

Satnam Singh

satnamsjob@gmail.com | +91 8860398075 | satnamsingh.in | linkedin.com/in/22satnam | github.com/22satnam | Gurugram

SUMMARY

Senior AI Engineer with 2.6 years of experience designing and shipping production **Generative AI systems** - multi-agent orchestration, **RAG pipelines**, and **LLM-powered applications**. Skilled in **LangGraph**, **Azure OpenAI**, and **open-source LLMs**, with a strong backend and data engineering foundation across **Python**, **FastAPI**, and cloud-native deployment. Portfolio satnamsingh.in.

WORK EXPERIENCE

ZS Associates

2024 – Present

Senior AI Engineer

Gurugram

- Architected a **multi-agent orchestration framework** using **LangGraph** and **FastAPI**, containerized with **Docker** and deployed via **Azure Container Apps**, unifying insights across **30+ reports and 150+ KPIs** with **95% analytical accuracy** and sub-12-second latency for Fortune 500 pharma clients.
- Engineered a **retrieval decision-making pipeline** with **Azure AI Search** and **Azure AI Foundry** delivering multi-modal responses – real-time text summaries, dynamic graphs, and business-insight methodologies.
- Built **conversational memory** on an Azure Redis-backed state layer, resolving complex follow-up questions by dynamically updating query parameters without re-running the full **RAG** pipeline.
- Developed **“Power Assist,”** a production **GenAI assistant** integrated with **PowerApps**, serving real-time text, graph, and insight responses to pharmaceutical KPI queries via a master-agent **LLM routing** design.
- Go-to engineer for AI and data architecture across workstreams; earned the **Client Impact Champion Award (2025)** for a first-of-its-kind Microsoft Fabric implementation that **cut delivery timelines by 40%** and became a reusable firm-wide accelerator.
- Established a **Medallion architecture** and automated DQM framework processing **120M+ records daily**, reducing data inconsistencies by **90%** and manual interventions by **60%**.
- Engineered scalable ETL/ELT pipelines with **Python**, **PySpark**, and **SQL**, achieving a **75% improvement** in data integration efficiency across Azure Synapse and Data Factory.

PROJECTS

Payer Policy Intelligence – Agentic Extraction Pipeline | Python, Groq, Pydantic, FastAPI

- Automated payer prior-authorization policy extraction from unstructured PDFs, **turning hours of manual analyst review per policy into seconds** via a production-grade multi-agent GenAI pipeline (~4.5K LOC) with 4 LLM agents (brand-check, extraction, validation, critique).
- Achieved **100% reproducible, byte-identical outputs** with a hallucination-resistant design where LLMs interpret policy language and deterministic Python owns all counting/scoring – eliminating the LLM inconsistency that breaks most GenAI extraction tools.
- Maximized extraction accuracy with 4 tunable accuracy tiers and **hybrid retrieval** (BM25 + dense embeddings, 5 backends); deployed **3 ways** (CLI, REST API, web) with full audit-trail logging for complete traceability.

Fabric Native RAG Chatbot | Python, FastAPI, Azure OpenAI, Azure AI Search, Microsoft Fabric, Redis

- Built a **retrieval-augmented chatbot** with a master-agent design, orchestrating downstream agents for SQL generation, execution, and response synthesis using **GPT-4o** for reasoning and section selection.
- Leveraged **Azure AI Search** for vector retrieval and a Fabric SQL endpoint for dynamic gold-layer queries; enabled secure, low-latency **LLM inference** via Azure Private Endpoints and Redis-backed caching.
- Built dynamic tool routing via a **master-agent model** with sub-agents for answer, graph, insight, and methodology generation and **semantic search** over unstructured documents.

AI Front-Desk Agent for Hotels | Python, Streamlit, RAG, Open-Source LLMs, Supabase, Vercel

- Built an **agentic front-desk assistant** with two roles – an admin console to upload hotel policy PDFs/docs for **RAG** ingestion, and a guest chat interface – backed by **Supabase** vector storage and **open-source LLMs**.
- Extended the agent beyond Q&A to **take real actions**: integrated a hotel management booking API so guests can create reservations directly from the chat, deployed on **Vercel**.

SKILLS

Generative AI & LLMs: Multi-Agent Orchestration, Agentic AI, LangGraph, RAG, Hybrid Retrieval, Prompt Engineering, Azure OpenAI, Open-Source LLMs (Llama 3), Semantic Search, Vector Retrieval

AI Platforms & Tooling: Azure AI Foundry, Azure AI Search, ChromaDB, sentence-transformers, Hugging Face, Pydantic

Programming: Python, SQL, FastAPI, C++, OOP, Data Structures & Algorithms

Data Engineering: Microsoft Fabric, PySpark, ETL/ELT, Delta Lake, Medallion Architecture, DQM, Lakehouse Architecture

Cloud & Deployment: Azure (Container Apps, Data Factory, Synapse, Redis), Docker, Vercel, Supabase, Streamlit

Tools & Visualization: Power BI, Plotly, Git, Azure DevOps, Postman

CERTIFICATIONS

- **Microsoft Certified: Fabric Data Engineer Associate** – Microsoft (Earned: April 2026) *Credential ID: 94CBB21B78A8E68E*

EDUCATION

- | | |
|--|---------------------------------------|
| Bharati Vidyapeeth College of Engineering
<i>B.Tech. Computer Science Engineering</i> | 2024
<i>New Delhi CGPA: 9.01</i> |
| Pusa Institute of Technology
<i>Diploma in Electronics and Communication Engineering</i> | 2021
<i>New Delhi CGPA: 8.8</i> |

ACHIEVEMENTS & PUBLICATIONS

- **Client Impact Champion Award (2025)** – ZS Associates, for architecting a Microsoft Fabric accelerator that reduced delivery timelines by 40%.
- **Research Excellence Recognition Award** – Dept. of CSE, BVCOE ● **Finalist**, Microsoft Engage 2022
- **Published:** Anomaly Detection in Heterogeneous Networks Using Graph Convolutional Networks & Non-Intrusive Machine Vision-Based Drowsy Driver Detection – IJFMR