Rushank Savant

| +91-9003517181| LinkedIn | GitHub | rushanksavant34@gmail.com | rushanksavant.github.io | dev.to |

AI Automation engineer with hands-on experience building RAG pipelines, and fine-tuned SLMs, agents, shipped with CI/CD, cloud deployment, and quantitative evals. Background in Python, data science, and Web3; previously managed a real-estate consultancy & investment firm.

SKILLS

Languages: Python, JavaScript

AI Automation: LangChain, LangGraph, LangSmith, LangFuse, RAG, RAGAS, QLoRA fine-tuning, BGE-M3, Unsloth, Ollama, MCP

Backend & DevOps: FastAPI, REST APIs, FastMCP, Docker, GitHub Actions, HuggingFace deployment

Databases & Storage: Qdrant (cloud + local), IPFS, vector databases (hybrid dense+sparse)

PROJECTS

SyncRAG - LangChain Docs Intelligent System (GitHub Repo) (Live Demo)

Python · LangGraph · FastAPI · FastMCP · Qdrant Cloud · Docker · GitHub Actions · RAGAS · LangSmith · Ollama

- Built an end-to-end RAG system over LangChain's official docs: hierarchical parent-child chunking, BGE-M3 hybrid (dense+sparse) embeddings, Qdrant Cloud indexing with RRF-based hybrid search, and a diff-based maintenance pipeline for continuous vector index sync without full re-indexing.
- Deployed an authenticated MCP server exposing retrieval as a reusable tool for developers, LLMs, and agentic workflows; extended by a LangGraph workflow for multi-step query planning, input filtering, and response synthesis, traced end-to-end with LangSmith.
- Engineered a RAGAS eval pipeline: knowledge-graph-based synthetic testset of 147 QA pairs (single + multi-hop) generated with Qwen 2.5 32B; achieved Faithfulness 0.79, Context Recall 0.77, scored using Command-R 35B as judge. Presented on a clean dashboard.
- Deployed on zero-cost infrastructure (HuggingFace Spaces) with CI/CD via GitHub Actions; Kaggle 2×T4 used to run synthetic testset generation and RAGAS evaluation.

Parametrix - Lightweight Financial PII Detection (GitHub Repo)

Python · QLoRA · Unsloth · PyTorch DDP · Kaggle 2×T4 · HuggingFace Hub

- Fine-tuned Llama 3.2 3B Instruct (nf4 quantized, LoRA rank=8) for financial PII detection across 29 entity types, addressing a niche but critical problem in finance where general-purpose LLMs are costly and over-engineered for structured entity extraction.
- Reduced training time by 50% using Unsloth on Kaggle 2 x T4 without quality loss; further optimized VRAM usage via gradient checkpointing, completing the full training run in ~4 hours on zero-cost hardware.
- Formulated a mathematical hardware estimation framework (VRAM/CPU RAM formulas for HuggingFace vs Unsloth) to predict training feasibility before running jobs, and a custom model testing framework to detect fine-tuned model limitations.
- Achieved 95.6% entity recall on 25,948 samples at 0.2% hallucination rate, demonstrating that a task-specific SLM can match general-purpose LLM accuracy at a fraction of inference cost; merged model published to HuggingFace Hub.

VastuLens - Floor Plan Vastu Auditor (GitHub Repo) (Live Demo)

Python · FastAPI · Gemini API · Fabric.js · Pydantic

- Identified the problem firsthand through architect and homebuyer interactions; built and deployed an AI solution for instant Vastu compliance checks on floor plan designs.
- Gemini auto-places fixture markers (entrance, stove, bed, etc.) on the floor plan; users drag to correct misdetections on a Fabric.js canvas, then a deterministic rule engine scores 8 fixture types across 9 directional zones with severity-graded findings and remedies.

Codebase Navigator - A multi-agent system (currently working)

Python · Ollama · LangGraph

Building an AI-powered developer assistant utilizing code-aware retrieval, program analysis, and agentic investigation for a persistent, auto-updating repository intelligence layer.

WORK EXPERIENCE

SRS Consultancy, Vadodara - *Managing Partner* JAN 2023 - AUG 2026

Directed operations across 50+ client relationships annually, overseeing investment portfolios and high-value real estate projects across multiple cities. Managed end-to-end legal documentation, dispute resolution, and cross-functional coordination with legal and financial advisors. (website)

InstaDapp, Remote - *Web3.0 Developer* JUN 2022 - AUG 2022

- Developed and deployed production smart contracts including a PoS Bridging Bot, DSA Manager Roles, and Long/Short trading contracts.
- Implemented formal verification of smart contracts using Certora Prover, ensuring correctness of critical on-chain logic.
- Built a Compound Resolver fetcher integrated with InstaDapp's DeFi infrastructure.

Nu10, Remote - *Blockchain Developer* FEB 2022 - MAY 2022

- Developed an ERC721-based NFT Marketplace smart contract with listing fees and royalties; integrated with a ReactJS frontend.
- Collaborated on casino payout infrastructure using Tatum payment APIs.

Other Experiences

Great Learning, Remote - *Data Science Intern* NOV 2021 - FEB 2022

Zebu Animation Studios, Remote - *Python Developer Intern* AUG 2021 - OCT 2021

EDUCATION

B.Tech, Electronics and Communication Engineering | VIT Vellore | 2018 - 2022 | CGPA: 8.01 | **Extra-curricular:** Technical Head, IET VIT