

Asgar Ali Sha

GENAI DEVELOPER | LLM & RAG ENGINEER | AGENTIC AI

📍 Chennai, Tamil Nadu | 📞 +91 7339661311 | ✉️ asgar156d@gmail.com | in LinkedIn | 🐙 GitHub | 📁 Portfolio

PROFILE SUMMARY

GenAI Developer with 5+ years of software engineering experience, the last two focused on shipping LLM-powered systems to production. Built RAG pipelines over vector stores, an agentic tool layer of 3 MCP servers giving non-technical users natural-language access to SQL, documents, and repositories, and prompt chains extracting structured data from 100+ document formats at 95%+ accuracy. Automated 80% of a manual data entry workflow through LLM-driven pipelines. Works across the full delivery path: prompt and retrieval design, Python service implementation with FastAPI and LangChain, and containerised deployment through CI/CD.

CORE COMPETENCIES

LLM Integration | RAG Pipelines | Prompt Engineering | Agentic AI & Tool Use | MCP Server Development | Embeddings & Vector Search | Function Calling | Structured Output Validation | NLP | Python Backend & APIs | Cloud Deployment | Agile / SDLC

TECHNICAL SKILLS

GenAI & LLM: OpenAI API, Claude API, Ollama (local/self-hosted models), Hugging Face, function calling, tool use, structured outputs.

Prompt Engineering: System prompt design, few-shot prompting, chain-of-thought reasoning, output schema enforcement with Pydantic, validation and retry loops, iterative prompt evaluation

RAG & Retrieval: Chunking strategies, embeddings, semantic and hybrid search, contextual retrieval, relevance thresholding, Qdrant, Chroma, Pinecone

Frameworks: LangChain, FastMCP (Model Context Protocol), FastAPI, Django, Flask, TensorFlow, PySpark

Languages & Data: Python, SQL, JavaScript, Pandas, NumPy, SQL Server, MySQL, PostgreSQL, REST APIs, Webhooks

Cloud & DevOps: Azure, GCP, Docker, Linux, Nginx, Git, Azure DevOps, CI/CD Pipelines

WORK EXPERIENCE

InteleANTS Virtual Workforce Pvt Ltd.

Specialist – Software Engineering

August 2024 – Present

- Built an **AI customer support chatbot** with a **RAG** knowledge-retrieval pipeline over a vector store, serving multi-channel support through Chatwoot webhook integration with automated escalation to human agents when retrieval confidence or intent fell outside supported scope.
- Developed 3 **MCP servers** (Model Context Protocol) with **FastMCP**, exposing tool-use interfaces for read-only SQL exploration and dependency tracing, hybrid semantic search over document collections with relevance thresholding and pagination, and Azure Git and work-item analysis, enabling non-technical users to query enterprise data in natural language.
- Designed **prompt chains** for structured data extraction reaching 95%+ accuracy through few-shot prompting, chain-of-thought reasoning, and **Pydantic** schema validation with automatic retry on malformed output, across 100+ heterogeneous document formats.
- Integrated **LLMs** (OpenAI, Ollama) via **LangChain** into production ETL pipelines for factory PO processing, firewall data extraction, and vendor data synchronisation, automating 80% of manual data entry and routing sensitive workloads to self-hosted models to control cost and data exposure.
- Shipped these systems as production **Python** services: a full-stack Call Audit QA platform (FastAPI, SQL Server, React) with 25+ features adopted by 30+ auditors, **TensorFlow** transformer-embedding models for call sentiment classification, and **PySpark** feature pipelines over millions of call records, containerised with Docker and released through Azure DevOps CI/CD.

Codentrix Technologies Pvt Ltd.

Software Engineer

September 2021 – July 2024

- Engineered the **Inventory Management System** (Django, React, MySQL) with REST APIs and schema migrations, handling inventory valuation across Raw Materials, WIP, and Finished Goods for multi-country operations.
- Created a **BOM calculation algorithm** in Python for FG and WIP items, processing 2M+ records and reducing turnaround time and manual effort by 70%.
- Developed the **Item Generation App** integrated with **Oracle ERP** for item lifecycle management, deployed across 4 countries (China, Thailand, Singapore, Japan) with role-based access control and secured API endpoints.
- Built **TracerBee** (Django, React), a part traceability application providing end-to-end tracking from Raw Material to Finished Goods, delivered through Agile sprints with peer code review and Git-based workflows.

ACHIEVEMENTS

- Recognised as the sole developer to independently deliver 5+ production-grade AI and full-stack applications within the first year at InteleANTS.
- Promoted to **‘Senior Software Engineer’** for independently managing a cross-border ERP integration project under tight deadlines with zero production incidents.
- Awarded **‘Customer Delight’** twice for delivering web applications with zero escalations.

EDUCATION

Bachelor of Engineering (Electrical Engineering) - Rajalakshmi Engineering College

August 2016 – September 2020 | First Class with 6.87 CGPA