LIKHITH V

likhithv02@gmail.com | +91 9480417715 | Bengaluru, India | likhithv.com | LinkedIn | GitHub

SUMMARY

AI Engineer with **2 years** of experience building and evaluating LLM systems, with deep expertise in **document intelligence** and **RAG**, which has been the core of my work across every role. At Networth Corp, built a **custom chunking module** that identifies a document's hierarchical structure (sections, sub-sections, tables, lists) and chunks along it, improving a RAG chatbot's response accuracy over insurance documents by **20%**; also built its evaluation and query-enrichment modules. Built a refund-processing support agent and tested it hard: evaluated on **83 golden conversations** on every pull request (**100%** policy-correct on Gemini and GPT), red-teamed with **120 LLM-generated attacks** (**0 breaches**), and load-tested to **2,500 concurrent users** with **0 failed requests**. Also shipped an on-device document assistant running Gemma 4 to Google Play.

WORK EXPERIENCE

AI Engineer | Networth Corp *October 2024 - Present*

- Developed a customized chunking module for a **RAG-based chatbot** over **unstructured insurance documents**, improving response accuracy by **20%** and reducing operational costs by **2%**.
- Built **LLM evaluation** and **query enrichment** modules to test retrieval quality, improve prompt/retrieval logic, and increase answer reliability in document-based AI workflows.
- Designed AI workflows around embeddings, vector retrieval, document parsing, and source-grounded generation to reduce hallucination in knowledge-base Q&A systems.
- Currently building a **synthetic data generation** application for structured, semi-structured, and unstructured data while **preserving correlations** with original datasets.

AI/ML Intern | PG-AGI | CERTIFICATE *December 2023 - June 2024*

- Led development of **6 POCs** and **2 production** GenAI projects, including HireXtra's **AI Recruiter**, with ownership from requirement understanding to implementation.
- Built **AI agents** and RAG pipelines using **Python, FastAPI, MongoDB, LangChain, and LlamaIndex** for production-oriented AI applications.

PROJECTS

Refund Agent - AI Customer-Support Agent | LIVE DEMO | CASE STUDY | GITHUB *2026*

- A support agent that approves, denies or escalates refunds, built with **LangGraph, FastAPI, Postgres and Redis**. The model handles the conversation, but every refund goes through a policy check in plain code, so it can't be talked into a bad one.
- Graded on **83 test conversations** on every pull request. It reaches the right policy decision on all of them with both Gemini and GPT.
- Red-teamed with **promptfoo**: 120 attacks written by another LLM, and none of them moved money or leaked data.
- Load-tested with **Locust** up to **2,500 concurrent users** with zero failed requests, after profiling found a bottleneck and fixing it raised throughput by 51%.

Anvit - On-Device AI Document Assistant | PLAY STORE *2026*

- Building a privacy-first **KMP** app for Android and iOS for local document intelligence with **PDF and DOCX** support, collections, chat sessions, source citations, streaming responses, and settings-driven model configuration.
- Built custom parsing and chunking code implementing **hierarchical chunking** based on document sections/subsections, tables, and lists — ensuring each chunk contains a complete, coherent unit of content.
- Integrated **LiteRTLM** for on-device inferencing with the **Gemma 4 LLM**, running completely on the local device without sending sensitive documents to external APIs.
- Implemented **agentic RAG** workflows with query routing, decomposition, corrective relevance evaluation, selective content reduction, native tool calling, and self-critique loops to reduce hallucination.

DeltaVec - Incremental Embedding Pipeline | DEMO | GITHUB *2026*

- Built an **incremental vector embedding pipeline** keeping Qdrant vector stores synchronized with mutating PostgreSQL sources using watermark change detection and SHA-256 content-hash diffing.
- Implemented **idempotent upserts** via deterministic chunk IDs, **orphan chunk deletion**, **tombstone propagation** for soft deletes, and automated reconciliation sweeps to eliminate index drift.
- Developed orchestration and data quality gates with **Dagster, FastAPI, and Docker**, featuring an interactive live showcase application deployed to GitHub Pages.

SKILLS

- **AI/LLM:** RAG, Agentic RAG, Agentic Workflows (ReAct, tool calling), Guardrails & Prompt-Injection Defense, Prompt Design, Hallucination Control, Source-Grounded Q&A, Semantic Search, Embeddings, Document Processing, On-Device LLMs
- **LLM Evaluation & Testing:** Golden-Set Evals in CI, LLM-as-Judge (Opik), Red Teaming (promptfoo), Retrieval Evaluation, Load Testing (Locust), pytest
- **Frameworks & Models:** LangGraph, LangChain, LlamaIndex, Autogen, OpenAI, Anthropic Claude, Google Gemini, Gemma (LiteRT-LM), Transformers, Sentence Transformers
- **Languages & Backend:** Python, SQL, Kotlin Multiplatform, TypeScript, FastAPI, Flask, SQLAlchemy, REST APIs, Server-Sent Events, React
- **Data & Vector Stores:** Qdrant, FAISS, Weaviate, Pinecone, Neo4j, PostgreSQL, MySQL, MongoDB, Redis, Dagster
- **Infrastructure & Observability:** Docker, Kubernetes, GCP (Cloud Run), Cloudflare, GitHub Actions, Prometheus, Opik Tracing, Git/GitHub, MCP

EDUCATION

BE in Artificial Intelligence and Machine Learning, Visvesvaraya Technological University, Bengaluru

Dec 2020 - May 2024