

Alex Morgan

AI / Machine Learning Engineer

San Jose, CA • (408) 555-0163 • alex.morgan@email.com • linkedin.com/in/alexmorgan-ml • github.com/amorgan-ai

SUMMARY

Machine Learning Engineer with 3.5 years of experience building and deploying production ML, deep learning, NLP, and LLM-powered systems. Shipped models serving 5M+ predictions per day, improved model accuracy by up to 18%, and cut inference costs by 40%. Experienced across the full ML lifecycle: data pipelines, training, evaluation, MLOps, monitoring, and cloud deployment.

TECHNICAL SKILLS

Languages: Python, SQL, C++, Bash

ML / DL: PyTorch, TensorFlow, Scikit-learn, XGBoost, Keras

NLP & GenAI: Hugging Face, LLMs, RAG, LangChain, Fine-Tuning, Prompt Engineering

Vector & Data: Pinecone, FAISS, Spark, Pandas, Kafka, PostgreSQL

MLOps: MLflow, Kubeflow, Docker, Kubernetes, Airflow, CI/CD, Weights & Biases

Cloud: AWS (SageMaker, S3, Lambda), GCP (Vertex AI), Azure ML

PROFESSIONAL EXPERIENCE

Machine Learning Engineer

Apr 2024 – Present

Vertex Intelligence Inc. — San Jose, CA

- Built a RAG-based enterprise search assistant (LLM + FAISS) answering 10K+ queries/day with 92% answer relevance, reducing support tickets by 35%.
- Fine-tuned transformer models (BERT, LLaMA) for document classification, improving F1 score from 0.78 to 0.92 (+18%).
- Deployed real-time recommendation models on SageMaker serving 5M+ predictions/day with p99 latency under 80 ms.
- Reduced inference costs by 40% through model quantization, distillation, and batch-serving optimization.
- Implemented MLflow tracking, automated retraining, and drift monitoring, cutting model release time from 3 weeks to 4 days.

Data Scientist / ML Engineer

Jul 2022 – Mar 2024

Lumina Analytics — Sunnyvale, CA

- Developed computer vision models (ResNet, YOLO) for defect detection with 96% precision, reducing manual inspection effort by 50%.
- Built NLP pipelines for sentiment and entity extraction on 2M+ documents using spaCy and Hugging Face.
- Created feature pipelines in Spark and Airflow processing 500GB+ of data daily with 99% reliability.
- Containerized models with Docker and Kubernetes; collaborated with engineering to ship 6 models to production.

PROJECTS

- Open-Source LLM Evaluation Toolkit:** Python library benchmarking LLM outputs for factuality and toxicity; 600+ GitHub stars.
- Fraud Detection Model:** Gradient-boosted ensemble achieving 0.97 AUC on imbalanced transaction data; deployed via FastAPI.

EDUCATION

M.S. in Computer Science (Machine Learning)

2022

San José State University | B.S. in Computer Engineering, UC Riverside (2020)

CERTIFICATIONS

AWS Certified Machine Learning – Specialty • TensorFlow Developer Certificate • DeepLearning.AI Deep Learning Specialization