Aswin K V

+91 9497210776 | aswin.kv202@gmail.com | aswinkv.com | linkedin.com/in/aswin-kv
github.com/Aswin-K-V | huggingface.co/Aswinkv07

Summary

Software Engineer with expertise in AI, experienced in building and shipping enterprise-grade LLM applications — agentic systems, multi-agent orchestration, RAG pipelines, and knowledge-graph reasoning. Combines strong backend engineering with hands-on GenAI development, from LLM fine-tuning and prompt engineering to scalable APIs and production deployment. Open-source contributor with published libraries and models, focused on production reliability, observability, and cost-efficient systems.

Technical Skills

Languages: Python, Java, C, SQL
GenAI & Agentic AI: LangChain, LangGraph, MCP, A2A, LLMs, LLM Fine-Tuning, Agentic Workflows, Prompt Engineering, Tool Calling
RAG, Search & Graphs: RAG Pipelines, FAISS, Chroma DB, Embeddings, Semantic Search, Neo4j, Cypher, Elasticsearch
LLM Reliability & Observability: Guardrails, Langfuse (Tracing, Token & Cost Tracking), Human-in-the-Loop Workflows
Backend & Tools: FastAPI, SQLAlchemy, Alembic, JWT/OAuth2, Docker, Git, AWS, MongoDB, Postgres

Experience

AI Engineer — Feb 2025 – Present
Tata Consultancy Services — *Hyderabad, India*

Agentic Decision-Intelligence Platform for Semiconductor Supply Chain
Python,Pandas,LangGraph, Neo4j, RAG, LiteLLM, Nvidea Nemo, FastAPI, Docker, Langfuse
- Delivered an agentic AI platform that acts as a command center for supply-chain executives — live dashboards, a conversational insights chatbot, proactive recommendations, and automated workflows — shortening the path from disruption to decision.
- Modeled the end-to-end supply chain — suppliers, fabs, assembly and test, logistics, and customers — as a dynamic Neo4j knowledge graph, combining multi-hop Cypher reasoning with RAG so agents surface risks, bottlenecks, and single-source dependencies in real time.
- Implemented Intelligent Choice Architecture (ICA): agents frame vetted decision options with trade-offs (e.g., alternate sourcing and mitigation actions) and trigger automated workflows, so executives act on recommendations rather than raw data.
- Hardened the platform for enterprise use with guardrails on agent outputs and Langfuse observability for end-to-end tracing, debugging, and per-query token and cost tracking.

Multi-Agent Orchestration Framework
LangGraph, MCP, Azure OpenAI, Python, FastAPI
- Engineered a framework-agnostic orchestration system where independent agents, tools, and workflows register dynamically and compose into pipelines — via MCP and custom communication protocols .
- Implemented human-in-the-loop approval gates for high-impact agent actions, making autonomous decisions auditable and safe for enterprise AI workflows.

Codebase Documentation RAG Chatbot
RAG, FAISS, FastAPI, Python, Azure OpenAI
- Built a RAG assistant that generates, indexes, and retrieves technical documentation directly from source-code repositories — owning ingestion, chunking, embeddings, and FAISS-backed semantic search end to end.
- Enabled natural-language querying of APIs, architecture, and code behavior, accelerating developer onboarding and cutting time spent navigating unfamiliar codebases.

Projects

cyphersmith — Open-Source Text-to-Cypher Python Library | **PyPI** | **GitHub**
Python, LiteLLM, Neo4j, CyVer,
- Authored and published a pip-installable library that converts natural-language questions into validated, read-only Cypher and executes them on Neo4j, with LiteLLM-based provider-agnostic support for all llm providers and local models.
- Engineered a layered safety pipeline — read-only keyword/procedure blocking plus CyVer syntax, schema, and property validation with retries — and shipped an interactive CLI with schema-aware prompting, injectable business context, and live progress logs.
- Fine-tuned and published a Qwen3.5-4B LoRA adapter for Neo4j text-to-Cypher generation on Hugging Face, complementing the library with an open model.

Content Publishing Platform | **GitHub**
FastAPI, Async SQLAlchemy, PostgreSQL, Alembic, JWT/OAuth2, Pydantic, AWS S3
- Built a fully async REST backend for a blog platform — users, posts, profile images, and password resets — with FastAPI and SQLAlchemy 2.x async ORM, using Pydantic validation and Alembic-managed schema migrations across PostgreSQL and SQLite.
- Implemented end-to-end security: Argon2 password hashing, OAuth2/JWT bearer authentication, and owner-only authorization on protected resources, plus background SMTP password-reset emails and Pillow image processing with S3-backed storage.

Multi-Agent Log Analysis System
Google vertex, RAG, FAISS, FastAPI, Elasticsearch
- Developed an AI-powered log intelligence platform where specialized agents collaborate on log parsing, anomaly detection, failure analysis, and root-cause investigation.
- Enabled natural-language investigation of system behavior, error patterns, and incidents through RAG over Elasticsearch-backed retrieval, reducing manual log triage effort.

Certifications

AWS Certified Cloud Practitioner

Education

Saintgits College of Engineering — Kottayam, Kerala
B.Tech in Computer Science and Engineering — *2020 – 2024*