

Abraham Guan

Berkeley, CA | abrahamwg123@gmail.com | +1 (425) 365-2439 | Portfolio | LinkedIn | GitHub

EDUCATION

University of California, Berkeley

Bachelor of Arts, Data Science

Berkeley, CA

Expected May 2027

Coursework: Data Structures & Algorithms, Machine Learning, Statistical Analysis, Exploratory Data Analysis, Probability

EXPERIENCE

NASA Ames Research Center

Jan 2026 – Aug 2026

Machine Learning Research Intern

Remote

- Raised spaceflight-signal detection from **0.5 (near-chance)** to **0.83–0.97 AUROC** across five tissue types, overturning the prior conclusion that spaceflight leaves no detectable signal in tissue by tracing that result to an evaluation flaw rather than a biological one.
- Built the team's reproducible **benchmarking harness**, the first fair comparison of foundation models on space biology data: five foundation models against classical baselines across three tasks on ACCESS GPU clusters, over a **6-stage preprocessing pipeline** unifying 2.3M samples with NASA spaceflight RNA-seq data into one comparable gene space.
- Diagnosed **batch-effect leakage** in the original random train/test splits, where models were scoring on which lab produced a sample instead of on spaceflight, and rebuilt the protocol on study-grouped **cross-validation** so every model is judged on biology.

Molex

May 2026 – Aug 2026

Machine Learning Engineer Intern

Fremont, CA

- Cut mirror swapping time by **74%** by training a ranking model using **supervised machine learning** algorithm LambdaMART, evaluated by nDCG@k, to predict the best performing parts, replacing a slow manual trial and error process.
- Built a custom **Gaussian Process kernel** to capture a repeating pattern in each optical module, cutting prediction error by **30%** over the team's hand-weighted model and removing days of testing from every new module.
- Owned **14 prediction models and pipeline** end to end, from scattered Excel files to one clean dataset, through feature engineering with ANOVA, correlation and clustering studies, to benchmarking nonlinear methods against simple baselines.

Umbo Energy

Sep 2025 – Dec 2025

Machine Learning Engineer Intern

Remote

- Designed an offline Reinforcement Learning (RL) system that achieved a 7% reduction in annual energy consumption using a three-model architecture system (Behavior Cloning, Surrogate Model, and CQL).
- Implemented Conservative Q-Learning (CQL) to enforce safety constraints, preventing equipment damage during model exploration.

Stockbit

Jul 2024 – Sep 2024

AI Engineer Intern

Jakarta, Indonesia

- Built the retrieval layer for a RAG assistant serving an investing platform, combining **BM25** keyword search with a query parser that resolves stock tickers and financial metric synonyms to canonical fields, reaching **0.87 Recall@100** and **0.89 F1** on a held-out query set.
- Designed a staged evaluation framework that separates retrieval failure from generation failure, measuring whether the chunk was retrieved at all, then scoring answer faithfulness with **LLM-as-a-Judge** validated against human labels.
- Engineered a BigQuery and Kubernetes data pipeline for real-time financial data, cutting research time from **5+ minutes to under 30 seconds**, and fine-tuned **Llama 3** for a customer FAQ assistant on the same stack.

PROJECTS

B2B Sales Propensity Model | Python, XGBoost, Scikit-Learn | GitHub

- Developed an XGBoost lead-ranking model to identify ready-to-buy B2B prospects, engineering high-value signal features and resolving a 1:9 class imbalance to achieve a 2.9x higher conversion rate than baselines.

IShowStream AI Co-Pilot for Streamers | Python, Go, Gemini, Google Cloud Run, WebSockets | GitHub

- Built a real-time Twitch AI co-pilot processing 1000+ messages/minute, deploying 5 Gemini agents to automatically extract stream highlights by analyzing chat velocity, semantic sentiment, and keyword spam.

GitHub Documentation Agent | Python, FastAPI, React, Google Gemini API | GitHub

- Architected a multi-agent AI system using **Model Context Protocol (MCP)** with 3 specialized agents in an orchestrator pattern to generate adaptive documentation, cutting generation time by 60% via parallel execution on an async **FastAPI** backend, with a **React** 3-level viewer and Docker/Cloud Run deployment handling repos with 10K+ files at 99% uptime.

SKILLS

- **Languages:** Python, SQL, Java, C++, HTML/CSS
- **AI/ML & Agents:** TensorFlow, PyTorch, RAG, Reinforcement Learning, MCP, NLP, Hugging Face, LangChain
- **Frameworks & Tools:** AWS, Pandas, NumPy, Node.js, RESTful APIs, Docker