Chi(Bruce) Cheng

Mountain View, CA | 858-305-0278 | bruceche@andrew.cmu.edu | linkedin.com/in/chi-cheng-779b4a259/ | github.com/Bruce0921

Summary

Mid-level AI product developer and product manager with a strong background in building AI-driven conversational solutions. Delivered an AI call assistant that reduced average handling time by 35% and cut intentmatching costs sixfold while guiding two full product cycles across finance and healthcare domains. Built a medical chatbot using FastAPI, LangChain, and Gemini that improved response speed by 40% and supported over 50 disease diagnoses. Seeking to apply product strategy and technical expertise to accelerate AI product innovation and drive measurable business outcomes.

EDUCATION

Carnegie Mellon University, Silicon Valley — **Aug 2025 - Dec 2026**
Master of Science, Software Engineering (GPA: 3.5) — *Mountain View*

- **Coursework:** Software Engineering, Foundation of Computer System, Functional Programming, AI in Industry

University of California San Diego — **Sep 2021 - Apr 2025**
Bachelor of Science, Mathematics-Computer Science (GPA: 3.6)

- **Coursework:** Software Engineering, Design & Analysis of Algorithms, Advanced Data Structures, Computer Systems, Parallel Computing, Computer Security, Online Database Applications, Recommendation Systems, Computer Vision

SKILLS

- **Programming Skills:** Python, C/C++, Java, JavaScript, Go, Node.js, React, HTML, CSS, MySQL, MongoDB, Flask, Git, swift, F#, Express.js, AWS, Pandas, NumPy, FastAPI, CI/CD Deployment, RAG, LLM APIs, Prompt Engineering, NestJS, TypeScript
- **Product and Business Skills:** A/B Testing, Product Roadmap Planning, User Research, Market Analysis, Figma, Google Workspace, Agile/Scrum

EXPERIENCE

Helport AI | ***AI Product Developer/ Product Manager*** — **Sep 2024 - Jun 2025**

- Led product strategy for an AI-driven call assistant, conducting client interviews and user research to uncover pain points across mortgage, healthcare, insurance, and government sectors-driving 2+ full product development cycles.
- Designed and deployed an AI-powered dialogue flow system using FastAPI and a curated database, delivering faster and more accurate customer service responses—cutting average handling time by 35%
- Overhauled the intent-matching and recommendation system for real-time agent guidance, migrating from Google Dialogflow to Transformer + ID with Gemini 2.0 and Vertex AI-reducing costs 6 times ($0.002 → $0.00031 per match).
- Partnered with engineering and R&D teams to define technical requirements, manage agile sprints, and align product vision with technical feasibility-reducing development time from 14 to 5 days.
- Conducted A/B testing and performance analysis, uncovering optimizations that boosted sales conversion by 15% and improved AI model accuracy by 35%
- Promoted from internship to full-time AI Product Developer in March 2025 after building a dataprocessing pipeline with Pandas and NumPy that improved preprocessing efficiency

Convoloo | ***Software Development Engineer Intern*** — **Jul 2024 - Sep 2024**

- Developed a web-based AI medical chatbot using FastAPI, LangChain, React, Google Gemini, and Google Cloud, enabling accurate identification of 50+ common diseases through context-aware medical data.
- Implemented RAG-based knowledge retrieval with secure cloud hosting-reducing average response time from 2.5s to 1.5s and improving answer relevance for follow-up questions.
- Integrated Google Calendar API for appointment booking and enhanced UI/UX through iterative design reviews, enabling 50+ bookings during pilot testing.
- Led backend integration with LangServe and Langraph, optimizing service deployment and chatbot performance.

PROJECTS

Parking Spot Locator | https://psl.fogx.link | ***CMU / BOSCH*** — **Aug 2025 - Dec 2025**

- Built a **Bosch-sponsored parking spot locator** using vision-language models to semantically identify parking availability from visual sensor data.
- Implemented a **VLMap-based pipeline** combining RGB, depth, and pose data with CLIP embeddings for spatial reasoning.
- Developed and deployed a FastAPI backend on AWS to support real-time parking queries and semantic search

Capitawise (2nd Place, $ 7,000 Prize) | ***Franklin Templeton*** — **Mar 2024 - Jun 2024**

- Led full-stack development of an AI-powered banking chatbot using GPT-4o, Node.js, Flask, and React, reducing customer service response time by 40%.
- Implemented NLP-based query handling with text/voice output, improving banking query resolution accuracy from 68% to 85%.
- Designed a dynamic real-time UI/UX, boosting user engagement and retention during pilot testing.

PawPrints (1st Place, $15,000 Prize) | ***Franklin Templeton*** — **Mar 2023 - Jun 2023**

- Built a blockchain-based medical data exchange platform using React.js, Express.js, MySQL, Web3.js, and JWT authentication to securely connect pet owners, clinics, and insurers.
- Implemented token-based on-chain record storage and role-specific dashboards, enabling secure, permission-based access.
- Optimized backend APIs to cut data retrieval time from 3s to <1s, improving overall platform responsiveness.