DEVENDRA CHAUDHARI

+61 481783399 | devachaudhari22@gmail.com | linkedin.com/in/devendra-chaudhari-40a79321a | github.com/DevaChaudhari | Portfolio

Professional Summary

AI/ML Engineer with 2+ years building and shipping generative AI, agentic AI, and RAG systems to production. Hands-on expertise in LangChain, LangGraph, MCP, A2A, FastAPI, Docker, Kubernetes, and Google Cloud Run. Published researcher with a track record of delivering autonomous multi-agent platforms from prototype to cloud deployment.

Education

Flinders University — Expected Dec 2027
Master of Data Science — *Adelaide, Australia*

Savitribai Phule Pune University — Feb 2021 – Jun 2024
Bachelor of Engineering in Computer Science | *CGPA: 9.13/10.00* — *Pune, India*

Work Experience

AI Developer Intern — Sep 2026 – Dec 2026
ExpertEase AI | *Adelaide, Australia*

- Building Chat and Voice AI Agents integrated with business systems through APIs, webhooks, and retrieval augmented generation for workflow automation.
- Implemented an autonomous operations agent harness using Hermes Agent, and benchmarked self hosted open weight LLMs for latency and response quality.
- Evaluating and improving RAG and chat quality across retrieval, embeddings, vector search, chunking strategy, context selection, prompting, and grounding to reduce hallucination in conversational agents.
- Supporting Voice AI and telephony integrations using ElevenLabs, Vapi, Twilio, and Vonage, covering the end to end call flow from telephony webhooks through the LLM and tools layer to voice response generation.
- Using Langfuse for LLM observability, tracing, and evaluation to identify and fix quality and performance issues in conversational agents.
- Debugging and resolving platform issues within a Git branch based team workflow, and building workflow automation with n8n.

AI Engineer — Jul 2024 – July 2026
Freelance | *Remote*

- Developed and deployed Generative AI applications including RAG chatbots, autonomous multi-agent systems, and AI-powered automation solutions using LangChain, LangGraph, FastAPI, and Python, reducing average client workflow time by about 35%.
- Engineered intelligent agent workflows with Model Context Protocol (MCP) and Agent-to-Agent (A2A) communication, cutting cross-agent integration time by about 50% compared to custom REST-based approaches.
- Built scalable AI platforms integrating vector databases, embeddings, semantic search, and web tools to support real-time information retrieval and task execution across 4+ client deployments.
- Designed production-ready backend services and REST APIs for AI applications, achieving $<$200ms median response latency across cloud environments.
- Deployed and monitored AI solutions using Docker, Microsoft Azure, GitHub, and LangSmith, reducing debugging and observability overhead by about 40% through structured tracing pipelines.

Associate Engineer — Dec 2024 – Mar 2025
L & T Technology Services | *India (concurrent with freelance AI Engineer role)*

- Performed as an AI Developer across LLMs, RAG, Python, NLP, generative AI, and agentic AI, while engineering enterprise-grade cloud applications on Microsoft Azure – including Azure Virtual Machines, Azure App Services, Azure SQL Database, and Azure AI services – for scalable production deployment.
- Architected a scalable Data Access Layer using Entity Framework Core with Database-First design, reducing backend data retrieval time by about 30% and improving system maintainability across 3+ modules.
- Optimized SQL Server performance by designing stored procedures, complex queries, and indexed database schemas, cutting average query execution time by about 45% on high-volume transactional data.
- Automated build and deployment processes through MLOps-aligned CI/CD pipelines in Azure DevOps, reducing manual deployment effort by about 4 hours per release cycle and enabling continuous integration across development teams.
- Completed advanced Microsoft Azure training encompassing Generative AI, NLP, Machine Learning, Computer Vision, and cloud-native AI application development.

Machine Learning Intern — Feb 2023 – Mar 2023
E-Cell IIT Hyderabad | *India*

- Built and evaluated machine learning models using Python, TensorFlow, and Scikit-learn to predict seasonal flu vaccine uptake, achieving a ROC-AUC of 0.84 on real-world healthcare datasets.
- Implemented and benchmarked Logistic Regression, Random Forest, and ensemble learning algorithms, selecting the best-performing model through systematic cross-validation and hyperparameter tuning.
- Applied end-to-end data preprocessing including feature selection, normalization, and class imbalance handling, improving model generalization F1 score by about 12% on skewed healthcare data.
- Evaluated model performance using confusion matrix analysis, ROC-AUC curves, precision-recall metrics, and F1 scores, ensuring rigorous and reproducible model assessment.

Projects

AustraliaJobSeeker AI – End-to-End AI Agent Job Platform Mar 2026 – Jun 2026

Python, LangGraph, LangSmith, Groq LLM, FastAPI, MCP, A2A, Docker, Kubernetes, Google Cloud Run, CI/CD

- **Live App** | **GitHub Repository**
- **Real-World Problem Solved:** Seek.com has no visa filter for international students – hours are wasted applying to citizen-only jobs. Built a full production AI agent system that scrapes live Seek jobs and filters them by visa type (500, 485, 482, PR) in real time.
- Architected a LangGraph multi-agent pipeline covering real-time job scraping, job enrichment, visa eligibility checking, resume match ranking, and AI-powered resume and cover letter generation in a single automated workflow.
- Built custom MCP (Model Context Protocol) servers exposing job search, visa check, resume tailoring, and cover letter generation as callable tools for any AI agent, and A2A (Agent-to-Agent) servers enabling true multi-agent collaboration and skill discovery across the system.
- Integrated LangSmith for real-time agent observability and tracing, providing full visibility into every agent decision, LLM call, and pipeline execution in production.
- Deployed on Google Cloud Run with Docker containerization, Kubernetes manifests for production-grade orchestration, and a CI/CD pipeline via Google Cloud Build for fully automated builds and deployments on every git push, migrating from Ollama to Groq for cloud-scale inference across an iterative Streamlit to Render to Cloud Run deployment path.

ProjectPilot-AI – Autonomous Multi-Agent System Mar 2026

Python, LangGraph, LangChain, FastAPI, Streamlit, Ollama, REST APIs

- **Live App** | **GitHub Repository**
- Architected a production-grade autonomous multi-agent AI platform using LangGraph, enabling dynamic task planning, multi-step LLM reasoning, and real-world automation across 5+ integrated tool agents.
- Engineered a modular AI agent orchestration framework with dedicated planner, executor, and tool-selection agents, streamlining LLM decision-making pipelines and improving end-to-end task orchestration efficiency by about 40%.
- Integrated 6+ external APIs and services – including GitHub API, REST APIs, file management systems, and web browsing tools – enabling fully automated workflow execution without human intervention.
- Deployed cost-efficient local LLM inference using Ollama, eliminating cloud API costs for agent reasoning while maintaining response quality comparable to hosted LLM providers.
- Built scalable RESTful backend services with FastAPI and an interactive real-time frontend with Streamlit, supporting concurrent agent execution and live task monitoring.

YouTube RAG Chatbot Jan 2026

Python, LangChain, Gemini AI, FAISS, Streamlit, Docker, Kubernetes

- **Live App** | **GitHub Repository**
- Built and deployed an end-to-end RAG chatbot enabling users to query YouTube video content in natural language using LangChain and Google Gemini LLM.
- Designed a 4-stage RAG pipeline covering video metadata extraction, transcript loading, FAISS vector indexing, and context-aware answer generation, achieving sub-second retrieval latency.
- Generated high-dimensional transcript embeddings using Google Gemini and indexed them in FAISS, enabling semantic vector search across 100K+ token transcripts with $<$500ms query time.
- Dockerized the full application and authored Kubernetes deployment and service manifests, ensuring cloud-native scalability and production deployment readiness.

Research Publication

Book Chapter: "A Real-Time Vehicle Counting System for Detecting Traffic Surveillance" – *Advancing Engineering and Technology through Interdisciplinary Approaches: A Pragmatic Approach*, Cambridge Scholars Publishing, 2026. Presented real-time deep learning-based vehicle detection, classification, and congestion analysis using YOLOv3, OpenCV, and Python for scalable smart traffic management.
Conference: ICATIA-24 – International Conference on Advanced Technologies and Intelligent Automation, May 2024. Presented research on computer vision and intelligent automation systems to an interdisciplinary engineering audience.

Technical Skills

Programming Languages: Python, SQL
AI/ML Technologies: Machine Learning, Deep Learning, Generative AI, Agentic AI, NLP, Computer Vision, LLMs, RAG Pipelines, Prompt Engineering, Fine-Tuning, Embeddings, Vector Search, AI Agent Orchestration, MLOps, Transfer Learning
Frameworks & Libraries: LangChain, LangGraph, LangSmith, TensorFlow, PyTorch, Scikit-learn, Hugging Face Transformers, FastAPI, Flask, React, Pandas, NumPy, OpenCV
Databases: SQL Server, MySQL, FAISS, Pinecone, ChromaDB, Vector Databases
Cloud & DevOps: Microsoft Azure, Azure AI Foundry, Azure DevOps, AWS (Amplify, ECR, ECS, ALB, CloudFront, IAM), Google Cloud Run, Docker, Kubernetes, CI/CD Pipelines
Developer Tools: GitHub, GitLab, VS Code, PyCharm, Jupyter Notebook
Concepts: MCP Servers, A2A Agents, Agent Harness (Hermes Agent), REST APIs, Microservices Architecture, Model Deployment, Feature Engineering, Hyperparameter Tuning, Model Evaluation

Certifications

Microsoft Certified: Azure AI Engineer Associate (AI-102) – Microsoft – Aug 2025
Microsoft Certified: Azure AI Fundamentals (AI-900) – Microsoft – Aug 2025
Machine Learning with Python – E-Cell IIT Hyderabad – Mar 2023
Problem Solving and Ideation Workshop – Entrepreneurship Development Startup & Innovation Cell – Nov 2021