

Lokesh Reddy V

Generative AI Engineer · RAG & Agentic Systems · LLM Evaluation & Guardrails

United States · +1 980-382-4245 · lokeshrveerabhadra@gmail.com · [linkedin.com/in/l0k35hr3ddy-d3v](https://www.linkedin.com/in/l0k35hr3ddy-d3v)

PROFESSIONAL SUMMARY

Engineer with 5+ years building production AI and machine learning systems across regulated financial services, SaaS platforms, and infrastructure automation — with the last three years focused on RAG pipelines, agentic workflows, and LLM evaluation and security controls. Earlier work spans an agentic scheduling platform, NLP classifiers on BERT and RoBERTa, and infrastructure automation with natural language interfaces. Builds evaluation, access control, privacy, and auditability into system design, testing, and release processes.

SKILLS

Core	Python, FastAPI, LangChain, LlamaIndex, LangGraph, tool calling, structured outputs, REST APIs
Models & Platforms	GPT-4o, Azure OpenAI, Llama 3, BERT, RoBERTa, LoRA, QLoRA, fine-tuning, prompt engineering
Retrieval	Pinecone, FAISS, ChromaDB, Weaviate, hybrid search, reranking, metadata filtering, RAG
Evaluation	RAGAS, DeepEval, CI/CD quality gates, task-specific test sets, regression testing
Cloud, Data & MLOps	Azure, Docker, Kubernetes, MLflow, Git, Databricks, PySpark, SQL
Security	Prompt injection defense, PII redaction, access-aware retrieval, audit logging, content controls

EXPERIENCE

Generative AI Engineer

Sep 2026 – Present

Tata Consultancy Services · United States

- Build RAG pipelines over approved internal knowledge sources, covering ingestion, chunking, metadata enrichment, access-aware filtering, hybrid retrieval, reranking, grounded generation, and source attribution.
- Design multi-step LLM workflows for document processing, summarization, Q&A, and compliance checks — explicit state management, structured outputs, validation, retries, and human review for sensitive or high-impact actions.
- Implement security controls as first-class engineering requirements: prompt injection defenses, PII redaction, authorization-aware retrieval, output validation, content controls, and audit logging.
- Develop evaluation suites measuring retrieval quality, groundedness, answer relevance, policy adherence, latency, and cost; integrate quality thresholds into release pipelines.
- Package AI capabilities as FastAPI services and deploy within approved cloud and container boundaries with versioning, monitoring, and rollback controls.

Senior AI Engineer — AI/ML & Agentic Systems

Jan 2024 – Aug 2026

Robert Half · United States

- Led development of an agentic AI platform for a scheduling and booking product, covering appointment management, intent resolution, cancellations, and exception handling; reduced booking-related escalations by over 20%.
- Built a production RAG pipeline with hybrid retrieval, metadata filtering, reranking, grounded generation, and source attribution; introduced GPT-4o and Pinecone as the platform matured.
- Designed LangGraph workflows with explicit state management, controlled tool execution, retries, validation, and human review for ambiguous or high-impact actions.
- Established automated LLM evaluation using RAGAS, DeepEval, and 200+ curated task-specific test cases, with CI/CD quality gates integrated into every release pipeline.
- Fine-tuned Llama 3 with LoRA and QLoRA for domain adaptation; benchmarked against prompt-based baselines across quality, latency, and serving cost.
- Implemented PII detection, content safety controls, observability, failure monitoring, and Kubernetes autoscaling, supporting traffic spikes up to 10x baseline without manual scaling.

Full Stack AI Engineer

Aug 2021 – Dec 2022

BT Group · United Kingdom

- Built full-stack AI capabilities for a developer-facing infrastructure automation product, including natural language interfaces for cloud provisioning and infrastructure-as-code generation.
- Developed a retrieval-backed knowledge assistant over engineering documentation and runbooks, improving access to infrastructure guidance for a team of 100+ engineers.
- Designed multi-step request processing with input validation, policy checks, structured outputs, and human approval before infrastructure actions were executed.
- Implemented safeguards for unsafe instructions and sensitive values through input validation, secrets masking, allowlists, output checks, and audit logging.
- Created repeatable evaluation datasets and regression tests for retrieval quality, generated responses, and infrastructure code before production releases.

Application Development Engineer — AI/ML Practice

Nov 2018 – Dec 2019

Tech Mahindra · India

- Developed NLP capabilities for financial services and retail projects, including document classification with BERT and sentiment analysis with RoBERTa during 2019.
- Built a transformer-based request routing workflow that classified service tickets and directed them to the appropriate support queues.
- Prepared training and evaluation datasets, compared model variants, performed error analysis, and tuned classification thresholds for production use cases.
- Standardized reusable MLflow patterns for experiment tracking, model versioning, packaging, and deployment across delivery workflows.
- Exposed model predictions through REST APIs and supported integration with enterprise applications and operational workflows.

EDUCATION

MS, Computer Science · Campbellsville University

2023 – 2024

MSc, Information Technology Security · Nottingham Trent University

Jan 2020 – Aug 2021

BTech, Electronics & Computer Science Engineering · KL University

2014 – 2018