Zhilang Gui

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

Education

Boston University **Boston, MA**
B.S. Computer Engineering Concentration: Technology Innovation *May 2025*
University of Sydney (Study Abroad) *Spring 2023*

Experiences

EasyBee AI | LLM 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 production RAG pipeline integrating Pinecone vector search (text-embedding-3-small) with OpenAI GPT-4o, implementing Server-Sent Events streaming that reduced end-to-end response time from 5s to 2-3s and improved booking conversion by 40%.
- Deployed containerized AI system with Docker + GitHub Actions CI/CD, implementing comprehensive monitoring via AWS CloudWatch.

Cadence Design Systems | Machine Learning Engineering Intern Austin, TX | May 2024 - August 2024

- POC Development: Led the technical evaluation of a new thermal modeling feature, building a functional prototype that validated performance gains of 5% for enterprise use-cases.
- Technical Demonstrations: Developed a custom 3D visualization tool to simplify complex model outputs, enabling non-technical stakeholders to make faster hardware validation decisions.
- Automated thermal data processing pipeline using Python (Pandas, NumPy) that processes 500GB of simulation 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 churn with 75% accuracy.
- Identifying UI friction points that improved engagement by 15% after implementing recommended onboarding changes.

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.
- Built full-stack SaaS: Chrome Extension (MV3), 7 serverless API endpoints, Stripe subscription integration, and Next.js marketing site.

Skills

- AI/ML: LangChain, LangGraph, PyTorch, RAG Pipelines, Prompt Engineering, Multi-Agent Systems, Vector DBs (Pinecone), OpenAI API, Computer Vision (U-Net)
- 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