Chi (Bruce) Cheng

San Francisco, CA | bruceche@andrew.cmu.edu | 858-305-0278 | bruceche.com | linkedin.com/in/chi-cheng921

PROFILE

AI engineer and product-minded technical project owner with experience shipping production LLM systems, automating workflows, and driving projects from implementation to delivery. CMU M.S. Software Engineering candidate (Dec 2026), with research in AI reliability and model evaluation.

EDUCATION

Carnegie Mellon University, Silicon Valley | M.S. Software Engineering | GPA 3.6 | Dec 2026

University of California San Diego | B.S. Mathematics–Computer Science | GPA 3.6 | Mar 2025

EXPERIENCE

Carnegie Mellon University — Graduate Researcher, Applied Generative AI (AGAI) | Sep 2026 – Present

- Identified and resolved flaws in LLM hallucination-detection experiments, improving detection precision by 6 percentage points through systematic evaluation and debugging.
- Evaluated factual support across TriviaQA and SQuAD with NLI-based verification; analyzed confident model errors and answer variations driving false positives and false negatives.

Helport AI — AI Product Developer / Product Management | Sep 2024 – Jun 2025

- Built Python/FastAPI services for an AI call assistant across mortgage, healthcare, insurance, and government workflows; shipped Docker/CI/CD deployments and real-time guidance APIs.
- Migrated intent matching to Gemini 2.0/Vertex AI, reducing cost 6× ($0.002 to $0.00031 per match); shortened development cycles from 14 to 5 days.
- Used A/B testing to guide AI workflow improvements; reported 35% accuracy improvement, 35% less handling time, and 15% higher sales conversion.
- Initially interned/worked part-time (Aug 2024–Mar 2025) and transitioned to full-time (Apr–Jun 2025), per LinkedIn employment detail; resume role dates begin Sep 2024.

Convoloo — Software Engineer Intern | Jul 2024 – Sep 2024

- Built an AI healthcare chatbot with React, FastAPI, LangChain, and GCP; developed MySQL-backed RAG retrieval, reducing response time from 2.5s to 1.5s.
- Integrated LangGraph/LangServe, FastAPI, and NestJS to support backend workflows and deployment.

UC San Diego — Instructional Assistant, Calculus / Differential Equations | Sep 2023 – Apr 2024

- Supported university mathematics instruction and student learning in calculus and differential equations.

SELECTED PROJECTS

NightyNight — Independent AI Product | Apr 2025 – Present

- Independently built an AI storytelling product with React/TypeScript, FastAPI, LangGraph agent workflows, SSE streaming, and ElevenLabs text-to-speech.

Capitawise — Franklin Templeton Hackathon, 2nd Place ($7,000) | Mar – Jun 2024

- Led full-stack development of a GPT-4o banking assistant using React, Node.js, and Flask; improved resolution from 68% to 85% and reduced response time by 40%.

Parking Spot Locator — CMU / Bosch Research Collaboration | Aug – Dec 2025

- Built multimodal parking search with CLIP and vision-language models; improved query latency 3×, from 45s to 15s.

Research Assistant Agent — CMU Course Project | Jan – Mar 2026

- Developed a LangGraph research agent with RAG over 30 papers; deployed on GKE with autoscaling.

SKILLS

Programming & Delivery: Python, TypeScript, JavaScript, Java, C/C++, Go, FastAPI, Flask, Node.js, React, Next.js, SQL, Git, CI/CD
AI & Cloud: LLM evaluation, NLI, RAG, LangGraph, LangChain, Gemini/Vertex AI, OpenAI API, PyTorch, GCP, AWS, Docker, Kubernetes/GKE, Terraform, MySQL, MongoDB

ADDITIONAL

Certification listed on LinkedIn: AI Fluency: Framework & Foundations; Python. Additional prior skills include F#, scikit-learn, TensorFlow, Pandas, and HTML/CSS.