Kai Breese

925-786-5550 | kai.breeze@me.com | linkedin.com/in/kaibreese | github.com/ninjakaib

Education

University of California San Diego — La Jolla, CA
Bachelor of Science in Data Science — *Sep. 2021 – March 2025*

Experience

Machine Learning Engineer Intern — June 2024 – Present
XiFin, Inc. — *San Diego, CA*

- Developed **IDE coding assistant** using open-source LLMs on AWS SageMaker, enabling context-aware code suggestions, inline editing, and real-time autocomplete.
- Built **automated insurance appeal letter system** with knowledge graphs and multimodal LLMs, cutting manual processing time.
- Optimized **insurance pricing pipeline** to achieve 90x throughput with multiprocessing and custom C backends.
- Redesigned pricing database to reduce data duplication by **75%** through normalization and optimized parsing.
- Created a **no-code deployment tool** for easy model deployment on AWS using custom Docker containers.
- Enhanced chatbot with streaming vision, reducing incorrect answer rate from **50% to 15%**.
- Developed a new **RAG pipeline** for agentic retrieval and visual document understanding.
- Designed an **AI SQL generator** with context-guided outputs for complex, accurate query formulation.

Software Developer — Mar. 2024 – July 2024
Clayton Health, LLC — *San Diego, CA*

- Built custom e-commerce platform for Shopify using Liquid driving **$100K in sales** within first 6 months
- Implemented automated email marketing campaigns leading to **8% conversion rate**
- Worked directly with CEO to rapidly bring new features from concept to production

Sales Consultant — Oct. 2023 – Apr. 2024
Verizon — *San Diego, CA*

- Ranked top 5 in San Diego district generating an average of over $100/hr in company profits
- Created long term relationships with high-value business clients to drive repeat sales

Computer Science Tutor — Mar. 2023 – Feb. 2024
TheCoderSchool — *San Diego, CA*

- Led one-on-one and classroom based sessions, teaching students in ways they can easily understand and relate to
- Helped students create over 100 unique projects including games, websites, and assignments

Projects

Machine Learning Hardware Accelerator Chip — Sept. 2024 – Present

- Implemented the **first ASIC-based L-Mul algorithm** in simulation, achieving up to **88.9% area**, **99.6% power**, and **80.7% delay** improvements versus IEEE-754 multipliers (FP8 to FP32) with ¡0.1% accuracy loss.
- Designed a full **accelerator core in Python** using PyRTL with a custom VLIW-style simulation framework and pipelined execution for efficient resource utilization.
- Developed a rudimentary **compiler** that converts a model's computational graph into an optimized instruction sequence via matrix tiling and topological sorting.
- Integrated RTL design, automated CI testing, and openROAD physical hardening to validate performance and guide design decisions.

AI Healthcare Assistant | *Spigot API, Java, Maven, TravisCI, Git* — June 2023 – Oct. 2023

- Developed a conversational AI phone call service before similar products from OpenAI came to market, achieving 80% API cost reduction over comparable solutions
- Designed SQL relational database with nutritional information scraped from the web
- Fine-tuned LLMs and deployed with AWS for real-time streaming and high-user load capacity

Technical Skills

Languages: Python, SQL, TypeScript, C/C++, Excel
Frameworks: FastAPI, Flask, Gradio, Svelte, Nginx, React, Liquid
Developer Tools: Git, Docker, Anaconda, Yosys, Jupyter, Jira, VS Code, ChatGPT
Libraries: pandas, NumPy, PyTorch, matplotlib, sklearn, transformers, langchain, Cython, PyRTL
Cloud: AWS Sagemaker, Lambda, S3, RDS, ECR, EC2, API Gateway, Oracle DB