AYUSH KHAMRUI

Bangalore, Karnataka | khamruiasok@gmail.com | Ph- 7001477696 | https://www.linkedin.com/in/ayush-khamrui/
Portfolio - https://portfolio-ayush-khamruis-projects.vercel.app/

Summary

Seasoned **AI/ML professional** with 3+ years of hands-on experience architecting and delivering production-grade Generative AI solutions for enterprise clients. Proven expertise in designing Agentic RAG networks, multi-agent orchestration frameworks, and LLM-driven chatbots on AWS (Bedrock, Lambda, Kendra), Azure OpenAI, and GCP (VertexAI). Adept at modernizing architectures through scalable microservices, Docker/Kubernetes deployments, and enterprise observability (Prometheus, Grafana, Splunk), achieving 99.99% uptime. Strong track record of leading cross-functional teams, driving prompt-engineering best practices, and automating AI pipelines to boost reliability by 30%. Passionate about translating cutting-edge research into robust GenAI products, optimizing retrieval, memory, and reasoning layers for next-generation solution architectures.

WORK EXPERIENCE

Senior Software Engineer (AI/ML) | Optum (United Health Group) — Apr 2025 - Present

- **Pioneered the architecture and development of an advanced Agentic AI-driven chatbot application,** seamlessly interfacing with legacy in-house databases via GraphQL to deliver dynamic, real-time statistical insights based on user-selected agents.
- **Engineered a bespoke Agentic RAG (Retrieval-Augmented Generation) network,** dramatically enhancing intelligent, context-aware response generation with customized memory, reasoning, and retrieval optimization layers.
- **Designed and built a Supervisor Agentic Framework using LangGraph, enabling dynamic orchestration, monitoring, and error correction across** 5+ autonomous AI agents, significantly boosting operational efficiency and collaborative task execution.
- **Developed a fully automated Agent Evaluation System by integrating VertexAI, RAGAS, and Playwright,** empowering continuous performance benchmarking, regression testing, and self-healing AI pipelines — driving a 30% improvement in agent reliability and response quality.
- **Spearheaded the backend architecture modernization** by leading the migration to a scalable microservices ecosystem, ensuring modular AI service development, rapid iteration cycles, and horizontal scalability.
- **Mastered cloud-native deployment** using Docker and Kubernetes to build highly resilient, autoscaling AI infrastructure, achieving 99.99% uptime and seamless cross-region failovers.
- **Established enterprise-grade observability** with Prometheus, Grafana, and Splunk, creating proactive monitoring, alerting, and intelligent incident analysis pipelines, accelerating issue resolution and optimizing system performance.
- **Championing innovation and operational excellence**, consistently pushing the boundaries of AI-driven architectures, model evaluation frameworks, and intelligent agent orchestration to deliver business-critical solutions at scale.

Senior Software Engineer | Capgemini — Mar 2022 – Apr 2025

- **Led a 5-member team in designing and architecting complex LLM systems (AI Agents, RAG & Fine-Tuning)** across AWS Bedrock, Azure OpenAI, and Hugging Face using Langchain, CrewAI; co-owned the team's technology vision, contributing significantly to its success.
- **Optimized LLM retrieval by employing various retrievers, prompt compression techniques, and guardrails**; crafted efficient prompts using context and zero-shot, one-shot and few-shot techniques; and applied advanced algorithms and data structures with a keen understanding of space and time complexities.
- **Enhanced code quality and maintainability** by applying object-oriented programming principles and design patterns to refactor, optimize, and debug code.
- **Developed efficient, maintainable code in Python , React and JavaScript (ES6+)**, adhering to industry coding standards to solve complex business problems.
- **Integrated Neo4j and ChromaDB as knowledge bases**, working extensively on Linux/Unix platforms for deployment and application management.
- **Created multiple APIs using Python Flask and FastAPI** integrating all services (Query, UML diagram generation, RAGAS Score, Carbon footprint calculation) to deliver robust backend solutions.
- **Integrated Azure Cache for Redis for caching** LLM responses and using it for faster retrieval enhancing application performance.
- **Followed SDLC in an agile environment**, collaborating with cross-functional teams to ensure on-time deliveries.
- **Developed and evaluated LLMs using RAGAS scores and MLflow**, demonstrating strong technical aptitude and solid computer science fundamentals.
- **Led frontend development of a GenAI Chatbot using JavaScript (ES6+), React and Python**, integrating AWS Bedrock LLMs using RAG.
- **Designed data models and low-level classes** for efficient chatbot data handling and processing.
- **Utilized object-oriented programming and design patterns** to develop scalable, maintainable codebase.
- **Integrated 5+ REST APIs** into the chatbot for user authentication, bot interactions, feedback, analytics, and export features.
- **Applied SOLID and DRY principles to refactor code**, improving readability and increasing efficiency by 12%.
- **Drove best practices** by actively participating in code reviews, design reviews, and architecture discussions.
- **Experimented with LangChain and Azure OpenAI**, promoting their adoption and measuring impact on project outcomes.
- **Led a 3-member team on an AWS GenAI pilot project**, using JavaScript, React and Python, achieving 30% GenAI adoption.

Technical SKILLS

- *Programming Languages:* **JavaScript**, **Python**
- *Front-end Technologies:* **React**
- *Back-end Technologies:* Flask, FastAPI
- *Generative AI / AI Tools:* **Amazon Q**, **Tabnine**, **Github Copilot**
- *AI/ML Libraries and Frameworks:* Numpy, Pandas, PyTorch, scikit-learn, TensorFlow
- *AWS services:* S3, AWS Lambda, AWS Textract, **AWS Bedrock**, DynamoDB, Route53, Cognito, AWS Kendra, AWS Step Function
- *Azure services:* **Azure OpenAI**, Storage accounts, Document intelligences, AI Search
- *GCP Services:* ***VertexAI****, BigQuery, Firestore, Google Kubernetes Engine*
- *Version control:* Git, Github, AWS CodeCommit
- *CI/CD:* Docker, Kubernetes, Github Actions
- *Logging/Monitoring Tools* – Grafana, Splunk

EDUCATION

Electronics and Communication Engineering | JIS College of Engineering **Jun 2018 - Jun 2022**

GPA: 9.17

- Published **3 Research Papers (1 in the Chemical domain and 2 in the IoT Domain)**
- Patent publication '*A CAR IGNITION SYSTEM BASED ON APPLICATION CONTROLLED BIOMETRIC SENSOR'* **(PATENT APPLICATION NO: 202031031645 A)**

AWARDS

- Awarded as the "**Star**" employee by Capgemini in July - September 2023.
- Awarded as the "**Rising Star**" employee by Capgemini in April - June 2023.
- Ranked **top-10** in AWS GenAI hackathon conducted by AWS across **35+ premium partners** by developing a Taxation UI in **JavaScript(ES6+), HTML, CSS, React, Bootstrap** within a record **2 hours** from scratch using **Amazon CodeWhisperer**.

LICENCES AND CERTIFICATIONS

- Certified **AWS Cloud Practitioner** (License No: HDZQ7MWCRJREQVW7, Issued by AWS)
- Capgemini Certified **Full Stack Developer - Level 1 & 2** (Issued by Capgemini)
- Certified **Mathematics-Basics to Advanced for Data Science and GenAI** (Udemy)
- Certified **Machine Learning with Python** (Coursera)