

PARAG PATIL

Generative AI Engineer | Multi-Agent Systems | LangGraph | RAG | Production LLM Platforms

Pune, India | +91-7622044515 | official.paragp91199@gmail.com

linkedin.com/in/the-parag-patil | github.com/TheParagPatil | paragpatil.dev

PROFESSIONAL SUMMARY

Generative AI Engineer with 3+ years building and operating production-grade multi-agent systems, LLM applications, and enterprise RAG platforms. Architect ReAct and Plan-and-Execute agent workflows in LangGraph with supervisor-coordinated routing, explicit inter-agent handoffs, tool calling via MCP servers, and layered fallback logic. Design agent memory across both tiers - short-term conversational state and long-term cross-session vector memory. Ship on async FastAPI microservices with JWT auth, Redis caching, and streaming responses, serving 1,000+ users at sub-400ms latency. Own production readiness end-to-end: LLM-as-judge evaluation for accuracy and task success, AgentOps observability, and security guardrails against hallucination, prompt injection, and PII leakage. Currently building MKCL's internal self-hosted AI platform, delivering data-sovereign inference and reusable agent infrastructure to engineering teams. Lead a team of 4 and mentor engineers across the agentic AI stack.

PROFESSIONAL EXPERIENCE

Senior Project Associate II — MKCL (Maharashtra Knowledge Corporation Ltd.), Pune

May 2023 – Present

Delivered 5 production AI systems across edtech, agritech, and public-sector domains, owning agentic architecture, multi-model orchestration, evaluation, and secure API design.

- **Led a team of 4 engineers** delivering production GenAI and Agentic AI systems; owned technical design decisions, code reviews, sprint planning, and production deployments.
- **Translated business use cases** from internal and client-facing stakeholders into reliable agent workflows with defined handoffs, fallback paths, and safety boundaries — shipping 3 AI POCs and MVPs to production.
- **Mentored engineers** on LangGraph orchestration, RAG architecture, prompt engineering, agent evaluation, and AI security practices.

KEY PROJECTS

► Internal AI Platform — Self-Hosted LLM & Agent Infrastructure

Mar 2026 – Present | Ollama · vLLM · OpenWebUI · Qwen · Gemma · FastAPI · Guardrails · Evaluation

- **Architected MKCL's internal AI platform** to move LLM workloads onto **self-hosted infrastructure** — eliminating per-token API spend, meeting **data-sovereignty and compliance mandates** for sensitive institutional data, and bringing model choice, versioning, and availability under direct organizational control.
- Deployed **self-hosted inference serving on Ollama and vLLM** running Qwen and Gemma model families, with OpenWebUI as the operator-facing interface for internal teams.
- Built a **shared model gateway** exposing a unified API across backing models, so product teams integrate against one contract while models are swapped, versioned, or routed underneath.
- Packaged a **reusable agent and RAG framework** — orchestration scaffolding, retrieval pipelines, and memory components — enabling engineering teams to stand up new agentic services without rebuilding core infrastructure.
- Delivered **guardrails and evaluation as shared platform components** (prompt injection detection, PII masking, LLM-as-judge scoring), making safety and quality measurement a default rather than a per-project effort.

► IntelliDocs — Enterprise Knowledge Base (Agentic RAG Platform)

Jan 2026 – Mar 2026 | LangGraph · Multi-Agent · Plan-and-Execute · pgvector · ChromaDB · FastAPI · AgentOps

- Designed a **multi-agent orchestration layer in LangGraph** where a **supervisor node routes queries** to specialized retrieval and synthesis agents, coordinating handoffs across departmental knowledge domains.
- Implemented a **Plan-and-Execute workflow** that decomposes complex multi-part queries into an explicit step plan, executes each step against the appropriate retrieval strategy, and synthesizes a grounded answer.
- Engineered **hybrid retrieval** combining **dense vector search**, **BM25 lexical matching**, and **CrossEncoder reranking** over pgvector and ChromaDB, with multi-tenant isolation enforced at the retrieval boundary.
- Built **long-term vector memory** persisting user context across sessions, alongside **short-term conversational state** — so agents carry prior interactions forward instead of restarting cold.
- Instrumented **AgentOps observability** across agent runs, token consumption, latency distribution, and failure modes, surfacing them in a real-time monitoring dashboard for production operations.
- Established an **LLM-as-judge evaluation harness** scoring response **accuracy, groundedness, and task success**, converting agent quality from subjective review into a measured, trackable signal.
- Enforced **AI security guardrails** — input sanitization, prompt injection detection, and PII masking — to keep LLM usage compliant in enterprise environments.
- Developed an **async FastAPI backend** with JWT authentication, Redis caching, and streaming LLM responses; **reduced cost and latency by 40%**.

- Shipped an embeddable chatbot widget and Vue.js admin dashboard for plug-and-play enterprise rollout.

► **Krushi AI — Agricultural Advisory Agentic Platform**

Oct 2025 – Jan 2026 | LangGraph · LangChain · ReAct · MCP · Multilingual RAG · Whisper · FastAPI

- Built a **multi-agent advisory system coordinated by a LangGraph orchestrator**, with specialized agents owning soil data retrieval, crop recommendation, and government scheme lookup — each handing structured results back to the coordinator for synthesis.
- Implemented **ReAct-style tool-using agents** on LangChain/LangGraph agent runtimes, reasoning iteratively over tool outputs to resolve advisory queries rather than following a fixed call sequence.
- Exposed external agricultural APIs to the agent layer through **MCP server endpoints**, enabling safe, schema-validated **structured data lookups** alongside unstructured document retrieval.
- Designed a **RAG pipeline** with embedding-based retrieval, **fallback strategy on retrieval miss**, and source traceability (filename + chunk) so every recommendation is auditable to its origin.
- Applied **guardrails** including response validation and domain-scoped prompt templates to **prevent hallucinated guidance** on sensitive agronomic advice.
- Delivered **multilingual voice I/O** via Whisper STT/TTS across English, Hindi, and Marathi.

► **Kranti AI — Multilingual Voice & Text Agent**

Apr 2025 – Oct 2025 | LangChain · GPT-4 · FastAPI · Microservices · Whisper · Redis

- Built an enterprise **multilingual conversational agent** (English, Hindi, Marathi) on LangChain with **tool calling** and managed conversation memory.
- Architected **FastAPI microservices** sustaining **500+ concurrent users at <400ms response latency and 99%+ uptime**.
- Implemented **conversation memory with context compression**, **cutting repeated queries by ~30%** while controlling token cost per turn.
- Integrated real-time Whisper-based voice I/O achieving **<600ms end-to-end voice response**.

► **Campus Live — University Management System**

Jan 2024 – Dec 2024 | Go Lang · Vue.js · REST APIs · MySQL · RBAC

- Developed 5 core modules (admissions, exams, fees, reporting, notifications) deployed across **20+ institutions** in Maharashtra.
- Implemented **role-based access control** ensuring secure, auditable handling of academic data.

TECHNICAL SKILLS

Agentic AI	LangGraph, LangChain, Multi-Agent Systems, ReAct, Plan-and-Execute, Supervisor & Coordinator Agents, Router Agents, Agent Handoffs, Tool Calling / Function Calling, Workflow Orchestration, AgentOps, Google A2A (familiarity)
LLMs & Foundation Models	OpenAI GPT-4, Qwen, Gemma, Prompt Engineering (reasoning, task execution, tool use), Streaming Responses, Context Compression, Hugging Face
LLM Serving & Infrastructure	Ollama, vLLM, OpenWebUI, Self-Hosted Inference, Model Gateway / Routing
RAG & Retrieval	Hybrid Search, Dense Retrieval, BM25, CrossEncoder Reranking, Semantic Search, Chunking Strategies, Source Traceability, Structured & Unstructured Data Retrieval
Agent Memory	Short-Term Conversational State, Long-Term Vector Memory, Cross-Session Persistence, Context Compression
Vector & Databases	pgvector, ChromaDB, MySQL, MongoDB, Redis
AI Evaluation & Observability	LLM-as-Judge Evaluation, Accuracy & Task Success Scoring, Groundedness Evaluation, Monitoring, Logging, Latency Optimization, Cost Optimization
AI Security & Guardrails	Prompt Injection Detection, PII Masking, Input Sanitization, Response Validation, Hallucination Prevention, Domain-Scoped Prompting, JWT Auth, RBAC
Backend & APIs	Python, FastAPI, REST APIs, MCP Servers, Microservices, Async Programming, WebSocket, Error Handling, Retry & Fallback Logic, Golang
DevOps & Tooling	Docker, Linux, Nginx, Git, CI/CD, Postman
Frontend	Vue.js, JavaScript, HTML/CSS, Bootstrap

CERTIFICATIONS & TRAINING

CDAC — Advanced Computing (AIT YCP) | Sep 2022 – Mar 2023

Java, J2EE, React JS, Programming Fundamentals

Web Development Certification — Internshala

EDUCATION

Bachelor of Engineering | Jul 2017 – Jun 2022

Shram Sadhana Bombay Trust, Jalgaon