

AMOS JAIMES

MA, USA | +1 (617) 934-9015 | amosjaimes01@gmail.com | LinkedIn

SUMMARY

AI/ML Engineer with around **5 years** of experience in building production **LLM** systems, **Generative AI** applications, and scalable **distributed backend infrastructure**. Hands-on across the full stack **RAG** pipelines, **model fine-tuning**, **agentic AI**, **microservices architecture**, and **cloud-native** deployments on **AWS**.

EXPERIENCE

Piper Sandler

Jun 2025 – Nov 2025

AI/ML Engineer

Boston, MA, USA

- Worked on a **RAG**-based document search tool using **LangChain** and **FAISS** over 10K+ SEC filings and earnings transcripts, cutting analyst lookup time from around 8 seconds to under 400ms.
- Contributed to a **Generative AI** portfolio assistant using **OpenAI API** and **Claude API**, helping the investment team reduce client memo prep time by roughly 6 hours per report.
- Supported **LLM** deployment on **AWS SageMaker** and tracked experiments with **MLflow**, helping the team move toward a structured **MLOps** workflow during a period of rapid AI adoption.
- Assisted in evaluating **NLP** sentiment classifiers on earnings call transcripts for the equity research pipeline, working in sprint cycles alongside senior **ML engineers**.

Dell Technologies

Aug 2022 – Aug 2024

Software Engineer

India

- Partnered with platform engineering and operations stakeholders to scope ambiguous requirements, translated them into a concrete technical design, and overhauled the **microservices** layer for Dell's OpenManage server platform using **Java Spring Boot** and **Kubernetes**, scaling concurrent session capacity from about 1,200 to 8,000.
- Built and owned a **REST API** gateway in **Python FastAPI** with **OAuth 2.0** auth and rate limiting, processing close to 45 million calls per month across 12 data center regions.
- Migrated a legacy firmware telemetry service to **Apache Kafka event-driven architecture**, reducing ingestion latency from 22 seconds to just over 1 second for 600K+ connected endpoints.
- Automated release workflows using **Jenkins** and **GitHub Actions** across 14 backend services, cutting deployment cycles from 3 weeks to about 4 days.
- Instrumented **Prometheus** metrics and **Grafana** dashboards across 8 backend services and presented findings directly to non-technical operations managers, translating alert thresholds into business-impact language that reduced mean time to detect production anomalies from about 40 minutes to under 6 minutes.
- Optimized **MySQL** query patterns for the server inventory service using **Redis** write-through caching and query batching, reducing average database read latency from 1.4 seconds to under 180ms across 250K+ asset records.

Airbnb

May 2020 – Jul 2022

Software Developer

India

- Translated business requirements from product and design into technical specs for host availability and messaging **REST API** endpoints in **Python** and **Django**, shipping through full code review and QA cycles as part of an 8-person engineering team.
- Reduced **PostgreSQL** query time on the search results page from about 3 seconds to under 800ms by profiling with EXPLAIN ANALYZE and adding composite indexes.
- Added **Redis** caching for high-traffic host listing availability queries, bringing repeat-request latency from 2.1 seconds to under 200ms and cutting database read load by roughly 3,000 queries per minute at peak.
- Shipped reusable **React** and **JavaScript** UI components for the host analytics dashboard, following Figma design specs across browsers and screen sizes.
- Worked directly with the host operations team to surface requirements and in-house constraints for a bulk-messaging feature, then implemented the back-end logic using queued **Python** background tasks, allowing hosts to notify up to 500 guests at once without blocking the main application thread.
- Wrote **unit tests** in **pytest** for new API endpoints and took part in weekly code reviews, helping the team catch regressions before they reached staging.

PROJECTS

Kube-LoRA: Distributed MLOps Fine-Tuning Platform | GitHub

- Created a **Kubernetes**-native fine-tuning platform driving training, scoring, and inference workloads from a **FastAPI** gateway with a dedicated ServiceAccount, RBAC scope, and a 14-endpoint REST surface.
- Optimized rank-32 **LoRA** fine-tuning of Stable Diffusion v1.5 in **PyTorch** and Diffusers with fp16 mixed precision, gradient checkpointing, SDPA, and 8-bit Adam, fitting training inside a 10.8GB budget on a 12GB RTX 3060.
- Integrated **Celery** and **Redis** workers to lift long-running provisioning, training, and scoring off the HTTP path, watching the full job lifecycle asynchronously.
- Instituted an automated quality gate that scores samples with a **TensorFlow** ResNet50 transfer classifier and promotes passing weights to a shared volume, then hands the GPU to a scale-to-zero inference server guarded by a 503-until-warm readiness probe.

ChestX-ray14 Classification & Monitoring Platform | GitHub

- Modeled 14-label thoracic pathology detection over the 112K-image NIH ChestX-ray14 set with a **PyTorch** DenseNet-121 trained under mixed precision and gradient accumulation.
- Engineered the training loop with class-weighted BCE loss, cosine and plateau schedulers, early stopping on validation AUC-ROC, and top-k checkpointing logged to **MLflow**.
- Developed three-layer drift detection combining Kolmogorov-Smirnov tests on pixel intensity, chi-squared tests on label distribution, and MMD with an RBF kernel over embeddings, alerting on rolling-window AUC decay.
- Implemented GradCAM overlays for top-k findings and per-class threshold tuning via Youden’s J, deployed through **Azure ML** with **Terraform** provisioning and blue-green canary promotion.

Dynamic Visual Ranker: Retrieval & Learning-to-Rank Engine | GitHub

- Constructed a two-tower retrieval stage in **PyTorch** using **CLIP** to embed product imagery and text metadata for large e-commerce catalogs
- Trained an **XGBoost** XGBRanker learning-to-rank model against NDCG, re-ordering results by user style profile and live inventory constraints
- Structured the feature pipeline in **PySpark** and served vectors from a **Redis** feature store under 100ms, with daily retraining scheduled in **Airflow** and runs logged to **MLflow**
- Launched the ranker as a containerized **FastAPI** microservice on **Docker**, wiring Evidently AI to watch prediction distributions and flag feature drift

Research Reasoning Agent: Autonomous ReAct System | GitHub

- Devised an autonomous ReAct agent on **LangGraph** that plans, calls tools, and iterates on its own observations, returning Pydantic-validated synthesis, follow-up questions, and a testable hypothesis with a confidence level.
- Minimized token spend with a tiered routing strategy, running the reasoning loop on **Claude Haiku** and reserving **Claude Sonnet** for final synthesis under tuned recursion limits and trimmed context.
- Established a local knowledge base through arXiv ingestion, PyPDF extraction, sentence-transformer embeddings, and a **FAISS** store, extended with live web search over the **Model Context Protocol** and graceful degradation when unavailable.

EDUCATION

Northeastern University <i>Master of Science, Applied Machine Intelligence</i>	Boston, MA, USA <i>Sep 2024 – Apr 2026</i>
Kristu Jayanti College <i>Bachelor of Science, Computer Science</i>	India <i>Jun 2019 – May 2022</i>

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, JavaScript, Go, SQL, Bash, Scala, C++
AI/ML & GenAI: Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), Agentic AI, PyTorch, TensorFlow, LangChain, LangGraph, LlamaIndex, Hugging Face Transformers, scikit-learn, FAISS, vLLM, ONNX Runtime, fine-tuning (LoRA, QLoRA, PEFT), prompt engineering
LLM APIs & Platforms: OpenAI API (GPT-4o), Anthropic Claude API, AWS Bedrock, Azure OpenAI, Google Vertex AI, function calling, structured outputs, LLM-as-judge, semantic caching
Cloud & MLOps: AWS, AWS SageMaker, AWS Lambda, AWS EKS, Docker, Kubernetes (K8s), Terraform, MLflow, Weights & Biases, Apache Airflow, Prometheus, Grafana, GitHub Actions, Jenkins, CI/CD Pipelines
Backend & Distributed Systems: Spring Boot, FastAPI, Django, REST APIs, GraphQL, gRPC, Apache Kafka, Apache Spark, microservices architecture, event-driven architecture, system design, distributed systems
Databases & Data: PostgreSQL, Redis, MySQL, MongoDB, Elasticsearch, Pinecone, ChromaDB, Weaviate, vector databases, Delta Lake, dbt

CERTIFICATIONS

Claude Certified Architect – Foundations | Anthropic