

# Arhum Savera

Silver Spring, MD | [arhum.a.savera@gmail.com](mailto:arhum.a.savera@gmail.com) | +1 973 769 9824 | [github.com/arhumsavera](https://github.com/arhumsavera) | [linkedin.com/in/arhumsavera](https://linkedin.com/in/arhumsavera)

## SUMMARY

---

Senior Software Engineer with 7+ years building scalable Python backend systems and ML/AI platforms. Architected event-driven microservices processing thousands of matching decisions per hour with Kafka, built AI pipelines with LangChain/OpenAI and vector search, and optimized computer vision models for edge deployment at Amazon. Led modernization of 15+ microservices across a high-traffic consumer marketplace.

## EXPERIENCE

---

**Senior Software Engineer**, ConsumerAffairs, Remote

*Nov 2021 — Present*

- Built an AI content platform from scratch using FastAPI and LangChain, integrating OpenAI APIs for review summarization and moderation with pgvector semantic search, achieving 10x faster filtering and automating 80% of manual review
- Architected Kafka event streaming across 8+ microservices for real-time analytics on matching decisions, lead scoring, review events, and budget tracking, improving system throughput by 5x
- Led platform modernization of 5+ services from Python 3.8/Django 3.x to Python 3.11/Django 4.2/Celery 5.4, standardizing tooling and eliminating security vulnerabilities
- Designed real-time budget enforcement and CPL pricing systems using async task architecture and query optimization, reducing database queries by 40% and Redis calls by 60%
- Built ML-driven referral scoring pipeline integrating Databricks models with multi-source data aggregation and Kafka publishing for real-time lead quality assessment

**Research Engineer**, Amazon, Boston, MA

*Jan 2020 — Sep 2021*

- Owned the computer vision pipeline for Amazon Dash Cart, optimizing TensorRT models on Nvidia edge devices deployed across 100+ production smart shopping carts
- Engineered GPU-accelerated video encoding achieving 40x size reduction (3.1 GB to 70 MB) and 90% faster uploads using Nvidia Jetpack, enabling real-time video analytics
- Delivered client-side encryption with AWS KMS in a 6-week sprint, coordinating across 4 teams to meet compliance deadline with zero production issues
- Built real-time MQTT messaging between services, automated DynamoDB/Lambda jobs for continuous model learning, and CloudWatch monitoring across 100+ carts

**Machine Learning Scientist**, Signify (Philips Lighting), Cambridge, MA

*Jun 2019 — Jan 2020*

- Returned post-masters to optimize deep learning models for edge deployment using iterative weight pruning and post-training quantization via TensorFlow, achieving 60% size reduction and 30% inference speedup
- Designed and trained multi-task models for satellite image segmentation using street view imagery
- Deployed containerized ML inference services using Flask and AWS ECR with automated model documentation and tracking

**Applied ML Intern**, Signify (Philips Lighting), Cambridge, MA

*May 2018 — Aug 2018*

- Built end-to-end ML pipeline for LightFinder using multi-output deep learning models (VGG, Mask-RCNN, Siamese networks) for image classification and retrieval, deploying on AWS EC2 with transfer learning
- Wrote custom web crawling and feature engineering scripts to build training datasets and stakeholder visualizations

**Software Engineer**, Techlogix, Karachi, PK

*Jul 2016 — Jun 2017*

- Developed SOAP services and middleware integrations using Java, IBM Integration Bus, and WebSphere MQ for one of the UAE's largest money exchange companies
- Built a web-based remittance system eliminating manual correspondent bank processing and delivering real-time transaction updates

## EDUCATION

---

**M.S. Computer Science**, University of Massachusetts Amherst — GPA: 3.9/4.0

*2017 — 2019*

**B.S. Computer Science**, National University of Computer & Emerging Sciences — GPA: 3.8/4.0

*2012 — 2016*

## SKILLS

---

### Languages

### ML / AI

### Backend

### Data & Streaming

### Cloud & Infrastructure

### Protocols & Tools

Python, C++, JavaScript, Java, SQL

LangChain, OpenAI, TensorFlow, TensorRT, pgvector, Keras, OpenCV

Django, FastAPI, Flask, Celery, Pydantic

Kafka, PostgreSQL, DynamoDB, Redis, Typesense, MySQL

AWS (Lambda, S3, KMS, EC2, ECR, CloudWatch), Docker, GCP, CI/CD

MQTT, Git, Linux, Ruff, Intel oneTBB