# 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 — Sept 2024 – Present
**University of Birmingham**, BS in Applied Mathematics with Information Computing Science — Sept 2020 – June 2024
**Jinan University**, BS in Information and Computing Science — Sept 2020 – June 2024

## Experience

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

- **Designed an end-to-end Retrieval-Augmented Generation (RAG) platform** that fuses a **Neo4j knowledge graph** with **FAISS vector search** and **LLMs**, lifting answer exact-match accuracy by **32%** on a neurology FAQ dataset.
- **Performed large-scale statistical analyses (Shapiro–Wilk, chi-square, Pearson/Spearman)** across **10,000+** patient records, isolating features that improved seizure-onset prediction AUC from **0.71** to **0.83** in cross-center validation.
- **Fine-tuned LLaMA-3 and Mistral-7B with LoRA and 8-bit quantization** on de-identified brain science Q&A data, achieving a **20%** F1 improvement
- **Prototyped & benchmarked reinforcement learning agents (DQN, PPO)** to predict seizure trajectories from time-series vitals, achieving **17%** higher cumulative reward compared to clinician-derived baselines.

**Data Team Assistant Intern**, Siemens – Guangzhou, China — 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

**Automatic prompt optimization for medical prompts** — Apr 2024

- **Designed an algorithm to automatically optimize medical prompts** using text-based gradient descent and momentum with Bayesian reverse validation, boosting LLMs on MedQA & PubMedQA by **20%** relative to **CoT baseline**.
- **Packaged the optimizer as a LangChain** tool-chain that layers advanced **prompt-engineering** patterns (dynamic system/instruction templates, function-calling, **RAG** fallback) and ships as a production-grade AI micro-service: **Flask REST API**, and a **Next.js** dashboard with live streaming and versioned prompt history.

**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

- **Built an open-source GPU kernel K-means accelerator in CUDA**, refactoring distance ops into sparse SpMM/SpMV on **cuBLAS** and **cuSPARSE**, achieved **1,000×** CPU speed-up and **2.6×** dense-CUDA baseline on MNIST/CIFAR-10.

**Liar's Bar: Bayesian Reinforcement Learning** — Dec 2024

- **Modeled the game as an imperfect-information Bayesian game**, solved for subgame-perfect Nash equilibria, and used those policies to warm-start a **DQN** agent—boosting win-rate **12%** over baseline bots in **10k** sims. The engine is exposed as a **Flask+PostgreSQL** microservice and auto-scaled on **Kubernetes** for multiplayer sessions.

**An Apex Legends AI Aimbot based on YOLO** — Dec 2023

- **Accelerated YOLOv8 with TensorRT+CUDA Graphs** to **60 FPS** & **<10 ms** on an RTX 3060 (mAP 0.89 on 1k frames) and built a lightweight desktop overlay that shows targets and triggers precise in-game aim.

**Asset Allocation Optimization Based on PSO and fixed point method** — Jun 2022

- **Reformulated the Markowitz mean-variance problem into an L1-regularised SSMP model** and solved it with two independent optimisers—**Particle Swarm Optimization** (global search) and a **proximal fixed-point gradient method** (local refinement)—boosting five-year back-test returns **2.6×** versus the vanilla MV solver.

## Technologies

**Languages:** Python, C++, Java, JavaScript, HTML/CSS, SQL
**Deep Learning Frameworks:** PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face, LangChain
**Data Analysis:** NumPy, Pandas, MySQL, PostgreSQL, MongoDB, Redis, Neo4j, FAISS
**Web & DevOps Tools:** Flask, FastAPI, React, Next.js, Vue.js, Kafka, Docker, Kubernetes, Git, Linux, CUDA, CI/CD, AWS/GCP
**Skills:** Machine Learning, Data structure, Algorithm Design, Recommendation system, Object-Oriented programming