

Sumit Vaise

Senior Applied ML Engineer | AI Developer

sumitvaise29@gmail.com | +1-263-366-0093 | [Linkedin](#) | [Github](#)

SUMMARY

Total 9+ years in AI/ML Engineering, specializing in model building (ML/DL/NLP, 5+ yrs), designing, developing, and deploying high-impact, AI-powered features and scalable systems. Experience building solutions with **LLM frameworks (LangChain, LlamaIndex, Hugging Face) to build reasoning systems and agentic workflows for asset and work intelligence.**, including deep familiarity with **prompt engineering, RAG**, and evaluation techniques. Proven track record of **architecting** and deploying **scalable**, production-grade AI applications on **AWS(SageMaker, Bedrock) and GCP (SageMaker, Vertex AI, EKS, GKE)**. Deep expertise in **MLOps**, end-to-end ML pipeline delivery, and leading cross-functional engineering teams to translate business requirements into high-impact AI solutions.

SKILLS

- **Programming & Development:**
 - **Languages & Frameworks:** Python, Scikit-Learn, TensorFlow, Keras, PyTorch, NLTK, Hugging Face Transformers, OpenCV, FastAPI, Flask, Streamlit.
 - **Tools & Libraries:** Pandas, NumPy, Matplotlib, PySpark, OpenAI API, REST APIs, Docker, Kubernetes, GitHub Actions, CI/CD Pipelines.
 - **Database Systems:** SQL, NoSQL, Vector Databases (Pinecone, Faiss, Chroma DB, OpenSearch).
- **Data Science & AI:**
 - **Core Skills:** Machine Learning, Deep Learning, Generative AI, **LLMs** (GPT, Gemini, LLaMA), Natural Language Processing (NLP), Computer Vision, Model Deployment, Prompt Engineering.
 - **Frameworks & Tools:** LangChain, **RAG** (Retrieval-Augmented Generation), LlamaIndex, LangGraph, LangSmith, AutoML.
 - **Applications:** AI-Driven Process Automation, Predictive Modeling, Statistical Analysis, Data Visualization.
- **Cloud & MLOps:**
 - **Platforms:** AWS (SageMaker, Lambda, Bedrock, EC2), **GCP** (GKE, Vertex AI, BigQuery, BQML, Artifact Registry), Heroku.
 - **Practices:** MLOps, LLMOps, AIOps, Cloud Machine Learning, **KubeFlow**, Orchestration Frameworks.
 - **Monitoring, Logging & Governance:** Grafana, Prometheus, ELK Stack concepts, MLflow, Model Versioning, Data Lineage, observability, IAM Security & Compliance.
- **Management & Operations:**
 - Agile Methodologies, Microservices Architecture, Product Management, Team Leadership, Stakeholder Management.

EXPERIENCE

Sr. ML Engineer | Fidelity Investments, Boston, MA, USA Nov 2024 - Dec 2025

- Built and deployed from POC to production a Multi-Document **RAG** application in AWS (**Bedrock, GPT, LangChain, HuggingFace, and OpenSearch, REST API**) to cut analyst research time by 40% across 500+ financial reports, working directly with finance stakeholders to understand and automate business processes.
- Decoupled embedding generation into an independent **EKS microservice** for separate **scaling**; integrated Amazon Bedrock (Llama 3) with prompt grounding; monitored via Datadog with full inference logging to S3 for audit and continuous improvement.
- Designed, built, and maintained core components of the internal **MLOps** pipeline, establishing robust CI/CD pipelines utilizing Jenkins for automated model testing and seamless deployment.
- Developed **end-to-end ETL and ML pipelines** for real-time communication error detection using **AWS, Airflow, Snowflake, Flask, and EKS**; built RESTful APIs (FastAPI) for seamless user interaction, containerized with Docker.
- Leveraged AI-assisted coding (GitHub Copilot) to accelerate SDLC, reducing boilerplate development time by 30% while maintaining high unit-test coverage in SONAR.

Sr. MLOps Developer | Definity Financial Corp, Waterloo ON May 2024 - Oct 2024 (Contract - Remote)

- Designed and maintained highly scalable GCP infrastructure for next-generation ML solutions, leveraging **Docker, Kubernetes (GKE), Vertex AI, and KubeFlow** for orchestrating production workloads.
- Built and maintained robust, low-latency cloud-native applications on **GCP (Cloud Run, Vertex AI, BigQuery)**, containerizing ML pipelines with **Docker and CI/CD** for seamless integration with production systems serving 14M+ records.
- Utilized BitBucket and JIRA for rigorous version control, deployment tracking, and task management, collaborating closely with product managers and software engineering teams.
- Successfully operationalized complex predictive pipelines into secure production environments, ensuring adherence to data privacy and governance best practices.

Sr. Machine Learning Engineer / Team Lead | Quantiphi Inc. July 2021 - Mar 2024 (Remote)

- Inventory Allocation Prediction (Team Lead, Team of 4): Optimized US shipment locations; **model building (Scikit-Learn) and scalable MLOps pipeline (KubeFlow, BigQuery, GKE, BQML) deployment into production, including model monitoring and evaluation tooling.**
- Insurance Claims Processing (Team of 12): Implemented a scalable **Deep Learning classification system on GCP, leveraging TensorFlow, OCR, PyTorch, and MLOps**, resulting in a 70% speeding up of claims processing and improving operational efficiency.

- Implemented comprehensive CI/CD pipelines (GitHub Actions) for ML workflows to support reproducibility and continuous delivery of models.
- Restaurant Kitchen Object Detection: Developed real-time **YOLOv5**-based object detection from video streams (PyTorch, AWS Lambda, S3, Step Functions, EC2) for improved kitchen efficiency.
- Led Federated Learning Proof of Concept (CNN-based, **TensorFlow** Lite, C++) for ARM boards.
- Developed and Operationalized NLP and Deep Learning models into production environments utilizing **Flask APIs, AWS Lambda, S3, EC2, and SageMaker**.
- Optimized and quantized heavy TensorFlow models using TensorFlow Lite to ensure low-latency inference for edge and mobile application deployments.
- Configured and managed distributed GPU training infrastructure on AWS to handle large-scale image classification tasks.

Research Assistant | Concordia University, Montreal *July 2019 - December 2020*

- Developed a Windows-based French pronunciation application using **Python, PyQt5, and Google Cloud Speech-to-Text/Text-to-Speech APIs**. Employed MVC design patterns and **Docker** containerization for maintainability and seamless deployment.

Computer Vision Engineer | Bosch India, Bangalore *August 2017 - December 2018*

- **Reduced the object detection model training time of a CNN** for an **autonomous driving project** by 22% without accuracy loss. Developed a Python algorithm to optimize dataset labeling (dense to sparse) for **GPU efficiency**. Trained the model on an **AWS EC2 instance**.
- **Built and trained a person re-identification and tracking system** using a CNN-based Siamese Network (**PyTorch**).

Software Engineer -Image Processing/Deep Learning | L&T Tech Services, Mysore *January 2014 - August 2017*

- Led projects in computer vision application development, including medical image stitching (**OpenCV, C++**) and thermal person detection (**Raspberry Pi, 82% accuracy**).
- Developed 3-class object detection using **transfer learning (VGG16) on NVIDIA Jetson-TK1 (96% accuracy)**.

KEY PROJECTS

RAG Knowledge Assistant (End-to-End MLOps Project)

- Designed a production-grade Retrieval-Augmented Generation (RAG) system adhering to strict MLOps and software engineering best practices.
- **Containerization & APIs:** Built a highly scalable RESTful API using FastAPI, fully containerizing the application and its dependencies using Docker and Docker Compose.
- **Automation:** Implemented a rigorous CI/CD pipeline via GitHub Actions for automated code quality checks (Flake8, Pytest), Docker image builds, and automated publishing to the GitHub Container Registry (GHCR).

Diabetic Retinopathy Classification Platform

- **Infrastructure & Monitoring:** Deployed model training infrastructure on AWS EC2 GPU instances, explicitly utilizing **Grafana** for real-time monitoring of system health, training statistics, and performance metrics.

CERTIFICATIONS

- [Deep Learning Specialization, Coursera](#)
- [Natural Language Processing in TensorFlow, Coursera](#)
- [Tensorflow, Coursera](#)
- [Convolutional Neural Networks with Tensorflow, Coursera](#)

EDUCATION

MEng in Electrical and Computer Engineering

Concordia University • Montreal, QC, CA

Specialization: Computer Vision and ML. Focused on AWS for model builds, training, and scalable deployments.

Bachelor's in Electronics and Communication Engineering

Dr APJ Abdul Kalam University • Ghaziabad, UP, India