Harikrishna R Nair

+971501563096 harirnair04@gmail.com
linkedin.com/in/harirnair04
Location: Abu Dhabi, UAE **Nationality:** Indian

SUMMARY

AI Engineer with 5+ years of experience designing and developing **Generative AI and Agentic AI solutions**, building scalable **backend services** using **Python, FastAPI, and Flask**. Skilled in **LLMs, RAG, multi-agent systems, AI evaluation, and AWS cloud services**. Proven ability to deliver **production-grade enterprise solutions** across domains including loan processing, intelligent data extraction, and AI-powered test automation, leveraging modern **agentic AI frameworks** and cloud-native architectures.

TECHNICAL SKILLS

- **GenAI & Agent Frameworks:** Amazon Bedrock Agent Core, Strands Agent, Claude Sonnet 4, GPT 4, CrewAI, Pydantic AI, Multi-agent orchestration, Supervisor agent pattern, OpenTelemetry tracing
- **Vector & Knowledge Systems:** Amazon Titan embeddings, Semantic search, RAG (Retrieval-Augmented Generation), Vector databases, S3 Vectors, Cosine similarity, Knowledge base integration
- **Languages & Frameworks:** Python, FastAPI, Flask, TypeScript, React.js, REST/GraphQL APIs, AWS CDK (Infrastructure as Code)
- **Cloud & AWS Services:** AWS Bedrock, Lambda, EKS (Kubernetes), DynamoDB, S3, Cognito, AppSync, CloudFront, WAF, ECR, EventBridge, Secrets Manager
- **DevOps & Infrastructure:** Docker, Kubernetes (EKS Auto Mode), AWS CDK, CI/CD (CodeBuild/CodePipeline), Multi-tenant SaaS architecture, Distributed tracing (OpenTelemetry)
- **AI/ML Integration:** Azure OpenAI, Azure Form Recognizer, SERP API, Tavily API, JIRA API, Multi-source data integration, Context-aware generation

EDUCATION

Sastra University **May 2022– May 2024**
Master of Computer Applications,

University of Calicut **May 2018– May 2021**
Bachelor of Computer Applications,

WORK EXPERIENCE

Tata Consultancy Services **Oct 2021– Present**
Software Engineer

- Developed **AI-powered automation systems** and backend services using **Python**, **FastAPI**, and **Flask**.
- Integrated **GenAI frameworks** (**CrewAI**, **Pydantic AI**, **Bedrock**, **Strands**) into enterprise workflows.
- Built cloud-native **multi-tenancy solutions** on **AWS** with **S3**, **DynamoDB**, and **Lambda**.
- Collaborated with cross-functional teams to improve scalability, automation, and developer efficiency.

PROJECTS

Agentic Test Orchestrator
Tech Stack: *AWS Bedrock Agent Core, Strands Agent Framework, Claude Sonnet 4, Flask, EKS Auto Mode, Cognito, DynamoDB, Multi-tenant SaaS*

- Architected enterprise **multi-tenant SaaS platform** using **AWS Bedrock Agent Core** with **Strands Agent Framework** and **Claude Sonnet 4** for intelligent test automation across isolated tenant environments.
- Engineered **supervisor agent pattern** with specialized tools for scenario generation, Gherkin syntax conversion, and Java/Selenium code generation, integrated with **OpenTelemetry** for distributed tracing.
- Built production-grade **Flask backend** with modular architecture (routes, services, managers) implementing **JWT authentication**, header-based tenant routing, and secure multi-tenancy with **DynamoDB** and **S3** storage.
- Deployed scalable infrastructure using **AWS CDK** with **EKS Auto Mode** (Kubernetes 1.33), **AppSync GraphQL API**, **Cognito** user pools, and **CodeBuild/CodePipeline** for automated tenant onboarding and CI/CD.
- Implemented **vector-based knowledge retrieval** using **Amazon Titan embeddings** with Lambda-powered semantic search (cosine similarity) stored in **S3 Vectors** for context-aware test generation.
- Secured platform with **AWS WAF**, comprehensive security headers, **CloudFront CDN**, TLS 1.2 encryption, rate limiting (2000 req/5min), and **AWS Secrets Manager** for credential management.

- Integrated **JIRA API workflows** for automated user story extraction via Lambda functions and created test execution engine with **Bedrock AgentCore runtime** for workflow orchestration (85% faster test generation).

LoanAgent - AI-Powered Loan Processing System
Tech Stack: *FastAPI, CrewAI, Azure OpenAI, PyMuPDF, Excel automation, Real-time streaming*

- Built and deployed backend services using **FastAPI** and **CrewAI agents** for automated loan notice processing with real-time data extraction.
- Implemented enhanced **PDF extraction** using **Azure AI Form Recognizer** and coordinate-based highlighting for 40% faster manual QA workflows.
- Integrated **SOP-based LiveLink agent** to handle duplicate detection and workflow management, reducing processing time by 60%.
- Designed **real-time streaming APIs** with FastAPI for live validation updates and multi-tenancy support across loan portfolios.
- Automated **Excel workflows** with LoanIQ agent for intelligent validation using multi-agent coordination.
- Enhanced performance with **Azure OpenAI integration** and automated deduplication reducing manual effort by 70%.

Research Assistant - Intelligent Web Research Automation
Tech Stack: *Strands Agent, Azure OpenAI, Tavily API, MCP (Model Context Protocol), Async processing*

- Created a **Strands Agent-powered** assistant using **MCP** for automated research and content synthesis.
- Implemented **Tavily API** integration for real-time web scraping and information gathering.
- Built **async content processing** with **Azure OpenAI** for 50K token summarization.
- Designed **intelligent query generation** strategies for comprehensive exploration.
- Developed **streaming response APIs** with FastAPI and **SSE** for real-time updates.
- Automated **report generation** with citation tracking and validation.

Info Retrieval Agent - Multi-Source Data Extraction System
Tech Stack: *Pydantic AI, FastAPI, WebSocket, Multi-agent coordination, RAG, Vector search*

- Developed a **multi-agent system** using **Pydantic AI** framework with Excel, Image, PDF, and SQL specialists for comprehensive analysis.
- Built **real-time log streaming** with **WebSocket** connections for transparent processing and contextual logging.
- Integrated **LLMs** with multi-tenancy support and automated research workflows reducing manual retrieval by 80%.
- Implemented **RAG (Retrieval-Augmented Generation)** with **FAISS** for enhanced information accuracy.
- Designed **contextual logging** and a **React TypeScript frontend**.
- Automated extraction from **structured/unstructured data** including databases, images, and documents.

AR Assembly Guide - Immersive Training Solution
Tech Stack: *Unity3D, AR Foundation, Real-time object tracking, Animation systems*

- Created an **AR-based assembly guide** system scanning products and overlaying animated instructions.
- Improved user experience and reduced setup errors by **90%**.
- Implemented **real-time object tracking** with **Unity3D**.
- Enhanced engagement with immersive **animations** and responsive design.

CERTIFICATIONS

- AWS Certified AI Practitioner
- Google Cloud Platform Associate Cloud Engineer
- Anthropic - Model Context Protocol
- Microsoft Power Platform Developer (PL-400)