Jayesh Koli

Enterprise GenAI Engineer | Backend Systems & AI Governance

Panvel, Navi Mumbai, India
+91 93724 68778
linkedin.com/in/jayesh-koli-042357227
jkoli6704@gmail.com
github.com/Jayesh12356

Professional Summary

Enterprise GenAI Engineer with a strong backend and cloud foundation, with 4+ years designing and deploying production-grade LLM systems for regulated enterprise workflows. Proven ability to integrate AI into finance and compliance-sensitive systems using human-in-the-loop controls, auditability, cost-aware architecture, and reliability-first design. Background in large-scale enterprise platforms serving 10,000+ users across finance, procurement, and corporate services.

Currently building production GenAI systems at Jio Platforms (Reliance) — LLM-assisted financial decision workflows with deterministic fallbacks, RAG over policy documents with role-based access controls, and a centralized LLM evaluation/governance framework with prompt versioning, drift detection, hallucination guards, and rollback. Deep focus on retrieval quality (hybrid search + reranking), structured LLM I/O (JSON mode, Pydantic), trace-level observability, and semantic-cache-driven cost control.

Enterprise GenAI Capabilities

- Designed LLM-powered decision support systems with deterministic fallbacks, confidence-based routing, and human approval workflows for compliance-sensitive use cases.
- Built Retrieval-Augmented Generation (RAG) pipelines over internal enterprise data with hybrid retrieval (BM25 + dense vectors), reranking, and strict role-based access controls.
- Implemented structured LLM outputs (JSON mode, Pydantic schemas) and SSE token streaming so model responses can flow safely into downstream backend systems and real-time approval consoles.
- Established LLM evaluation and governance frameworks: offline regression suites, LLM-as-a-Judge scoring, drift and hallucination detection, prompt/configuration versioning, and rollback strategies.
- Integrated GenAI services into event-driven backend architectures with retries, circuit isolation, semantic caching for cost control, rate limiting, and full trace-level observability across prompt, retrieval, and generation steps.
- Deployed cloud-native GenAI applications with audit logging, PII redaction, jailbreak and topic-boundary guardrails, and granular cost and token-budget enforcement.

Work Experience

Software Engineer — Enterprise Backend & GenAI Systems
Jun 2022 — Present

Jio Platforms Limited (Reliance Corporate Park), Navi Mumbai, India

- Designed and deployed enterprise-grade backend systems for finance, procurement, and corporate services workflows, improving system response times by 30% through architectural and performance optimizations.
- Built automated financial decision workflows integrating LLM-assisted classification with deterministic validation rules and SAP-based approval systems, enabling auditable, human-in-the-loop processing for compliance-sensitive expense proposals and improving workflow efficiency by 25%.
- Architected a scalable financial management platform that reduced procurement cycle times by 20%, with extensibility for future commercialization and strict data integrity guarantees.
- Developed event-driven enterprise services using RabbitMQ and Node.js to power real-time dashboards and notifications, improving information delivery latency by 35% while maintaining system reliability.
- Shipped production GenAI surfaces (RAG-grounded policy assistant + LLM Ops platform) to internal users, integrating prompt versioning, hallucination guards, and trace-level observability into the standard enterprise release lifecycle.
- Implemented role-based access control (RBAC) and permission management across internal platforms, enforcing strict data boundaries and eliminating duplicate or inconsistent authorization states.
- Collaborated with cross-functional teams across engineering, finance, and operations to deliver secure, scalable systems used by 10,000+ employees.

Projects

Enterprise Financial Policy Assistant (GenAI + RAG)
GenAI · RAG · Hybrid Retrieval · Human-in-the-Loop · LLM Governance

- Designed an LLM-assisted financial analysis system grounded in internal policy documents using Retrieval-Augmented Generation (RAG), ensuring responses were constrained to approved enterprise knowledge sources.
- Implemented role-based document retrieval, metadata filtering, and structured prompt templates to prevent unauthorized access and reduce hallucinations in compliance-critical use cases.
- Improved retrieval quality with hybrid search (BM25 + dense vectors over pgvector) and a reranking stage, plus tuned chunking strategies (recursive + semantic) to keep grounding tight on adversarial finance queries.
- Streamed structured JSON outputs (Pydantic-validated) directly into the approval console via Server-Sent Events, with token-level streaming for sub-second perceived latency on long answers.
- Introduced confidence-based routing and human-in-the-loop approval workflows, maintaining full audit trails of prompts, retrieved context, model outputs, and final decisions.
- Evaluated model reliability using curated test cases and production feedback signals, monitoring correctness, latency, and token usage to optimize cost-performance tradeoffs.

LLM Evaluation, Monitoring & Governance Framework
Centralized LLM Ops · Trace Observability · LLM-as-a-Judge · Cost Budgeting

- Built a centralized framework for testing, monitoring, and governing LLM behavior across enterprise applications.
- Implemented prompt and configuration versioning, treating GenAI behavior as code with traceable changes and rollback capability.
- Designed automated response quality scoring pipelines using LLM-as-a-Judge rubrics (correctness, faithfulness, policy compliance) alongside heuristic checks to detect hallucinations, retrieval failures, and behavioral drift over time.
- Instrumented end-to-end trace logging across prompt, retrieval, generation, and output (LangSmith / Langfuse-style spans) to enable post-hoc replay and root-cause analysis on failed runs.
- Monitored latency, token usage, and error rates with semantic caching and per-tenant budgets to enforce cost and reliability constraints for production GenAI systems.

Education

M.Tech — Software Systems — *CGPA 8.1 · Jun 2022 — Jun 2026*
BITS Pilani (WILP)

B.Sc. Computer Science — *CGPA 9.0 · Jun 2019 — Mar 2022*
Pillai College of Arts, Commerce & Science (Autonomous)

Higher Secondary — Science (CBSE) — *90% · Jun 2017 — Mar 2019*
Kendriya Vidyalaya Jr. College

Secondary School (CBSE) — *92.8% · Jun 2016 — Mar 2017*
Datta Meghe World Academy School

Skills

GenAI & LLM Engineering: LLMs (OpenAI GPT-4 / 4o, Anthropic Claude), Embeddings (text-embedding-3, BGE), Prompt Engineering (Few-shot, Chain-of-Thought, ReAct), Function Calling / Tool Use, Structured Outputs (JSON Mode, Pydantic schemas), Streaming Responses (SSE), Human-in-the-Loop Workflows

RAG & Retrieval: Retrieval-Augmented Generation, Hybrid Search (BM25 + Vector), Reranking (Cohere Rerank, BGE rerankers), Query Rewriting / HyDE, Chunking Strategies (Recursive, Semantic, Sliding Window), Metadata Filtering, Role-based Retrieval, Semantic Caching

LLM Ops & Governance: Prompt & Configuration Versioning, Offline Regression Suites, LLM-as-a-Judge Evaluation, Drift Detection, Hallucination Guards, Trace-level Observability (LangSmith / Langfuse-style), Token & Cost Budgeting, Rate Limiting, Audit Logging, Rollback Strategies, Guardrails (Guardrails.ai, NeMo Guardrails), PII Redaction, Jailbreak & Topic-boundary Defense

Frameworks & Orchestration: LangChain, LlamaIndex, LangGraph, OpenAI SDK, Anthropic SDK

Backend & Systems: Python, Node.js, FastAPI, Express.js, REST APIs, OpenAPI / Swagger, Async & Event-Driven Architecture (Kafka, RabbitMQ), Microservices, Secure API Design

Data & Storage: PostgreSQL, MongoDB, MariaDB, Oracle, Redis, pgvector, FAISS, Pinecone, Qdrant, Weaviate

Frontend: JavaScript, TypeScript, React.js, Next.js, Redux, Tailwind CSS

Enterprise Platforms: SAP Integrations, Frappe Framework, RBAC & Access Control, Compliance Workflows

Engineering Practices: Git, Agile, CI/CD, Performance Optimization, Audit Logging, Cost Monitoring, Reliability-first Design

Certifications

IBM Generative AI Engineering — *Coursera*
Server Side Development with Node.js, Express.js & MongoDB — *Coursera*
The Ultimate React Course — React, Next.js, Redux & More — *Udemy*