Brian Chu

(415) - 969 - 0765 | brian.chu1030@gmail.com | linkedin.com/in/brian-chu123

WORK EXPERIENCE

Stealth Startup | Remote **June 2024 - Present**

AI Engineer

- Deployed multi-provider AI agent system serving 3 LLM providers by building tool-calling agentic loops with structured Pydantic output via FastAPI microservices, replacing manual machinist quoting with autonomous AI-driven pricing across 7 cost categories
- Achieved 40% higher engineering drawing analysis accuracy by developing a VLM-based preprocessing pipeline using Gemini Vision for classification, assembly trees, and BOM extraction
- Consolidated three separate quote workflows into a unified microservices architecture (React, FastAPI, PostgreSQL) with recursive assembly quoting across arbitrarily nested BOMs, implementing async communication with retries and structured logging

Vulcan Engineering Solutions | Irvine, CA **June 2025 - Present**

ML Operations Engineer

- Accelerated model review cycles by 60% by implementing end-to-end MLOps pipeline with CI/CD workflows for model deployment, serving, monitoring, and versioning
- Streamlined daily evaluation workflows by designing LLM prompt engineering pipelines using Groq and LangChain, enabling structured output for multiple automated validations

Software Engineer Intern **January 2025 - June 2025**

- Improved structural defect detection accuracy by 45% by fine-tuning R-CNN models with systematic hyperparameter optimization using PyTorch and OpenCV
- Reduced average response time by 30% by designing RESTful API services using Python FastAPI

Jacobs School of Engineering, UCSD | La Jolla, CA **January 2023 - June 2023**

Process Engineer

- Achieved 30% higher energy efficiency and 95% conversion of methanol using Aspen Plus & COMSOL to design and simulate industrial scale dimethylamine reactor and heat exchanger system
- Reduced design review cycles by 25% by developing automated verification scripts for comprehensive Piping and Instrumentation Diagram (P&ID) documentation in Aspen Plus

PROJECTS

Speed Demon - F1 AI Racing Technology Automation **September 2025 - Present**

- Engineered real-time telemetry system using OpenTelemetry, with a time-series database on AWS EC2, integrating LangChain agents utilizing skills and MCP with GraphRAG to predict driver style
- Orchestrated a distributed data processing backend by implementing batch and stream processing using PySpark on cloud infrastructure with centralized monitoring

NLP Drug-Drug Interaction Prediction System **May 2025 - August 2025**

- Predicted combined drug effects with 82% accuracy by developing a Knowledge Graph Neural Network for 10K+ descriptions with TensorFlow and Neo4j
- Processed pharmacological text data by tokenizing with spaCy and generating feature pipelines using Hugging Face Transformers, improving model input quality
- Benchmarked KGNN performance against classical ML models using scikit-learn, including Logistic Regression, SVM, Random Forest, and KNN, improving evaluation rigor and comparison

EDUCATION

University of California, San Diego | San Diego, CA

Masters of Science in Chemical Engineering **August 2024 - December 2025**

Bachelor of Science in Computer Science and Chemical Engineering **August 2020 - June 2024**

SKILLS

Languages | Frameworks: Python | C++ | SQL | TypeScript | LangChain | TensorFlow | FastAPI | Pydantic
Tools | Infrastructure: AWS | PostgreSQL | Neo4j | CI/CD | Kubernetes | Docker | PySpark