Chi(Bruce) Cheng

Mountain View, CA | 858-305-0278 | bruceche@andrew.cmu.edu | LinkedIn | Github | bruceche.com

SUMMARY

MS student in Software Engineering at CMU with experience in AI systems, agentic workflows, and full-stack software development. Passionate about learning. Actively seeking Summer 2026 internships in software engineering, AI/agentic systems, or product-focused roles.

EDUCATION

Carnegie Mellon University, Silicon Valley **Aug 2025 - Dec 2026**
Master of Science, Software Engineering (GPA: 3.6) *Mountain View*

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

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

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

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, TypeScript, GCP, Next.js, Streamlit, LangGraph, LangChain, Kubernetes, Docker, Socket.io, Web3, Orchestration, FastMCP, Milvus, TensorFlow, Relational Database, Testing, Terraform, Hugging Face, vLLM, OpenAI API, PyTorch
- **Product and Business Skills:** A/B Testing, Product Roadmap Planning, User Research, Market Analysis, Figma, Agile/Scrum, LangSmith

EXPERIENCE

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

- Worked closely with R&D and engineering teams to design and implement backend systems for an AI-driven call assistant across mortgage, healthcare, insurance, and government sectors, using python, Docker, and CI/CD pipelines, which enhanced system reliability and reduced latency
- Built and deployed an AI-powered dialogue flow system using **FastAPI and a structured database**, improving response accuracy and reducing average handling time by **35%**.
- Redesigned the **intent-matching and recommendation pipeline**, migrating from Google Dialogflow to a **Transformer-based system using Gemini 2.0 and Vertex AI**, reducing inference cost by **6×** ($0.002 → $0.00031 per match).
- Implemented backend services and APIs to support real-time agent guidance, collaborating with cross-functional teams to ship production features and reduce development cycles from **14 days to 5 days**.
- Developed a data-preprocessing pipeline using Pandas and NumPy to optimize data ingestion and model input preparation, which accelerated data processing and shortened model training cycles; promoted from intern to full-time in March 2025
- Conducted **A/B testing and system performance analysis**, improving AI model accuracy by **35%** and increasing sales conversion by **15%**.

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 on secure cloud infrastructure using Terraform and MySQL, reducing average response time from 2.5 s to 1.5 s and increasing 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 of the chatbot with LangServe and LangGraph using FastAPI and NestJS, which streamlined service deployment and improved overall chatbot performance

PROJECTS

NightyNight | Demo Link | ***Independent AI Project*** **Apr 2025 - Present**

- **Built NightyNight**, a **full-stack** AI bedtime science story generator deployed on Render, with a React/TypeScript frontend and FastAPI backend communicating over **Server-Sent Events (SSE)** for real-time streaming progress.
- **Designed a LangGraph multi-agent pipeline** that parallelizes chapter writing across independent LLM nodes, then assembles and polishes into a coherent long-form narrative; supports age-adaptive prompting across 4 audience profiles (children ages 4–6 through adults).
- Defined the **product direction** around soothing narration, factual science content, child-friendly readability, and low-friction bedtime use cases.
- **Integrated ElevenLabs Flash v2.5 TTS** with one-shot generation (up to 40,000 chars) across 7 curated narrator voices

Research Assistant Agent | Project Code Demo | ***CMU Course Project*** **Jan 2026 - Mar 2026**

- Built a **LangGraph-based AI agent** with a **RAG pipeline** to answer questions from 30 indexed academic papers.
- Developed a **Streamlit web app** for paper upload, multilingual querying, and context-grounded answers with references.
- Deployed containerized services (**AI agent, translation API, vector DB**) on **Google Kubernetes Engine with auto-scaling**.
- Implemented **semantic retrieval using vector embeddings** and compared indexing algorithms (**HNSW, IVF_PQ, DiskANN**)

Parking Spot Locator | https://psl.fogx.link | ***CMU / BOSCH Research Collaborator*** **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
- **Optimized semantic search pipeline**, reducing parking availability query latency **from 45s to 15s (3× speedup)** through improved embedding retrieval and backend query processing.

Capitawise (2nd Place, $ 7,000 Prize) | ***Franklin Templeton (Investment firm)*** **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.