

# Antony Ibrahim

(909) 257-1065 | antonyibrahim0@gmail.com | antonyibrahim.com  
github.com/AntonyI1 | linkedin.com/in/antony-ibrahim

## Summary

AI-focused Software Engineer skilled in building intelligent systems, RAG-powered chatbots, and cloud-native automation using **Python**, **AWS**, and modern LLM frameworks. Experienced in optimizing pipelines, deploying scalable AI services, and integrating retrieval-based architectures.

## Experience

### Genesis Healthcare

AI & Cloud Systems Engineer / Software Engineer

Jun. 2024 – Present

Remote — California, USA & Japan

- Built enterprise chatbot systems on **AWS Bedrock**, integrating **Claude 3.5 Sonnet** and **GPT-4o** with a **RAG framework** to help marketing teams query company data conversationally.
- Developed optimized knowledge bases and embedding pipelines to improve context matching and ensure responses aligned with business data.
- Refactored multi-database **ETL pipelines** handling **21M+ rows**, reducing runtime from 2 hours to 15 minutes (**88% faster**) through cleaner code structure and efficient data handling.
- Automated **cloud deployment** and data workflows with **Apache Airflow**, **Python**, and **AWS Lambda**, saving 2+ hours per release and improving reliability.
- Cut monthly infrastructure costs by **85%** by migrating analytics from Azure VMs to a secure **AWS EC2** environment with proper network isolation and access controls.

## Projects

### Daygent — AI Stock Analysis Assistant

Jan 2025 – Jun 2025

JavaScript, Python, LM Studio, Hugging Face, Claude, DeepSeek, OpenAI API, FAISS, Web Speech API

- Developed a local AI assistant for stock trend analysis and query handling using **LM Studio** with open-source **Hugging Face models**.
- Integrated **FAISS**-based vector search and retrieval workflows to provide relevant financial insights and contextual responses.
- Experimented with and compared performance across **Claude**, **DeepSeek**, and **OpenAI** models for reasoning accuracy and latency.
- Deployed and optimized inference workloads on local and **AWS EC2** environments for performance and cost benchmarking.
- Designed a browser interface using **Web Speech API** for hands-free query interaction and result narration.

## Technical Skills

**Languages:** C++, Python, C#, Java, JavaScript, TypeScript, SQL, HTML, CSS, TailwindCSS, MySQL, PostgreSQL

**Technologies:** AWS, .NET, Vue, Angular, Linux, Git, React, Node.js, Express.js, Amazon Aurora, MongoDB, RESTful API

**Developer Tools:** VS Code, Visual Studio, JupyterHub, Metabase, Docker, Apache Airflow

**Libraries:** Pandas, NumPy, DESeq2, Bedrock Vector Search

## Education

### University of California, Irvine

Bachelor of Science in Software Engineering

Sep. 2021 – June 2025

Irvine, CA

**Relevant Coursework:** Data Structures and Algorithms, Operating Systems, Database Management Systems (SQL/NoSQL), Distributed Systems, Agile Software Development, Web Applications (React, Angular, Vue)