YASHASVI UDAYAN

AI Systems Architect | Agentic Workflows & Infrastructure

Lucknow, Uttar Pradesh | +91 9369380649 | yashasviudayan@gmail.com | LinkedIn | GitHub | X (Twitter)

EXPERIENCE

Independent AI Systems Developer *January 2026 – Present*

Self-Directed | Lucknow, India

- Designed and shipped four end-to-end AI infrastructure projects spanning multi-agent orchestration, RAG memory engines, autonomous DevOps, and concurrent web research — all on local compute.
- Mastered Apple Silicon (Metal) GPU acceleration to run 8B parameter LLMs locally, achieving $0 cloud API cost while maintaining production-grade performance.
- Established CI/CD pipelines via GitHub Actions to automate test suites and publish multi-platform Docker images; applied security-first credential redaction across all vector ingestion workflows.

TECHNICAL SKILLS

AI & Machine Learning: LangGraph, RAG, Vector Databases (ChromaDB), Agentic Orchestration, Local LLMs (Ollama, Llama-3-8B, Nomic), Prompt Engineering

Backend & Infrastructure: Python, FastAPI, Redis, Docker (Multi-platform), GitHub Actions (CI/CD), Pydantic v2

Hardware & Optimization: Apple Silicon (Metal) acceleration, Local inference optimization (24GB RAM / 16-core GPU tuning)

Tools: Git, GitHub CLI, Python Watchdog, SQLite, Pytest

AI ENGINEERING PROJECTS

The Orchestrator | *Python, LangGraph, Redis, Ollama, FastAPI* *Feb 2026 – Present*

- Architected a 100% local, multi-agent orchestration system using LangGraph to dynamically route complex software development tasks to specialized AI agents.
- Engineered a Redis-backed shared memory “blackboard” for sub-second data handoffs and context sharing between asynchronous agent nodes.
- Designed a Human-in-the-Loop (HITL) approval gateway with 4 risk classification levels, intercepting critical operations before deployment.
- Deployed an async FastAPI backend with SSE streaming to monitor system health and task iterations in real-time.

Context Core (Personal RAG Engine) | *ChromaDB, Docker, GitHub Actions, Pytest* *Feb 2026*

- Developed a local-first RAG knowledge base utilizing ChromaDB to persist long-term memory and contextual data across isolated AI workflows.
- Built event-driven ingestors using Python Watchdog to monitor and index terminal logs, clipboard history, and file system changes with near-zero latency.
- Implemented automated regex filtering to redact 15+ types of sensitive credentials (API keys, JWTs) prior to vector embedding.
- Achieved 100% test coverage and automated CI/CD pipelines to build and publish multi-platform (amd64/arm64) Docker images to Docker Hub.

PR-Agent (Autonomous Repo Maintainer) | *Llama 3, GitHub CLI, Docker, Webhooks* *Jan 2026*

- Built a webhook-driven DevOps pipeline to autonomously triage GitHub issues, scan repo architecture, and generate functional code fixes.
- Reduced cloud API expenditure to $0 by routing code reasoning and diff verifications to a local Llama-3-8B model on 16-core GPU unified memory.
- Automated Git state transitions, branch creation, and PR submissions via subprocess integration with the GitHub CLI.

The Autonomous Researcher | *Crawl4AI, DuckDuckGo API, Markdown* *Jan 2026*

- Engineered a concurrent web discovery agent integrating LangGraph and Crawl4AI for deep, multi-angle data extraction.
- Synthesized 276,000+ characters of unstructured web data into fully cited technical reports in under 15 seconds.

PROFESSIONAL SUMMARY

Innovative AI Systems Architect specializing in fully local, production-grade agentic infrastructure. Expert in multi-agent orchestration (LangGraph), RAG pipelines, and autonomous DevOps — all running on-device with zero cloud dependency. Combines deep backend engineering with hands-on LLM optimization to deliver intelligent systems that are fast, private, and cost-efficient.