Youwei Zhen

69 Brown St, Providence, Rhode Island

☎ 917-882-0066 ✉ youwei_zhen@brown.edu linkedin.com/in/youwei-zhen-a8b662213 github.com/AntoDono

Education

Brown University **Expected Graduation: 2028**

Bachelor of Science in Applied Mathematics and Computer Science, ***GPA:*** *4.0* *Providence, Rhode Island*

- **Relevant Coursework:** Statistical Inference I, Linear Algebra, Foundations of AI, Computer Systems, Deeplearning, Ordinary Differential Equation

Technical Skills

Computer Skills: Python, Java, Javascript, Node.js, C#, C++, R, HTML, CSS, Vue.js, Nuxt.js, React.js, Next.js, MongoDB, Nginx, Typescript, Linux/Ubuntu, Git, Github, Docker, Natural Language Processing, Pytorch, Tensorflow, Flask, Express, Socket.io, Canvas API.

Languages: Chinese (Mandarin and Cantonese), English

Awards & Certifications

United States of America Computing Olympiad - Gold Level (Platinum Division) *February 2023*

1st Place Winner - Emergent AI Conference Competition ($5,000 Prize) *May 2025*

- Developed OSCE AI, an AI-powered virtual patient interviewer for medical OSCE exam preparation using **multi-agent architecture** with **custom fine-tuned LLMs** for realistic patient simulation, real-time clinical assessment, and personalized feedback delivery using **RAG** and **vector embeddings**.

1st Place Winner - HackMIT (Cerebras Systems Track) *September 2025*

- Built tempoRoll, a real-time EEG-based music therapy system combining a **custom-trained-from-scratch brainwave classification model** with **Cerebras API** (**20× faster than GPUs**) to detect mental health conditions and generate personalized adaptive music therapy, achieving **85% accuracy** in real-time diagnosis for schizophrenia, bipolar disorder, and ADHD using **NeuroSky EEG hardware** and **Google Cloud Speech API**.

2nd Place Overall Winner - Cornell BigRedHacks *September 2025*

- Created Duelingo, the world's first non-deterministic language learning game where AI evaluates infinite creative word combinations in real-time (**under 200ms**) with mandatory pronunciation verification across multiple languages using **Groq LLM API** and **Google Cloud Speech API**.

Industry Experience

Refine.dev **March 2025 - Present**

Machine Learning Engineer *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.
- Engineered advanced tool-use frameworks and **vector embeddings** for sophisticated code understanding.
- Implemented **LoRA** fine-tuning and optimized **LLM performance** for enterprise software automation.
- Developed **multi-agent system** with specialized research and implementation agents.

Index Tax & Financial **December 2024 - Present**

Project Manager *Remote*

- Developed high-performance AI agent with sub-millisecond **RAG** capabilities for complex client queries.
- Built an **autonomous systems** for client communication automation including voice AI and email processing.
- Created **RPA** solution for **automated** tax document management and filing workflows.
- Directed physical **computing architecture** design with custom **NVIDIA** GPU and **AMD** Threadripper CPU configurations.
- Managed distributed systems and hardware **deployment across Qingdao, Shanghai, and Los Angeles.**

Brown Database group **June 2025 - August 2025**

Research Assistant *Providence, Rhode Island*

- Researching 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.**

SammyGPT | sammy.siths.tech **February 2023 - August 2023**

Project Lead *Staten Island, New York*

- Created an AI Assistant **using 4-bit quantization, LLMs, Vector Database, Transformers, BERT, and NLP.**
- Features: AI detection tool, multilingual chatbot, 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).