

Andy Rosales-Elias

Senior AI/ML & Full Stack Engineer

+1 2097039755 | rosaleseliasandy@gmail.com | <https://www.linkedin.com/in/andyrosales/> | Brooklyn,
New York

SUMMARY

Senior Backend & AI/ML Engineer with 8+ years of experience building production-grade backend systems and large-scale ML platforms. Expertise in distributed systems, scalable ML infrastructure, model monitoring, and real-time AI applications. Proven track record designing Kubernetes-based training pipelines, anomaly detection systems, and Retrieval-Augmented Generation (RAG) AI assistants. Strong technical leader and collaborator, mentoring teams and aligning AI/ML development with business objectives.

EXPERIENCE

Senior Software Engineer, ML Platform

Lyft | New York, NY

04/2023 – Present

- Designed and launched a production AI Assistant Platform using RAG, handling thousands of internal queries daily.
- Built a company-wide anomaly detection system with Airflow, Spark, and Fugue, identifying both spikes and gradual shifts in incident rates.
- Integrated anomaly detection with LyftLearn ML training infrastructure, enabling automated feedback loops for model incidents.

- Developed Model Health Monitoring system calculating unified KPIs across deployed ML models, improving observability, reliability, and MTTR.
 - Led architecture for ML services emphasizing scalability, fault tolerance, and operational simplicity.
-

Software Engineer, ML Platform

Lyft | San Francisco Bay Area

04/2020 - 04/2023

- Lead engineer for LyftLearn Kubernetes-based ML training platform, supporting multiple teams (dispatch, pricing, fraud detection).
 - Built and scaled distributed model training pipelines using Kubernetes, Spark, and Fugue for scikit-learn, LightGBM, XGBoost, PyTorch, TensorFlow.
 - Standardized training workflows and resource management, increasing platform reliability and developer velocity.
 - Partnered with ML teams to productionize models and streamline ML lifecycle processes, including model deployment, monitoring, and retraining pipelines.
-

Software Engineering Intern

Lyft | San Francisco Bay Area

07/2019 - 04/2020

- Designed backend services for ML observability tools, improving model monitoring and debugging.
 - Integrated real-time metrics pipelines for tracking online ML model performance.
-

DevOps Intern

InTouch Health | Santa Barbara, CA

04/2019 - 07/2019

- Enhanced backend/frontend of Device Manager platform supporting medical devices via REST APIs.
 - Improved system reliability and usability through iterative feature development.
-

Undergraduate Researcher

UC Santa Barbara | Santa Barbara, CA

06/2016 - 07/2019

- Developed CNN-based computer vision models for wildlife detection in SmartFarm IoT platform.
 - Built OCR system extracting temperature readings with 100% accuracy, achieving 5× performance improvement.
 - Applied supervised learning and data labeling techniques to improve model robustness.
-

Additional Experience (Internships & Projects)

- Machine Learning Intern, GitHub: Built NLP-based code duplication detection tool; curated large-scale datasets.
 - ML Software Engineering Intern: Prototyped deep learning-driven interactive fitness applications.
 - Summer Institute of Math and Science Intern, CNSI: Developed unsupervised OCR model achieving 96% accuracy.
 - AI Campus Ambassador, InTouch Health: Presented AI research and Intel AI technologies to university community.
-

EDUCATION

Bachelor of Science (BS), Computer Science

UC Santa Barbara | 08/2015 – 05/2019

SKILLS

AI / ML: TensorFlow, PyTorch, Hugging Face, LangChain, RAG, Scikit-Learn, Kubeflow, MLflow, Model Bias Detection, A/B Testing, TensorFlow.js, OpenCV

Cloud & MLOps: AWS (SageMaker, Bedrock, ECS, Lambda, EC2, S3, Fargate), GCP (Vertex AI, Dataflow, Cloud Run, BigQuery), Azure (Cognitive Services, Functions, DevOps, ML), Terraform, Chaos Engineering, Istio Service Mesh, Kubernetes, Helm, Docker Containerization, Jenkins, CI/CD Pipelines

Programming Languages: Python, Go (Golang), TypeScript / Javascript, C#, PHP, CSS

Data Engineering & Databases: Snowflake, PostgreSQL, MySQL, MongoDB, Redis (Caching, Pub/Sub, Semantic), GCP, Azure SQL, Weaviate, Milvus, Pinecone, ETL, Data Modeling, Schema Design, Vector, Embeddings, Plots, Seaborn, Pandas

API & Backend: RESTful APIs, GraphQL, Node.js (Express), Go (Golang), C# (.Net Core), PHP(Laravel), Python, FastAPI, Django, API Design, API Gateways, Authentication, Authorization, Performance Tuning, WebSockets, Kafka, CORS, JWT / OAuth2

Frontend: React, Redux, Tanstack, Zustand, TypeScript / Javascript, Next.js, Material UI, Tailwind CSS, Context API

Security & Compliance: HIPAA Compliance, Data Encryption, Authentication (OAuth2, JWT, SSo), Authorization, Secure API Design, Rate Limiting, AWS Security Hub, GCP Security Command Center, Azure Security Center

LLM Evaluation: RAGAS, TruLens, Promptfoo/Garak, SHAP, LIME, Captum, Red-teaming, Bias/Drift
QA