shihao Yang, et al. Uncovering Epileptic Seizure Propagation Using Knowledge Graph-based Reinforcement Learning, Under review, 2025
Shihao Yang, **Zirui Wen**, Wenxin Zhan, et al. Knowledge Graph Representation of the Mappings between Seizure Semiology and Epileptogenic Zones, Accepted by Scientific Reports, 2025.
Zirui Wen, Wensheng Gan, et al. Automatic Prompt Optimization for Medicine, Under review, 2024
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

Smarter Doctor agent Oct 2025

- Built and deployed a real-time voice-based **medical agent** on **Google Cloud** using a FastAPI backend and Next.js frontend with WebSocket streaming and tool orchestration, integrating Twilio, **Elasticsearch**, **BigQuery**, and **Gemini** through **Vertex AI**, with CI CD via Docker and GitHub Actions.

Recommendation service Aug 2025

- Built a real-time **recommendation system** on AWS with **Kafka**, **Feast**, and **Redis** features, and a Dockerized FastAPI service serving **XGBoost**, productionized with automated retraining and deployment, **MLflow** and **Optuna** tuning, **Airflow** pipelines, and Kubernetes monitoring and autoscaling.

Kernel K-Means GPU Accelerator May 2025

- Rewrote kernel k-means as sparse linear algebra and accelerated it on GPUs using cuSPARSE and cuBLAS, optimizing memory access and occupancy with **coalesced global accesses**, **shared-memory tiling** to outperform CPU and dense GPU baselines on MNIST and CIFAR 10.

Automatic prompt optimization for medical prompts Jan 2024

- Built an automatic **medical prompt optimizer** using text gradient style updates with momentum and Bayesian validation, shipped as a **LangChain service** with a **hybrid retrieval** and reranking improved MedQA accuracy by **20%** vs CoT
- Added an evaluation and observability stack with offline RAGAS scoring, LangSmith tracing, OpenTelemetry dashboards, and a low-latency gRPC scoring path.

Technologies

Languages: Python, C++, Java, JavaScript, HTML/CSS, SQL, R
LLM & Training: PyTorch, LoRA/QLoRA, FSDP/DeepSpeed, RLHF, RAG/GraphRAG
Serving & LLMOps: Triton/TensorRT/ONNX, vLLM/TGI, FastAPI/gRPC, KServe, Prometheus/Grafana, A/B testing
Data & Platform: Spark/Ray, Airflow, MLflow, Docker/K8s, AWS/GCP, Terraform, SQL/NoSQL, Vector DB (FAISS/Milvus)