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*

- Building custom inference containers to serve ML models on AWS and deploying them as secure API endpoints
- Improved insurance pricing data pipeline throughput by over 100x with highly-parallelized C backends
- Designed AI agent using fine-tuned LLMs and knowledge graphs to automatically appeal denied insurance claims
- Trained and productionized a transformer-based model to parse, ingest, and autofill medical insurance forms
- Enhanced company chatbot with multimodal document RAG, reducing incorrect answer rates from 52% to 17%
- Reduced pricing database storage requirements by 75% by designing an efficient, highly normalized schema
- Developed internal coding assistant using open-source LLMs featuring chat, editing, and autocomplete

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

- Built custom e-commerce platform for Shopify using LiquidJS, 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

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

Deep Learning Hardware Accelerator — Sept. 2024 – Present

- Implemented energy-efficient floating point multipliers in hardware achieving 89.1% area, 94.3% power, and 66.0% delay improvements vs IEEE-754 multipliers with <0.1% model inference accuracy loss
- Designed accelerator core with pipelined matrix multiplication engine, accumulator memory, and activation unit
- Developed basic compiler to transform models into instructions through matrix tiling and graph optimization
- Created Python package to design and simulate accelerators using composable and configurable hardware blocks
- Integrated automated functional verification, and OpenROAD physical hardening to validate performance metrics

Credit Risk Modeling ML Competition — Mar. 2024 – June 2024

- Won 1st place with 0.903 AUC score, indicating better predictive power for loan defaults than actual FICO scores
- Trained and optimized low-depth boosted decision trees to predict loan default risk based on transaction history for customers without credit scores
- Engineered 4000+ features from time-series banking data using multi-layered statistical aggregations

AI Healthcare Coach — June 2023 – Oct. 2023

- Developed real-time conversational AI phone service for 80% cheaper than Twilio API (before GPT-4o released)
- Designed SQL database with nutritional information for common foods and restaurants scraped from the web
- Fine-tuned LLMs and deployed on AWS EC2 for real-time streaming and high-user load capacity

Technical Skills

Languages: Python, SQL, C/C++, Verilog, TypeScript
Developer Tools: Docker, Git, Huggingface, Yosys, OpenROAD
Cloud: AWS Sagemaker, Lambda, S3, RDS, ECR, EC2, API Gateway, GitHub Actions, Oracle DB
Libraries & Frameworks: PyRTL, PyTorch, transformers, vLLM, TGI, peft, Unsloth, pandas, sklearn, NumPy, Numba, Dask, Cython, FastAPI, gradio, langchain, boto3, Apache TVM