

Rohit Raj AI Engineer

📍 Bengaluru, India

☎ 8709287493

🌐 <https://github.com/RoCode8>

✉ rohitraj0830@gmail.com

🌐 www.linkedin.com/in/rohitraj9805

AI solutions developer with more than 3 years of experience spanning Generative AI, RAG, fine-tuning, machine learning, data engineering, cloud architecture, and solution architecture. Proven track record of leading enterprise chatbot and AI projects across diverse domains, from design to deployment. Skilled in leveraging major cloud platforms, including Azure (AI Studio, ML, Data-Bricks) and Google Cloud (Workbench, Vector AI, Cloud Run, BigQuery, Cloud Storage), to build scalable and secure solutions. Proficient in Python, PySpark, and SQL for developing robust data and AI pipelines. Adept at cross-team collaboration, delivering strategic recommendations, and optimizing performance and security. Passionate about applying cutting-edge AI/ML and LLM technologies to drive innovation and solve complex business challenges.

🧠 Skills

Gen AI: Proficient in developing Gen AI applications, including building and optimizing RAG systems with document processing pipelines, vector store implementation, and retrieval optimization. Experienced in fine-tuning open-source language models for specific use cases, crafting efficient prompt engineering strategies, and leveraging advanced image generation models like SDXL, ControlNet, LoRA, and Karras Samplers for image synthesis.

Cloud: Proficient in **Azure Cloud** (AI/ML Studio, Azure OpenAI Service, Cognitive Services, Functions, ADF, Databricks, ADLS) and **Google Cloud Platform** (Vertex AI Workbench, Vector Search, Cloud Run, BigQuery, Cloud Storage) for building, deploying, and managing scalable AI solutions and data pipelines.

DevOps: Experienced in **Docker** and implementing CI/CD pipelines with **Azure DevOps**, **GitHub Actions**, and **Google Cloud Build**. Skilled in managing deployments using Azure Kubernetes Service (AKS) and Google Kubernetes Engine (GKE). Proficient in managing Azure resources with **Azure CLI** and ARM/Bicep templates, as well as Google Cloud resources with **gcloud CLI** and Deployment Manager. Expertise in GitHub, version control practices, and automation workflows across multi-cloud environments.

Programming Languages, Tools and Frameworks: Python | Langchain | Langsmith | Langflow | Promptflow | Databricks | Microsoft Fabric | MySQL | Pytorch | TensorFlow

📁 Work Experience

AI Engineer, Prudential Assurance

05/2025 – Present | Remote

Critical Illness Claims Adjudication Prototype – Developed an AI-driven prototype to automate the adjudication of critical illness claims (heart attack, kidney failure, cancer) for an insurance provider. Designed a three-stage agentic workflow: (1) **Extraction Engine** to parse medical claim documents from cloud storage, extracting key-value pairs and summaries; (2) **Assessor Engine** to evaluate claims against SOPs and extracted data, generating AI rationale and classification into approved, rejected, or manual review; and (3) **Re-Evaluator Engine** to verify assessment accuracy before final decision. Leveraged cloud storage, document processing, LLM prompting, and multi-agent orchestration to streamline and accelerate the claims process.

Delphi Consulting

07/2022 – 05/2025

Enterprise Chatbot Development - Client : Dubai Holding Real Estate.

- Designed and deployed an advanced multilingual Enterprise Chatbot leveraging Microsoft's Cognitive Search and Azure OpenAI, specializing in prompt engineering and fine-tuning for HR and procurement policy queries.
- Seamlessly integrated the chatbot into the client's infrastructure, collaborating with Microsoft Power Platform developers to enhance functionality for employee self-service tasks, including streamlined pantry ordering.

- Automated SharePoint document indexing through Azure Functions with blob triggers, improving document accessibility and operational efficiency.
- Enhanced system security by implementing private endpoints for Azure Storage and optimizing IP configurations for Azure Web App services.
- Recommended and deployed a cost-effective Azure App Service plan tailored to client needs, ensuring scalability and performance.
- Delivered measurable outcomes, including improved employee engagement, reduced dependency on HR and procurement teams, and increased productivity through task automation.

AI-Powered Cold-Calling Recruitment Assistant - Client : Shobha Realty

- Designed and implemented an AI-driven voice assistant leveraging Twilio API and OpenAI for natural language conversations to automate candidate outreach and engagement.
- Integrated VAPI platform for call flow management, incorporating NLP for advanced speech-to-text and text-to-speech capabilities, enabling seamless real-time transcription.
- Developed robust REST API and webhook integrations for data exchange with existing HR systems, ensuring smooth process alignment and automation.
- Built a comprehensive analytics and reporting module, capturing call logs, summaries, and key metrics to optimize recruitment strategies and monitor performance.
- Delivered measurable outcomes, including increased recruitment efficiency, enhanced candidate engagement, and reduced hiring costs by automating initial screening and minimizing reliance on third-party recruiting services.

PoCs & Projects

AI Powered Product Validation Web Application, Client : Al futtaim

- Developed AI-powered product validation application using Python, Streamlit, and Azure OpenAI GPT-4 Vision API for e-commerce verification processes.
- Implemented advanced image analysis and comparison with user-provided descriptions, generating detailed validation reports.
- Designed a flexible product categorization system supporting multiple categories (Electronics, Clothing, Grocery) with specific validation parameters.
- Engineered AI algorithms for the automated generation of SEO-optimized product descriptions, titles, and tags.
- Created a user-friendly web interface using Streamlit for efficient product detail input and image upload.
- Deployed application in Azure container using Docker, ensuring scalability and reliability.

AI-Powered Service Request Automation, Client : WASL

- Led development of two AI prototypes using Azure OpenAI GPT-4 Vision, GPT-3.5, and Azure Vision API, improving image categorisation accuracy from 70% to 95%.
- Integrated optimized solution into client's application, demonstrating significant performance enhancement.
- Implemented seamless photo upload and camera capture functionalities, collaborating with Flutter App developer.
- Engineered AI algorithms for automated categorisation of service requests and generation of relevant descriptions.
- Streamlined request process, reducing user effort and submission time while improving service quality through enhanced accuracy in classification and description generation.

Education

Bachelor of Engineering(B.E) in ECE,

BMS Institute of Technology and Management 

Electronics and Communication Engineering

2021 | Bangalore, India

Certificate

Microsoft Certified: Azure AI Engineer, *Certificate Link* 

08/12/2024

AI Engineer: Certificate of Top 5% professional by Pro5.ai,
Certificate Link - 

06/2024