

Raof Naushad

+971-502704724 | raofnaushad.7@gmail.com | [Website](#) | [LinkedIn](#) | [GitHub](#) | [Medium](#) | [ResearchGate](#)

PROFILE

AI Solutions Architect with 8+ years of experience designing and delivering enterprise-grade AI systems — multi-agent platforms, secure LLM gateways, Document Intelligence pipelines, and RAG-based products. Lead architect for regulated, compliance-heavy AI solutions spanning data privacy, LLM security, microservices, vector search, and model orchestration. Proven leadership scaling and managing engineering teams to deliver production-ready AI products.

SKILLS

Programming Languages: Python, Go, SQL, Bash, R, IDL, ReactJS, NextJS

Generative AI: LangChain, LangGraph, LlamaIndex, Vector Databases, Azure AI Foundry, MCP, vLLM, Ollama, Hugging Face

LLM Security & Governance: PII/SPII Anonymization, Prompt-Injection Defense, Data Classification & Model Routing, RBAC, Guardrails, Grounding Evaluation

Open / Frontier Models: Anthropic, OpenAI, Qwen, Kimi, DeepSeek, Llama, Mistral

Machine Learning & Data: PyTorch, TensorFlow, NumPy, Pandas, Matplotlib, NLTK, spaCy, Scikit-Learn

Databases: Postgres (pgvector), MongoDB, Pinecone, RDS, MySQL

MLOps & Cloud: MLflow, Docker, FastAPI, Flask, Nginx, Gunicorn, GitHub Actions, Azure (Functions, Service Bus, Event Grid, Container Apps), AWS, Kubernetes

Concepts: Retrieval Augmented Generation (RAG), Multi-Agent Workflows, Fine-tuning, Prompt Engineering, Distributed Systems, Event-driven Pipelines, Stream Processing (Kafka, Spark), Micro-services, CI/CD, Agile

WORK EXPERIENCE

AI Architect | Full-time

Sep 2025 – Present

AsbiTech (Watar Partners) – Dubai, UAE

- **Lead a 9+ member AI engineering team** building the AI Core (Document Intelligence, RAG Gateway, Agentic Workflows, MCPs) of a wealth-management platform serving family offices and institutions across the GCC (UAE, KSA, Kuwait).
- **Architected a secure LLM Gateway** providing PII/SPII anonymization, prompt-injection and jailbreak defense, and policy-driven model routing that selects LLMs based on invoked tools and data classification — meeting the compliance and regulatory demands of GCC wealth management.
- **Built a Claude-style agentic assistant** operating entirely over private, internal MCP tools across our module suite — able to create and edit documents/reports, generate dynamic dashboards, and schedule tasks in a fully private and secure environment.
- **Implemented fine-grained RBAC** across all agentic use cases, scoping data retrieval and tool access to the identity and entitlements of the requesting user.
- **Deployed latest open-source models (Qwen, Kimi, DeepSeek)** for internal use cases, enabling private VNET/on-prem inference and reducing dependency and cost on external providers.
- Architected Azure-native pipelines using Functions, Service Bus, Event Grid, AI Search, and Document Intelligence; orchestrated models via Azure AI Foundry (OpenAI, Anthropic, Qwen) with streaming, retries, and dynamic routing.
- Built a vector-first hybrid RAG system with optimized chunking, re-ranking, metadata strategy, and grounding evaluations; designed multi-agent reasoning workflows (LangGraph) for financial insights, compliance summarization, and memo generation.
- Owned platform cost, reliability, and security (VNETs, Private Endpoints, Key Vault) with IaC across dev → staging → prod, and defined the product AI roadmap and engineering standards with the CEO and product lead.

Data & AI Architect | Full-time

Nov 2023 – Sep 2025

Royal Bank of Canada (RBC) - Toronto, Canada

- **Patent-pending GenAI solution:** invented and led a GenAI-based Personal Statement generator for Personal Banking users, producing narrative financial insights from client data — filed as a patent-pending solution at RBC.
- Implemented the SafeAI Adoption Process, digitizing the SafeAI onboarding framework to ensure compliance with enterprise architecture, security, and 2LoD risk approvals.
- Developed an AI Onboarding Pattern to standardize AI deployment, starting with a private LLM inference pattern for secure, efficient model hosting.
- Conducted data mining on 5M+ client records, enhancing LLMs with prompt engineering and RAG (chain-of-thought, RAPTOR), reducing pipeline processing time by 23%.
- Designed an end-to-end MLOps pipeline with MLflow for experiment tracking, model versioning, and deployment; CI/CD via GitHub Actions, cloud deployment on AWS, agile tracking in JIRA.
- Built a PoC with open-source models (Hermes-2-Pro-Llama-3-8B, GritLM-7B), fine-tuning Meta Llama 3 8B for function calling to replace an OpenAI-based client application — a potential 96.3% reduction in price per score.

Senior Analyst (Data Science) | Full-time

Nov 2022 – August 2023

Factspan Analytics - Bangalore, India

- Developed PySpark data-loading functions on Azure Databricks to load and transform 5M daily incremental records, significantly improving processing efficiency.
- Used 86+ complex SQL queries to validate business logic and resolve source-data discrepancies; analyzed data across four billing systems to shape the billing data architecture.
- Automated the testing framework via a scheduled PySpark Databricks notebook for continuous monitoring and alerting, saving 10% in resource time. Received a spot award for innovation.

Data & ML Engineer | Full-time

Dec 2019 – Aug 2022

UST Global - Kerala, India

- Built an intent and entity recognition NLP system (BERT) to automate the client's IVR system, reducing manual labor by 8%.
- Deployed deep learning models on AWS (GPU/CPU) with server provisioning for retraining to ensure performance and scalability.
- Designed a scalable data engineering pipeline on AWS using Step Functions and Lambda; optimized the Python codebase (Pandas, multiprocessing, multithreading) for a 60% reduction in data load time. Received the UStar Award.

EDUCATION**Masters of Computing & Data Analytics | 4.2/4.3 GPA**

September 2023 – December 2024

*Saint Mary's University, Halifax, NS***Bachelor of Technology - Electronics and Computer Engineering | 8.64/10 GPA**

May 2015 – July 2019

*APJ Abdul Kalam Technological University, Kerala, India***Research & Publications**

- Naushad, R.; Kaur, T.; Ghaderpour, E. Deep Transfer Learning for Land Use and Land Cover Classification: A Comparative Study. *Sensors* 2021, 21, 8083. <https://doi.org/10.3390/s21238083>
- Raoof Naushad, M. Harshita Vijayam, A. Mohammed Hassan, T. Safwan Rashid, S. Asha, "Plastic waste segregation using Raman spectroscopy," AIP Conference Proceeding; <https://doi.org/10.1063/5.0004262>
- [Open Source Contributor - LlamaIndex](#)