

GIRISH VUYYURU

Houston, TX | +1 346 434 8296 | girishsap45@gmail.com | linkedin.com/in/girish-genai-engineer | github.com/gireeshvuyyuru501-design | gireeshvuyyuru501-design.github.io/girish-ai-engineering-portfolio

Senior AI/ML Engineer | Generative AI · Agentic AI (LangGraph, LangChain, MCP) · RAG · LLM Evaluation | Python · FastAPI · MLOps | AWS · Azure · GCP · Docker · Kubernetes

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with 6+ years building and deploying production AI for financial-services and federal clients. Designs agentic AI and multi-agent systems (LangGraph, LangChain, Google ADK) with tool calling, shared state, memory and human-in-the-loop approval; builds MCP and FastMCP servers; and builds RAG and GraphRAG pipelines evaluated with RAGAS. Raised RAG faithfulness from 0.57 to 0.91 and cut hallucinations by 43% at the U.S. Department of Veterans Affairs. At Credit Union of Texas, raised answer accuracy from 75% to 88%, cut response time from 8s to 4s, and saved about 125 hours a month for 25 users. Works with OpenAI, Azure OpenAI, AWS Bedrock and Google Gemini on Vertex AI, and with vector databases (pgvector, Pinecone, ChromaDB, FAISS). Applies AI governance: guardrails, PII/DLP and compliance controls for regulated data. Uses Claude Code daily.

TECHNICAL SKILLS

Generative AI & LLMs: LLMs, Large Language Models, Prompt Engineering, Context Engineering, OpenAI APIs, Azure OpenAI, AWS Bedrock, Google Gemini, Vertex AI, Anthropic Claude, Claude Code

Agentic AI & Orchestration: LangGraph, LangChain, Google ADK (Agent Development Kit), MCP, FastMCP, Multi-Agent Systems, AI Agents, Agent Orchestration, Agent-to-Agent (A2A), Tool Calling, Function Calling, Conditional Routing, Shared State, Agent Memory, Human-in-the-Loop, ReAct, LlamaIndex, AutoGen, CrewAI, Semantic Kernel, OpenAI Agents SDK, Claude Agent SDK

RAG & LLM Evaluation: RAG, GraphRAG, Retrieval-Augmented Generation, Embeddings, Semantic Search, Hybrid Retrieval, Reranking, pgvector (PostgreSQL), Pinecone, ChromaDB, FAISS, RAGAS, LLM Evaluation, LLM-as-a-Judge, Evaluation Harnesses, Langfuse, LangSmith, Hallucination Detection

Machine Learning & Backend: Python, FastAPI, Pydantic, REST APIs, Microservices, asyncio, SQL, PostgreSQL, PyTorch, TensorFlow, Scikit-learn, XGBoost, Hugging Face Transformers, NLP, Deep Learning, PySpark, Pandas, NumPy

Cloud, MLOps & Observability: AWS (Lambda, SageMaker, ECS, EKS, Bedrock, S3, IAM, CloudWatch), Azure (Azure OpenAI, AI Search), GCP (Vertex AI), Docker, Kubernetes, MLflow, MLOps, LLMOps, CI/CD, Jenkins, GitHub Actions, Airflow, Terraform, OpenTelemetry, Prometheus, Grafana

AI Governance & Security: AI Governance, Guardrails, PII Protection, Data Loss Prevention (DLP), Compliance Controls, Responsible AI, IAM Least Privilege, OAuth 2.0, Audit Logging, FISMA

PROFESSIONAL EXPERIENCE

Senior AI/ML Engineer | Credit Union of Texas | Dec 2025 – Present | Houston, TX

- **RAG & GraphRAG pipelines:** Built RAG and GraphRAG pipelines (ingestion, embeddings, semantic search on pgvector, Pinecone and ChromaDB, reranking) for financial document intelligence over about 2,000 documents. 25 users across 3 teams run about 150 queries a day, saving roughly 125 hours a month.
- **LLM evaluation & monitoring:** Built a RAGAS evaluation framework (faithfulness, answer relevancy, context precision and recall) on a 100-question set, raising accuracy from 75% to 88% and cutting response time from 8s to 4s. Added LLM-as-a-judge and deterministic assertion checks to CI/CD, and Prometheus and Grafana dashboards tracking response quality drift, API latency and token use.
- **MCP servers & tool access:** Built FastMCP servers used by 5+ applications behind Python asyncio and FastAPI services, with PII/DLP guardrails, audit logging and OAuth 2.0 / IAM least privilege.
- **Multi-agent workflows:** Designed LangGraph and Google ADK multi-agent workflows with tool calling, conditional routing, shared state, memory and human-in-the-loop approval, using Gemini on Vertex AI, OpenAI and AWS Bedrock.
- **Guardrails on regulated data:** Implemented PII masking, DLP controls, input and output validation and audit logging for LLM applications on regulated financial data.
- **Stakeholder delivery:** Gathered requirements from business stakeholders, demoed working prototypes to non-technical users, and built their feedback into production.
- **APIs & CI/CD:** Built FastAPI microservices connecting LangGraph agents, LangChain RAG pipelines and MCP services to enterprise applications, documented with OpenAPI/Swagger, with SonarQube quality gates in CI/CD.

AI Engineer / ML Engineer | U.S. Department of Veterans Affairs | Jan 2023 – Nov 2025 | Washington, D.C.

- **RAG quality:** Raised RAG answer faithfulness from 0.57 to 0.91 (RAGAS) and cut hallucinations by 43%, with Langfuse and LangSmith tracing and evaluation.

- **LangChain RAG & LLM applications:** Built Python LangChain RAG applications with OpenAI APIs for enterprise knowledge management: document retrieval, semantic search and LLM response pipelines for veteran services workflows, under federal security requirements (FISMA).
- **AI services on AWS:** Built FastAPI microservices deploying AI/ML and LLM workloads on AWS SageMaker, Lambda, ECS and EKS, with IAM least privilege and OAuth 2.0.
- **NLP & document intelligence:** Built NLP classification, entity extraction and semantic document pipelines with Hugging Face Transformers, with SageMaker batch inference.
- **Monitoring & drift detection:** Built CloudWatch and Prometheus monitoring with drift alerts for LLM response quality, model accuracy and API latency.
- **MLOps:** Containerized applications with Docker and shipped them through Jenkins and GitHub Actions CI/CD, with MLflow experiment tracking and model versioning.

Data Scientist | Radixweb | Jun 2020 – Dec 2021 | India

- **Model APIs:** Built FastAPI REST APIs (Pydantic, OpenAPI/Swagger) exposing ML and NLP models for real-time prediction, containerized with Docker and deployed on AWS SageMaker, with experiments and model versions tracked in MLflow.
- **NLP & text modeling:** Built NLP text classification, semantic similarity and embedding-based retrieval pipelines with PyTorch, TensorFlow, Hugging Face Transformers, Scikit-learn and XGBoost.
- **Model evaluation:** Evaluated models with precision, recall, F1 and ROC-AUC, using cross-validation and hyperparameter tuning to improve generalization.
- **Data pipelines:** Built Pandas, NumPy, SQL and PySpark pipelines for preprocessing and feature engineering, scheduled with Apache Airflow, with Jenkins CI/CD for model validation and release.

Data Science Intern | Radixweb | Jun 2019 – Jun 2020 | India

- Built Python data preprocessing pipelines and Scikit-learn models validated with train/test splits and cross-validation; performed EDA and feature engineering; used Docker, Git and Jupyter.

PROJECTS

- **MCP Employee Knowledge Assistant:** Enterprise knowledge assistant built on Model Context Protocol (MCP) with controlled tool access, retrieval and agent workflows. github.com/gireeshvuyyuru501-design/enterprise-employee-knowledge-assistant
- **Enterprise Finance Agent:** Multi-agent finance system covering portfolio, budget, expense, risk and reporting workflows. github.com/gireeshvuyyuru501-design/enterprise-finance-agent
- **Enterprise LLM Chat Assistant:** Production-minded LLM application showing reusable enterprise AI integration patterns. github.com/gireeshvuyyuru501-design/enterprise-llm-chat-assistant

EDUCATION & CERTIFICATIONS

M.S. Computer Science, Texas A&M University-Kingsville | Jan 2022 – Dec 2022 | GPA 4.0/4.0

B.Tech. Computer Science & Engineering, Vignan's University, India | Jul 2016 – May 2020

Claude API Fundamentals · Agent Skills · Claude Code · MCP Advanced Topics · Deep Learning Specialization