Yash Bhavsar

Gmail: yash.ynb@gmail.com | Mob.:+91 91678 81955

Professional Summary

AI-focused Software Engineer with hands-on experience designing, building, and deploying LLM-powered agents and RAG systems on Azure. Delivered multi-agent automation, prompt engineering, and model evaluation workflows for enterprise clients, partnering with stakeholders from requirements through production support. Proficient in Python, LangGraph, LangChain, Azure OpenAI, and Azure AI Foundry, with strong practice in monitoring, reliability, and secure deployment using Docker and Kubernetes. Effective at translating business workflows into AI solutions, documenting deliverables, and enabling adoption across technical and non-technical partners.

Skills

AI / Agents & LLMs: LLMs, Agentic AI, Multi-Agent Systems, RAG (incl. Graph RAG), Prompt Engineering, Model Evaluation, Guardrails, LangGraph, LangChain, Azure OpenAI, Azure AI Foundry, Microsoft Copilot, MCP Integrations, Intent Classification, NER

Languages & APIs: Python (primary), JavaScript, Shell, REST APIs, Flask, FastAPI

Automation & Integration: Workflow Automation, n8n, RPA, OCR, Process Optimization

Cloud, DevOps & SDLC: Azure, Google Cloud, Docker, Kubernetes, CI/CD Pipelines, Git, Linux, Agile/Scrum

Data & Monitoring: Vector DBs (Milvus, ChromaDB), PostgreSQL, MongoDB, MSSQL; Grafana, Prometheus

Professional Experience

GenAI Consultant

Independent project work and consulting October 2025 – August 2026

- Designed, built, and deployed Azure-based AI agent and RAG solutions using Python, Azure AI services, Docker, Kubernetes, and REST APIs to automate enterprise workflows.
- Partnered with client stakeholders to identify automation use cases, capture requirements, and translate business processes into functional AI solution designs.
- Implemented RAG and vector-search pipelines (document ingestion, embeddings, retrieval optimization) and multi-agent evaluation workflows to improve accuracy, reliability, and production readiness.
- Maintained production AI systems end-to-end: monitoring, troubleshooting integrations, performance tuning, and adherence to security constraints.
- Documented solution design and operational guidance to support handoff, governance-friendly delivery, and user adoption.

Software Engineer II (AI)

Indicus Software, Pune, India February 2025 – September 2025

- Developed agentic workflows and GenAI features for NeoPilot through the full feature lifecycle—requirements alignment, design, development, testing, and release—in agile sprints.
- Engineered a production LLM-RAG customer support agent with context management and prompt engineering for autonomous, context-aware query resolution.
- Integrated Python-based Model Context Protocol (MCP) client and server for secure, scalable tool and agent interactions; improved API responsiveness by ∼40% via Flask optimization.
- Led a cross-functional team of 4+ AI engineers, coordinating delivery, compliance for production AI systems, status updates, and on-time releases.
- Contributed to evaluation and reliability practices so agent outputs met quality bars before production rollout.

Software Engineer

vPhrase, Mumbai, India June 2019 – September 2022

- Designed and developed a portfolio restructuring web application for a corporate financial firm, scaling to 500,000+ active monthly users.
- Led a cross-functional team of 4 developers across multiple client projects, managing deliverables, stakeholder communication, and timely delivery.
- Implemented an automated data refresh pipeline processing 10+ GB of data daily using MSSQL and Python, reducing manual operational effort.
- Delivered a model deployment path for large ML models using TorchServe, cutting time-to-production by 3× and strengthening production SDLC discipline.

Key Projects

Multi-Agent Automation and Evaluation System Python, LangGraph, LangChain, OpenAI, n8n

- Built multi-agent workflows to automate code-review style processes, routing work across specialized agents for security checks, style compliance, and unit test generation.
- Integrated guardrails and evaluation checkpoints into agent loops to improve determinism, reduce hallucinated errors, and support reliable automation at scale.
- Designed workflows for maintainability and iterative improvement—monitoring failure modes and refining agent prompts and routing logic.

RAG-based Knowledge Assistant Python, LangChain, Flask, Milvus, Docker

- Designed and developed a customer support knowledge agent using LangChain and Flask, applying RAG to automate context-aware query resolution over enterprise content.
- Engineered data ingestion pipelines with optimized chunking, Milvus vector storage, and OpenAI embeddings; improved retrieval precision via query routing and semantic search tuning.
- Reduced response hallucinations through retrieval and prompt iteration; containerized services with Docker for reproducible dev-to-production deployment.

Azure GenAI Assistant Python, Azure OpenAI, Azure AI Foundry, Azure ML, Docker, FastAPI

- Designed an end-to-end GenAI assistant on Azure AI Foundry to automate enterprise workflows and data-oriented tasks from ideation through deployment.
- Implemented Prompt Flow evaluations to systematically refine LLM outputs for reliability, consistency, and enterprise-ready quality bars.
- Containerized a FastAPI orchestration layer with Docker, exposing secure API endpoints for parallel sessions and asynchronous LLM streaming; supported ongoing performance optimization.

RPA Travel Reimbursement Receipt Processing Automation Anywhere (A360), Excel, Email, OCR

- Built an unattended bot to process reimbursement emails, extract receipt fields (date, merchant, amount, currency) via OCR, and write structured claims data to Excel.
- Implemented validation rules for missing fields, duplicates, and amount thresholds, with exception queues and logging for controlled manual review—mirroring operational risk controls.
- Documented bot flow and reusable actions to support maintainability and handoff, aligned with enterprise automation lifecycle practices.

Education

Bachelor of Engineering in Computer Engineering August 2015–June 2019
KJ Somaiya Institute of Technology, Mumbai, India

Publication

- Web-Based Intelligent Traffic Management System for Varied Weather Conditions and Emergency Vehicles (2025)