Youwei Zhen

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

Education

Brown University May 2027
Bachelor of Science in Applied Mathematics and Computer Science, ***GPA: 4.0***

Experience

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

- Engineered production observability pipeline to detect **attention sinks** degradation in **long-context LLMs**
- Architected a distributed **code indexing system** using Pinecone and Redis caching, achieving **sub-100ms latency retrieval** across **150+ repositories** and reducing **LLM token consumption by 20%**.
- Optimized inference throughput to **2000+ token/second** by implementing **asynchronous API orchestration**, and **load-balancing** across multiple model providers.
- Designed and implemented **multi-agent system** with **cost-aware routing layer** and real-time prompt complexity classification, dynamically dispatching requests based on **semantic difficulty scoring** and **latency-cost tradeoff**.

Lead Software Engineer December 2024 – Present
Index Tax & Financial *Remote*

- Engineered **full-stack platform** for tax financial services firm with **Django REST**, **Nuxt.js**, **Electron.js**, and **Expo**, reducing CPA workflow time by **50%** through automated data aggregation and cross-referencing.
- Architected fully autonomous **RAG pipeline** with **Redis** caching, **Celery** worker distribution, and custom **vector database** to process **500GB+** of client tax documents concurrently.
- Built **CI/CD pipeline** with **Docker** for automated deployment, remote updates, and silent software distribution.
- Designed **concurrency control system** with **dynamic resource allocation** across **2 GPUs** and **24 cores**, preventing database deadlocks during real-time file ingestion. Secured infrastructure via **reverse tunneling** with restricted port exposure.

Founding 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**, reducing staff workload by **23%** through **multilingual support** (Chinese, Russian, English).
- Built **web scraping pipeline** with **Celery** cron jobs and **ChromaDB** caching, indexing **1,000+ pages** for **sub-200ms retrieval** with real-time school data updates.
- Implemented **modular inference architecture** with **4-bit quantization** and **BERT**, enabling plug-and-play model swapping with real-time streaming via custom **WebSocket** protocol.
- Secured **$10,000 sponsorship** for dedicated **Lenovo ML server**, and featured in **NASSP Leadership magazine** (December 2023).

Projects & Awards

EdgeCart – AI Food Waste Prevention | *Python, OpenCV, ResNet, Gemini API* November 2025

- **4x Winner @ HackPrinceton (Capital One, Knot API, Gemini, YC Runner-Up)**: Developed an edge-computing vision system to detect produce freshness using custom ResNet-50 models
- Implemented high-precision **blemish segmentation** using Google Gemini, improving detection accuracy by **19.5%**
- Implemented **intelligent discount matching** using **Knot API** and **xAI** to target customers, cutting waste by **180M tons annually**

TempoRoll – Real-Time EEG Music Therapy | *Python, PyTorch, Cerebras Systems* September 2025

- **1st Place @ HackMIT (Cerebras Track)**: Engineered a real-time brainwave classification system using a hybrid CNN/ResNet architecture
- Optimized Wafer-Scale Engines inference, achieving **85% accuracy** on EEG streams with **sub 200ms latency**
- Integrated wireless **BLE** EEG communication and generated personalized music to guide brainwaves to baseline

USACO Gold Level (Platinum Division) February 2023

Technical Skills

Languages: Python (Advanced), C++, Java, TypeScript, JavaScript, SQL, R, C#
AI & Machine Learning: PyTorch, TensorFlow, vLLM, RAG, Vector Search, Transformers, BERT, CNNs, ResNet
Backend & Infrastructure: FastAPI, Django, Flask, Docker, Kubernetes, Nginx, PostgreSQL, Redis, MongoDB, CI/CD
Web Technologies: Next.js, React, Vue.js, Nuxt.js, Node.js, Nginx, WebSockets, REST APIs
Spoken Languages: English, Chinese (Mandarin and Cantonese)