# KARRI VAMSI KRISHNA

AI Engineer | Agentic AI · Forward-Deployed AI Engineering · LLM Applications · Production RAG

Hyderabad, India · +91 8074749058 · Vamsikv28@gmail.com · LinkedIn · github.com/kvcops · Portfolio

## SUMMARY

Forward-deployed **Agentic AI Engineer** (**1.5+ years**) who embeds with real workflows and takes AI pilots to production: multi-agent systems (LangGraph), tool-calling and browser agents, production RAG / Graph-RAG and FastAPI services. Replaced manual patent and legal workflows with agentic pipelines at **95%+ accuracy** and **90–95% lower cost**. **1st Place, IKDD Agentic AI Challenge (CODS 2025)**; open-source author on PyPI.

## SKILLS

**Agentic AI:** AI Agents, Multi-Agent Orchestration, LangGraph, LangChain, Agent State Machines, Tool / Function Calling, Model Context Protocol (MCP), Browser Agents (Browser Use, Playwright), Human-in-the-Loop, Agent Memory, Supervisor / Planner Patterns

**LLM Engineering:** Prompt & Context Engineering, Structured Outputs (JSON), Multi-LLM Routing & Failover, Token Streaming, Cost/Latency Optimization, LLM Evals & Observability (LangSmith); OpenAI, Anthropic Claude, Gemini, OpenRouter, Groq, Mistral, Ollama, Whisper

**RAG & Retrieval:** Production RAG, Graph-RAG (LightRAG, Neo4j), Hybrid Search (BM25 + Dense + RRF), Reranking, Coreference-Aware Chunking, Embeddings (Jina), Vector Search (FAISS), Citations / Grounding

**Document Intelligence:** Azure OCR, Docling, PP-DocLayout, Multimodal LLM Extraction, OpenCV, PDF Parsing, Layout & Coordinate Extraction

**Backend & Deployment:** Python, FastAPI, Flask, WebSockets, REST APIs, Async Pipelines, PostgreSQL, Neo4j, SQLite, Docker, Git, Selenium, Streamlit, Google Cloud, Firebase, Render, Vercel; also SQL, JavaScript, HTML/CSS, C++

## EXPERIENCE

**SciTech Patent Art Services Pvt. Ltd.** — *Junior Data Scientist* **Mar 2026 – Present** · Hyderabad

- **Patent extraction pipeline (US/EP/JP), 3–5x lower cost:** built OCR + non-OCR pipeline with Azure OCR, PP-DocLayout/Docling (CPU/GPU-optimized) and Gemini bounding boxes for image cropping and LaTeX formulas, with line-level coordinates and real-time WebSocket collaboration; made server-to-server transfer **fully async with auto-matching**, cutting manual matching work.
- **Chart & claim construction, 95%+ accuracy, 90–95% lower cost:** replaced a manual Adobe Acrobat workflow (30–40% accuracy, hours per document) with an automated pipeline that processes thousands of pages in minutes; built a CPU-only OpenCV rectangle-merge method to extract SMask-embedded images instead of a heavy layout model, plus multi-agent text-spillover detection and citation extraction.
- **Non-table claim chart pipeline at $0.80 / 1,000 pages** with exact per-line coordinate extraction, robust across widely varying layouts.
- **Legal RAG chatbot:** parallelized ingestion and leaner agent orchestration, reducing query latency and LLM token spend (FastAPI, PostgreSQL, LangChain/LangGraph).
- Built a Selenium webpage-to-PDF engine fixing broken Print-to-PDF images/layouts, and a video/GIF pipeline that dedupes keyframes (24/30 FPS) with Whisper transcripts + LLM summaries.

**Gyan Data Pvt. Ltd.** — *AI Software Engineer L1* **Aug 2025 – Mar 2026** · Chennai

- **Telemetry AI dashboard** for real-time turbine/industrial monitoring: autonomous agents for natural-language queries, streamed LLM tokens to cut perceived latency, live Plotly charts (Streamlit).
- **Graph-RAG chatbot backend** with FastAPI + LightRAG + Neo4j, using domain-constrained graph selection for retrieval; customized Frappe ERP with Playwright end-to-end tests.
- **DB-agnostic AI database chatbot** (LangChain/LangGraph, FastAPI, Google Auth) with pluggable inference across OpenRouter, Gemini, OpenAI and local Ollama; monitored and tuned with LangSmith.

**Gyan Data Pvt. Ltd.** — *Data Science Intern* **Jan 2025 – Jul 2025** · Chennai

- ERP customization and AI chatbot development; **converted to a full-time role** based on performance.

## PROJECTS

**ReferentWeave** | *Reference-aware RAG · Python, Rust, FastAPI, Gemini, Jina v5, BM25* **GitHub ↗**

- RAG framework that resolves cross-chunk references (e.g. "it" → entity) in one document-level pass **before** retrieval, without changing source text; hybrid dense + BM25 retrieval, RRF and Gemini rerank with chunk citations.
- **Recall@1: 63.3% vs 18.3%** for standard RAG (N=60, 20 docs) and 60% vs 33% (N=15); R@5 100%; ingestion under $0.01 per 100 pages.

**JobHunterX** | *Multi-Agent System · LangGraph, Browser Use, Playwright, FastAPI* **GitHub ↗**

- Autonomous **6-agent LangGraph pipeline**: parses a resume, plans searches, scrapes ATS boards (Greenhouse, Ashby, Lever), scores jobs and tailors a resume PDF per job (Jinja2 + xhtml2pdf).
- Stealth Playwright browser agent with human-in-the-loop takeover for CAPTCHA/login/MFA, streamed over CDP to a FastAPI + WebSocket dashboard; multi-LLM router (Gemini/Gemma, Groq Llama, Mistral) with failover and budget caps; **85 automated tests**.

**Atrophy** | *Open-source CLI · Python, SQLite* **GitHub ↗ · PyPI ↗ · Site ↗**

- CLI that separates human vs AI-generated git commits and tracks skill decay across 10 disciplines; owned the Python/SQLite data layer.

**AI Deep Research Agent** | *Python, Gemini API, Web Scraping, ReportLab · 29 stars* **GitHub ↗ · Demo ↗**

- Research agent: plans queries, scrapes multiple engines concurrently, synthesizes with Gemini and auto-generates PDF reports.

**NeurIQ AI** | *Python, FastAPI, Gemini API, PostgreSQL* **Live ↗**

- AI tutor platform with dynamic roadmaps, adaptive learning engine, SRS flashcards and context-aware AI explanations.

## EDUCATION

**B.Tech, Computer Science & Engineering** — Gayatri Vidya Parishad College of Engineering (A) 2021 – 2025 · **CGPA 8.97**

**Intermediate (MPC)** — Sri Chaitanya Junior College, Steel Plant 2019 – 2021 · **97.7%**

**SSC** — Sri Chaitanya EM School, Anakapalle 2019 · **GPA 9.8**

## ACHIEVEMENTS & CERTIFICATIONS

- **1st Place — IKDD Agentic AI Challenge (AssetOpsBench), CODS 2025**, IISER Pune: multi-agent predictive-maintenance AI solution.
- **Hackathons & Leadership:** 2nd Round, Adobe GenSolve Hackathon · IBC National Hackathon · Organizer, IEEE Mystical Code.
- **Certifications:** CCSK Cloud Security (Cohort-5) · Google Cloud Gen AI · Juniper Networks · Infosys AI · Salesforce Developer.
- **Languages:** English, Telugu, Hindi · **Interests:** Competitive Programming, Chess, Photography.