

Zhilang Gui

857-343-4019 | lang.gui.bu@gmail.com | San Francisco, CA | buildwithlang.com

Education

Boston University

B.S. Computer Engineering Concentration: Technology Innovation

University of Sydney (Study Abroad)

Boston, MA

May 2025

Spring 2023

Experiences

EasyBee AI | AI Engineer

San Francisco, CA | July 2025 - Present

- Engineered a stateful LangGraph supervisor agent with 3 specialized sub-agents (retrieval, validation, response generation), implementing agent pooling that reduced cold-start latency from 5s to 150ms - enabling real-time user interactions.
- Built RAG pipeline integrating Pinecone vector search (text-embedding-3-small) with OpenAI model, implementing Server-Sent Events streaming that reduced end-to-end response time from 5s to 2-3s and improved booking speed by 40%.
- Deployed production system on AWS EC2 with Docker, CI/CD pipelines, and AI-assisted code review (CodeRabbit).

Cadence Design Systems | Machine Learning Engineering Intern

Austin, TX | May 2024 - August 2024

- Developed and validated U-Net thermal prediction model achieving 5% efficiency improvement over baseline, processing multi-layer thermal maps for data center optimization.
- Built 3D visualization pipeline (Paraview) for CNN feature map interpretation, enabling engineers to diagnose thermal prediction errors and accelerate model iteration cycles.
- Built ETL pipeline (Python, Pandas, NumPy) for U-Net model processing 500GB of data center thermal layer data in overnight batch jobs, eliminating 8 hours/week of manual CSV analysis by senior engineers.

Terrier Motorsports | GLV Team Lead

Boston, MA | 2023-2024

- Led 5-engineer team to deliver 3 embedded subsystems (steering control, telemetry, driver alerts) using C++ and CAN bus protocols, reducing input latency from 120ms to 72ms. Completed integration 1 week ahead of the competition deadline using Agile sprints.

Human Computer Interaction Research Assistant

Sydney, Australia | Feb 2023 - July 2023

- Built Python models analyzing smartwatch sensor data (heart rate, step count, sleep patterns) from 200+ users, predicting user workout patterns with 75% accuracy.
- Identified UI friction points and redesigned onboarding flow, improving engagement by 15%.

Projects

Semi-Autonomous Tricycle for the Visually Impaired (CoE Societal Impact Award)

Sep 2024 - May 2025

- Built autonomous navigation system using Oak-D stereo cameras and ROS2 achieving 92% real-time obstacle detection accuracy.
- Integrated Google Maps API with haptic feedback vest and voice guidance, reducing route deviation by 60% during pilot tests with 5 visually impaired users.
- Won CoE Society Impact Award (top among 70 senior design teams) for technical innovation and real-world impact.

SaaS Prototype Suite (5 weeks 5 apps)

San Francisco, CA | July 2025 - Dec 2025

- Developed AI-powered news aggregator processing 20+ SF sources with automated weekly summaries via Vercel Cron, and photo-based restaurant recommendation app using OpenAI Vision API for real-time menu analysis. Both deployed on Vercel with PostgreSQL/Prisma backend.
- Shipped production Chrome extension with parallel AI processing pipeline executing 4 concurrent LLM calls in 3-5s, reducing average article reading time by 12 minutes.

Skills

- AI/ML: LangChain, LangGraph, PyTorch, RAG Pipelines, Prompt Engineering, Multi-Agent Systems, Vector DBs (Pinecone), OpenAI Model API, Computer Vision, Embedding, Agentic AI, Inference Optimization.
- Software & DevOps: Python (FastAPI/Flask), C++, SQL (PostgreSQL), REST APIs, Async Programming, Docker, GitHub Actions CI/CD, AWS (CloudWatch/EC2), Nginx
- Tools & Data: Git, ROS2, Pandas, NumPy, Three.js, Paraview