VAISHNAVI PATEL
Automation Engineer | Gen AI / Agentic AI Lead | Data Scientist
USA| Permanent Resident
vaishnavikrishnakantt@gmail.com | (203)-215-1561

SUMMARY

Generative AI Engineer and Agentic AI Lead with 6+ years of experience designing, fine-tuning, and deploying multi-agent AI systems, RAG pipelines, and autonomous workflows using LangGraph, CrewAI, and AutoGen across AWS, Azure, and GCP environments in regulated industries (legal tech, fintech, EV infrastructure). Skilled in building RAG pipelines (LangChain, AWS Kendra, Pinecone), fine-tuning LLMs (Llama-2, Mistral, GPT-4) and small language models (SLMs) using LoRA/PEFT, pruning, and teacher–student distillation. Proven track record in scaling GenAI systems from MVP to enterprise deployment across AWS, Azure, and GCP. Adept at architecting agentic frameworks for multi-step reasoning while ensuring Responsible AI principles—fairness, transparency, and compliance—are embedded into every solution.

KEY SKILLS

Programming & Frameworks
Python, Node.js, FastAPI, Flask, TypeScript, React.js / Next.js, **AngularJS (basic)**, LangChain, LangGraph, AutoGen, Crew AI, **GitFlow, PR Workflows**

AI & Machine Learning
Hugging Face Transformers, PyTorch, TensorFlow, LoRA / QLoRA / PEFT, **Plugin & Tool Calling**, **Agent Memory Management**, **Task Planning**, **Evaluation Metrics (RAGAS, DeepEval, FMeval)**, RAG Pipelines, Prompt Engineering, Observability, Bias Mitigation

Databases & Vector Stores
FAISS, Pinecone, Weaviate, **CosmosDB**, **MongoDB**, Azure Cognitive Search, PostgreSQL, MySQL, AWS Redshift

Cloud Platforms
AWS (Bedrock, SageMaker, Lambda, Glue, Redshift, S3), Azure (OpenAI, Monitor, Cognitive Search), GCP (Vertex AI)

MLOps & DevOps
LLMOps (prompt versioning, caching, monitoring, cost tracking), Docker, GitHub Actions, Jenkins, Terraform, **ETL & Data Pipelines (batch and streaming)**, **GitFlow**, **Code Reviews**, **Branching Strategies**, CI/CD Pipelines

APIs & Integrations
REST, GraphQL, Salesforce, n8n, Power Automate, Zapier

Data & Visualization
Tableau, Power BI, Data Reporting, Real-Time Analytics

Business & Leadership
AI Governance, Model Safety, Requirement Gathering, Project Scoping, Market Intelligence, Stakeholder Management, Strategic Planning, Agile / Scrum, Quality Assurance, Customer-Centric Analytics

PROFESSIONAL EXPERIENCE

Chaffin Luhana LLP – USA
AI Technical Lead (GenAI Architect)   *April 2025 – Present*

- **Spearheading the design and deployment of GenAI-powered legal automation solutions to optimize personal injury case intake, document review, and client communications, reducing processing time by 45%.**

- **Designed and deployed GenAI pipelines with human-in-the-loop feedback cycles to ensure accuracy, compliance, and continuous model improvement.**

- **Built scalable AI solutions aligned with Responsible AI principles, ensuring fairness, transparency, and data privacy compliance in regulated domains.**

- **Architected RAG pipelines integrating LangChain, Pinecone, and Azure Cognitive Search with real-time ETL workflows, enabling faster legal research and case summarization accuracy by 40%.**

- **Developed agentic AI systems (Crew AI, AutoGen) for dynamic task orchestration across paralegal workflows, ensuring seamless collaboration between AI agents and human teams.**

- **Implemented AI governance frameworks to monitor prompt integrity, bias mitigation, and model observability, ensuring compliance with legal and ethical standards.**

- **Implemented evaluation frameworks (RAGAS, DeepEval, FMeval) to monitor LLM performance, reliability, and**

accuracy across agent workflows.

- **Led scaling of GenAI systems from MVP to enterprise production across multi-cloud environments (AWS/GCP/Azure), supporting millions of requests.**
- **Conducted performance tuning and observability audits to ensure latency optimization, fairness, and safety of production GenAI agents.**
- **Collaborated using GitFlow branching and code review practices to maintain production-quality AI deployments.**
- **Automated end-to-end workflows (n8n, Power Automate, Zapier) integrating Salesforce, case management tools, and document databases, increasing operational efficiency by 50%.**
- **Deployed scalable AI APIs via FastAPI and Docker on Azure & AWS, enabling integration of LLM-based solutions into enterprise applications with reduced latency and cost tracking.**
- **Collaborated with attorneys, paralegals, and data engineers to design ML-driven litigation strategy insights, leveraging predictive analytics to estimate case outcomes and optimize settlements.**
- **Championed innovation initiatives, introducing multi-modal AI for processing medical records, evidence images, and structured claims data to improve fact discovery and case analysis.**

Sacred Heart University, Connecticut, USA

AI Solutions Engineer (GenAI Architect)   August 2024 – Mar 25

- o **Engineered an AI-driven grant recommendation system** by implementing **deep learning models (CNNs, RNNs, and Transformer-based LLMs)** to optimize researcher-grant matching, improving accuracy by **35%**.
- o **Developed an autonomous AI agent** leveraging **Reinforcement Learning (RLHF)** and **prompt engineering** to refine grant-matching recommendations dynamically, reducing researcher effort by **50%**.
- o **Automated web scraping and data ingestion pipelines** using **Python, Scrapy, and Selenium**, integrating **AI-driven entity recognition (NER) with spaCy and BERT** to extract structured insights, reducing manual effort by **60%**.
- o Designed **agentic architectures with persistent memory and planning capabilities**, improving reasoning depth and decision reliability in AI-powered grant recommendations.
- o Optimized LLMs and small language models (SLMs) for cost-efficient deployments in production, applying pruning, LoRA, and quantization.
- o **Implemented multi-modal deep learning models** (combining **text embeddings and vectorized metadata**) to enhance **semantic search and recommendation precision**, increasing retrieval efficiency by **40%**.
- o **Integrated LLM-powered embeddings** (using **OpenAI, LangChain, and Pinecone**) with **vector databases** to power high-dimensional search, achieving **90% accuracy** in researcher recommendations.
- o **Deployed Flask-based APIs** supporting **MERN stack** front-end integration, enabling seamless interaction with AI models while reducing system latency by **50%**.
- o **Architected agentic AI frameworks (LangGraph, ReAct-style agents) to enable multi-step reasoning and autonomous task execution**.
- o **Implemented knowledge graphs and entity linking** with **Neo4j and Graph Neural Networks (GNNs)** to establish deep semantic relationships between research domains and funding opportunities.
- o **Enhanced predictive grant allocations** using **XGBoost, Random Forest, and deep reinforcement learning (DQN)**, increasing researcher success rates by **20%**.

EVRE

AI Automation Engineer – EV Charging Solutions   May 2020 –Jan 2024

- o Designed and deployed **AI/ML models** (Time Series Forecasting, XGBoost, Deep Learning) to optimize EV charging demand prediction, improving station utilization accuracy by 30%.
- o Applied teacher–student distillation techniques to compress large models into lightweight, edge-deployable AI systems without performance loss.
- o Built **autonomous scheduling and load-balancing agents** using Generative AI and Reinforcement Learning to dynamically allocate charging resources and reduce peak-hour congestion by 40%.
- o Implemented **computer vision and NLP models** for monitoring charging equipment, analyzing fault logs, and predicting maintenance needs, reducing downtime by 25%.
- o Developed **real-time RAG pipelines** combining IoT sensor data, weather inputs, and traffic APIs to improve predictive analytics for EV charging patterns.

- o Engineered **ETL data pipelines** (batch and streaming) integrating IoT, RAG, and vectorized analytics to optimize predictive performance across EV infrastructure networks.Integrated **AI-powered dashboards** (Tableau, Power BI) providing stakeholders with live insights into charging station performance, revenue tracking, and anomaly detection.
- o Enhanced **fraud detection and payment anomaly models** with LLM-driven analytics, safeguarding EV charging transactions and ensuring regulatory compliance.
- o Applied **Agile & SCRUM methodologies** to deliver AI-driven EV infrastructure automation projects on time and within scope.

Randstad Technologies
AI Automation Engineer – Talent AI Systems   Jan 2019 – May 2020
- o **Contributed to automation initiatives** at **Randstad Offshore Services (now Randstad Digital)**, assisting in streamlining the **hiring and sourcing process** through **RPA and data-driven solutions**.
- o Assisted in developing **end-to-end automation tools** using **Python, UiPath, and SQL**, reducing **manual sourcing efforts by 60%** and improving candidate pipeline efficiency.
- o Participated in designing and deploying **Flask APIs** to integrate with internal **ATS (Applicant Tracking System)** and **CRM databases**, optimizing data workflows and reducing processing time by **25%**.
- o Helped automate **resume parsing, job matching, and candidate shortlisting** using **NLP, machine learning models, and web scraping**, improving recruitment accuracy and speed.
- o Assisted in building **real-time dashboards in Tableau and Power BI**, transforming raw hiring data into actionable insights for strategic decision-making.
- o Worked on enhancing **data management** by automating **Google Sheets and SQL databases**, reducing redundancy and improving data accuracy.
- o Contributed to delivering **5+ automation projects ahead of schedule**, leveraging **ETL pipelines, API integrations, and reusable RPA components**, resulting in **$5M annual savings** and increased recruiter productivity.

Genpact
Data Analyst- Contract   Dec 2017 –Dec 2018
- o Spearheaded AI/ML initiatives to enhance **WhatsApp Payments infrastructure**, focusing on fraud detection, payment anomaly prediction, and transaction optimization.

- o Built **deep learning models (RNNs, Transformers, and Graph Neural Networks)** to detect fraudulent payment behaviors in real time, reducing false positives by 35% and improving fraud catch-rate by 45%.
- o Developed **Generative AI and LLM-powered chatbots** to provide secure, conversational payment support inside WhatsApp, streamlining user interactions and reducing support costs by 40%.
- o Designed and deployed **risk-scoring pipelines** using SageMaker and TensorFlow to assess transactions in real time, enabling faster payment approvals while maintaining regulatory compliance.
- o Implemented **entity recognition (NER) and anomaly detection** using BERT and PyTorch to extract insights from payment metadata, chat logs, and transaction histories.
- o Automated **end-to-end payment processing workflows** with FastAPI, Docker, and CI/CD pipelines, ensuring scalable integration with WhatsApp and backend banking APIs.
- o Created **AI governance and observability frameworks** to monitor LLM and ML-based payment models, ensuring transparency, auditability, and compliance with PCI-DSS and financial regulations.
- o Collaborated with cross-functional teams (engineers, compliance officers, banking partners) using Agile/SCRUM methodologies to deliver AI-driven financial solutions ahead of deadlines.

Projects

- o

- o **Autonomous Research Agent (LangGraph + Pinecone):** Built a multi-agent framework with persistent memory and feedback loops for research summarization.
- o **Document QA Agent (AutoGen + Azure Search):** Created RAG-powered assistant for legal document summarization and compliance validation.

- o **Open-Source Contributor:** Shared experimental notebooks on LangChain agent evaluation and RAGAS-based scoring.

CERTIFICATIONS

Microsoft Certified: Azure AI Fundamentals, Google Data Analytics, MTA: Introduction to Programming Using Python - Certified 2019, Excel Skills for Data Analytics and Visualization Specialization, **Certified Scrum Product Owner (CSPO) – In Progress**, **Certified Scrum Practitioner (CSP) – Expected 2025**

EDUCATION

- o Sacred Heart University, USA: Master of Science in Computer Science, Data Science Track (Mar 2025)
- o NMIMS,: Master of Business Administration, IT and HRM, (Dec 2023)
- o RTMNU University,: Bachelor of Engineering, Information Technology, (Dec 2020)

REFERENCES

1. Robert Mccloud – Dean – Sacred Heart University
2. Jeff Hinke – Account Manager – Randstad Technologies