Youwei Zhen

69 Brown St, Providence, Rhode Island
917-882-0066 | 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. 2023 – May 2027*

Experience

Machine Learning Engineer — March 2025 – Present
Refine.dev - YC 23 — *Remote*

- Researched **attention dilution** and developed **multi-agent architecture** to mitigate generation degradation
- Created scalable **RAG database** with real-time **code indexing and semantic search** for LLM context learning
- Implemented **LoRA fine-tuning** and optimized **LLM performance** for enterprise software automation
- Developed **multi-agent system** with specialized research and implementation agents

Project Manager — December 2024 – Present
Index Tax & Financial — *Remote*

- Built full-stack platform with **Django backend**, **Nuxt.js frontend**, **Electron.js** desktop and **Expo** mobile apps
- Developed high-performance AI agent with sub-millisecond **RAG capabilities** for complex client queries
- Built **autonomous systems** for client communication automation including voice AI and email processing
- Created **RPA solution** for **automated** tax document management and filing workflows
- Managed **distributed systems** and hardware **deployment across Qingdao, Shanghai, and Los Angeles**

Research Assistant — June 2025 – August 2025
Brown Database Group — *Providence, RI*

- Researched LLM **attention degradation and dilution**, developing **multi-agent systems** to redact clinical records
- Developed models for **clinical text de-identification**, filtering PHI from unstructured records
- Engineered **record segmentation**, enabling **high accuracy** and **parallel redaction**

Project Lead — February 2023 – August 2023
SammyGPT (sammy.siths.tech) — *Staten Island, NY*

- Created an AI Assistant **using 4-bit quantization, LLMs, Vector Database, BERT, and NLP**
- Developed features including AI detection tool, multilingual chatbot, and school information provider
- Secured **$10,000 sponsorship** funding for dedicated **Lenovo Machine Learning** server infrastructure
- Achieved national recognition with feature **publication in NASSP Leadership magazine** (December 2023)

Projects & Awards

1st Place – Emergent AI Conference | *$5,000 Prize* — May 2025

- Developed OSCE AI using **multi-agent architecture** with **custom fine-tuned LLMs** for medical exam preparation, featuring realistic patient simulation and personalized feedback delivery using **RAG** and **vector embeddings**

1st Place – HackMIT (Cerebras Systems Track) — September 2025

- Built tempoRoll, a real-time EEG-based music therapy system with **custom brainwave classification model** and **Cerebras API** (**20× faster than GPUs**), achieving **85% accuracy** in diagnosing schizophrenia, bipolar disorder, and ADHD

2nd Place – Cornell BigRedHacks — September 2025

- Created Duelingo, a **non-deterministic language learning game** with AI evaluation of infinite word combinations in **real-time (under 200ms)** and pronunciation verification using **Groq LLM API** and **Google Cloud Speech API**

USACO Gold Level (Platinum Division) — February 2023

Technical Skills

Languages: Python, Java, JavaScript, C#, C++, R, HTML, CSS, TypeScript
Frameworks: Vue.js, Nuxt.js, React.js, Next.js, Node.js, Flask, Express, PyTorch, TensorFlow
Developer Tools: Git, GitHub, Docker, Linux/Ubuntu, Nginx, MongoDB, Socket.io, Canvas API
Libraries & Technologies: Natural Language Processing, RAG, Vector Embeddings, Transformers, BERT
Spoken Languages: English, Chinese (Mandarin and Cantonese)