Youwei Zhen

69 Brown St, Providence, Rhode Island
US Citizen | 917-882-0066 | youweizhen.com | youwei_zhen@brown.edu | linkedin.com/in/youwei-zhen | github.com/AntoDono

Education

Brown University — Providence, RI
Bachelor of Science in Applied Mathematics and Computer Science, GPA: 4.0 — *Aug. 2024 – May 2027*

Experience

Machine Learning Engineer — March 2025 – Present
Refine.dev - YC 23 — *Remote*
- Analyzed **attention score distributions** to identify performance degradation in **long-context prompts**
- Built **multi-agent system** with specialized sub-agents using different models for PM and development tasks
- Developed **code indexing system** enabling **sub-100ms retrieval** across thousands of files in **150+ projects**
- Achieved **2000+ token/second generation speed** through optimized **API orchestration** and model selection
- Built **Next.js** frontend with **Langfuse** integration and real-time **WebSocket** streaming backend

Lead Software Developer — December 2024 – Present
Index Tax & Financial — *Remote*
- Building **full-stack platform** for tax financial services firm with **Django REST**, **Nuxt.js**, **Electron.js**, and **Expo**
- Developed **RAG system** with custom **vector database** to process **500GB+** of client tax documents
- Reduced CPA workflow time by **50%** through automated data aggregation and cross-referencing
- Built **CI/CD pipeline** with **Docker** for automated deployment and updates across client devices

Research Assistant — June 2025 – August 2025
Brown Database Group — *Providence, RI*
- Developed **chunking system** to segment clinical records, avoiding **attention degradation** in long documents
- Built **parallel redaction pipeline** reducing processing time from **1800 seconds to 30 seconds** per document
- Evaluated multiple models (**DeepSeek-R3, Gemma3, Qwen**) for clinical text de-identification accuracy

Lead Software Engineer — February 2023 – August 2023
SammyGPT - Staten Island Technical High School — *Staten Island, NY*
- Led development of **full-stack AI assistant** for **1,500+ students and staff** with **Nuxt.js** and **Flask**
- Built **web scraping pipeline** and **vector database** indexing 1,000+ pages for **sub-200ms retrieval**
- Implemented **4-bit quantization** and **BERT** with real-time streaming responses via custom **WebSocket** protocol
- Secured **$10,000 sponsorship** for dedicated **Lenovo ML server** infrastructure
- Achieved national recognition with feature **publication in NASSP Leadership magazine** (December 2023)

Projects & Awards

4x Winner – HackPrinceton | *1st Capital One, 1st Knot API, Gemini, YC Runner-Up* — November 2025
- Built EdgeCart, an **AI-powered food waste prevention system** reducing grocery emissions by **19.5%**
- Developed **custom ResNet model** to detect fruit freshness with **Google Gemini segmentation mask**
- Implemented **intelligent discount matching** using **Knot API** and **xAI** to target customers, cutting waste by **180M tons annually**

1st Place – HackMIT (Cerebras Systems Track) — September 2025
- Built tempoRoll, a **music therapy system** using **real-time EEG monitoring** for mental health conditions
- Developed hybrid **CNN/ResNet** brainwave classifier with **Cerebras inference** achieving **85% accuracy**
- Integrated wireless **BLE** EEG communication and generated personalized music to guide brainwaves to baseline

2nd Place – Cornell BigRedHacks — September 2025
- Created Duelingo, real-time multiplayer language game with **AI-powered non-deterministic** vocabulary discovery
- Built **on-the-fly evaluation system** assessing linguistic and cultural context with **sub-200ms response times**
- Implemented **caching system** reducing token usage by **90%** using **Groq LLM** and **Google Cloud Speech API**

USACO Gold Level (Platinum Division) — February 2023

Technical Skills

Languages: Python, Java, JavaScript, C#, C++, R, HTML, CSS, TypeScript, SQL
Frameworks: Vue.js, Nuxt.js, React.js, Next.js, Node.js, Flask, Express, Django, PyTorch, TensorFlow
Developer Tools: Git, Docker, Linux, Nginx, MongoDB, PostgreSQL, WebSocket, REST APIs, CI/CD, Langfuse
ML/AI: Deep Learning, NLP, Computer Vision, RAG, Vector Embeddings, Transformers, BERT, CNNs, ResNet
Spoken Languages: English, Chinese (Mandarin and Cantonese)