

SHASHANK SINGH

+91-7077421913 | singhshasank2003@gmail.com | LinkedIn | GitHub | Portfolio

Professional Summary

AI/ML Engineer with **2+ years of software development experience** building **Generative AI, Agentic AI, RAG, and applied ML** systems for enterprise workflows. Architected a multi-agent **Text-to-SQL Copilot** improving SQL accuracy from **50–60% to 90–95%**, engineered Python/FastAPI/MCP workflows, built enterprise AI assistants, and trained a domain-specific **BERT-based SLM on 1,000+ FSD documents**. Strong in Python, SQL, LLMs, LangChain/LangGraph, PyTorch, Hugging Face Transformers, prompt engineering, embeddings, semantic/hybrid retrieval, vector databases, and model evaluation.

Technical Skills

Languages	Python, Java, SQL, JavaScript
GenAI / Agentic	LLMs, Generative AI, RAG, Prompt Engineering, LangChain, LangGraph, Agent Orchestration, Multi-Agent Systems, MCP, Tool Calling, Function Calling, Structured Output, LLM Evaluation
Retrieval / Vector	Embeddings, Semantic Search, Hybrid Retrieval, BM25, FAISS, Qdrant, Pinecone, Vector Databases, Reranking, Document Chunking, Schema Retrieval, Vector Indexing, GraphRAG
ML / NLP	Machine Learning, Deep Learning, NLP, Transformers, PyTorch, Hugging Face Transformers, Scikit-Learn, BERT, Classification, Regression, Model Training, Model Evaluation
Backend / Data	FastAPI, Flask, Pydantic, REST APIs, Microservices, Python SDK Development, Java/Spring Boot Integration, ETL/ELT, Data Pipelines, MySQL, MongoDB, Redis, Rasa
Cloud / Engineering	AWS, Docker, Git, GitHub, Test Automation, Intent Detection, Entity Extraction, Dialogue Management, Action Flows, Multi-Turn Conversations

Experience

Tata Consultancy Services Ltd.

Oct 2024 – Present

AI/ML Developer

Gurugram, Delhi NCR

- Engineered **production RAG pipelines** for enterprise knowledge workflows, combining document ingestion, structure-aware recursive chunking, embeddings, semantic search, vector indexing, prompt engineering, validation, and grounded generation.
- Orchestrated **domain-specific FSD assistant workflows** using intent detection, entity extraction, multi-turn dialogue, retrieval, API integration, and agentic routing to connect user intents with enterprise actions.
- Architected **LangGraph-based multi-agent workflows with MCP** for stateful routing, tool calling, retrieval, validation, conditional transitions, and integration with enterprise AI services.
- Engineered and operationalized a hybrid-retrieval **Known Error Database (KEDB)** using **BM25, FAISS, and Redis caching**, surfacing matches across **100K+ historical bug reports in under 2 seconds**, reducing duplicate defects by **27%**, and saving **120+ engineering hours/month** in triage effort.
- Engineered and deployed an AI-powered **application log analysis platform**, adopted across multiple engineering teams and saving engineers **4–5 hours/week**; converted application logs into structured error summaries, root-cause analysis, and recommended solutions using **Qwen3.5 and LLaMA 3**.
- Trained a domain-specific **BERT-based SLM on 1,000+ FSD documents** using masked language modeling with PyTorch and Hugging Face Transformers for domain-specific understanding and downstream inference.
- Developed an **Agentic Payment System** as an AI hackathon solution using multi-agent orchestration and workflow automation.

Soven Developer

Jun 2024 – Sept 2024

Software Developer

Remote

- Delivered frontend and backend features using **JavaScript and Python**, integrating REST APIs, backend services, data-processing workflows, and AI-enabled functionality.
- Implemented reusable components and API-driven workflows while supporting debugging, integration, and maintainable feature delivery.

Projects

TCS | Enterprise AI SQL Copilot GitHub

Python, FastAPI, Agentic AI, RAG, GraphRAG, Embeddings, FAISS, NVIDIA GPT OSS-20B

- Architected and internally deployed a **planner-first, read-only Text-to-SQL Copilot** with the architecture: **question → intent → entity resolution → schema linking → join discovery → query planning → validated SQL**, replacing repetitive manual query-authoring with a controlled generation workflow.
- Improved SQL-generation accuracy from approximately **50–60% to 90–95%** – a **30–45 percentage-point improvement** – through schema-aware retrieval, deterministic planning, prompt engineering, validation, error correction, and confidence-based controls for more reliable query generation.
- Trained and evaluated the system across a capital-markets schema spanning **18–20 related business tables** with explicit table/column relationships, supplemented by a schema dataset containing **100K+ columns** for broader multi-table coverage.
- Combined **BM25 and FAISS retrieval** with schema-aware table/column chunking, GraphRAG relationship traversal, and multi-hop join discovery to rank relevant tables/columns and resolve foreign-key paths.
- Designed a **7-stage LangGraph workflow** for intent detection, schema linking, query planning, SQL generation, validation, correction, and confidence scoring; implemented **4-layer guardrails** with an **80-point confidence threshold** and up to **3 self-correction retries**.
- Prototyped retrieval with FAISS, migrated the Text-to-SQL retrieval layer to **Qdrant**, and used **Pinecone** in additional vector-retrieval implementations; assessed SQL correctness through schema validation and query-success checks.

TCS | Enterprise FSD & Capital-Markets Conversational Platform

Rasa, FastAPI, Python SDKs, MCP, AWS

- Architected an enterprise **Rasa-based conversational AI workflow** with the architecture: **user intent → intent detection → action mapping → API discovery → tool invocation → data summary**, using structured dialogue, multi-turn handling, and controlled action routing for reliable task execution.
- Converted **Java microservices into reusable Python SDKs**, exposed them through FastAPI, and connected them to Rasa action flows so detected intents trigger the required enterprise APIs instead of repetitive manual application/API lookup.
- Integrated an **AWS-hosted AI tool** through a **multi-agent MCP Streamable HTTP** workflow for dynamic tool discovery, direct API invocation, application/API discovery, and capital-markets data summaries.
- Automated a **4-stage workflow** — **user intent → API discovery → tool invocation → data summary** — replacing repetitive manual navigation with intent-driven service access across FSD and capital-markets workflows.

Certifications

Agentic AI Certified Foundations Associate – Oracle | Claude Certified Architect – Professional – Anthropic | Build with Gemini – Google | AWS Certified AI Practitioner – Amazon Web Services

Education

Siksha 'O' Anusandhan University

Aug 2020 – Jul 2024

B.Tech in Computer Science & Engineering | CGPA: 8.50/10

Bhubaneswar, Odisha