

PAVUN KUMAR R

AI Engineer · Agentic AI Developer · Full Stack Developer

Chennai, India · rpavun57@gmail.com · +91-7338996189 · [GitHub](#) · [LinkedIn](#) · [Portfolio](#)

PROFESSIONAL SUMMARY

AI Engineer and Full Stack Developer with 1.5+ years of hands-on experience designing and shipping production-grade Retrieval-Augmented Generation (RAG) systems, multi-agent AI workflows, voice AI applications, and cloud-native full stack platforms. Reduced document retrieval latency by 42% and improved semantic search accuracy by 38% through optimized vector database pipelines and agent orchestration. Proficient in LangChain, Agno, OpenAI APIs, PostgreSQL pgvector, FastAPI, Docker, and CI/CD automation. Active open-source contributor and LLM security researcher specializing in Agentic AI architecture and Model Context Protocol (MCP) integration.

TECHNICAL SKILLS

AI & Agentic Systems: LangChain, Agno, OpenAI API, Retrieval-Augmented Generation (RAG), Multi-Agent Systems, Model Context Protocol (MCP), Prompt Engineering, Semantic Search, Vector Databases, pgvector, LLM Security, AI Agent Orchestration
Languages: Python, JavaScript, TypeScript, SQL
Backend: Node.js, Express.js, FastAPI, Firebase, PostgreSQL, MongoDB, REST APIs
Frontend: React.js, Next.js, Vue.js, Tailwind CSS, HTML5, CSS3
Cloud & DevOps: Docker, GitHub Actions, CI/CD Pipelines, Google Cloud Platform (GCP), AWS, Microsoft Azure, Vercel, Cloudflare
Automation & Tooling: n8n, Make.com, Zapier · AI Coding: Cursor, Windsurf, Cline, Rocode, Augment

PROFESSIONAL EXPERIENCE

- Associate Innovation Developer · ThoughtSeed – Bangalore (Hybrid)

Jul 2025 – Present · 1 yr

 - Engineered production-ready multi-agent RAG systems using LangChain and Agno, improving document retrieval accuracy by 38% over keyword-based search through optimized vector embedding and semantic ranking strategies.
 - Designed and deployed Model Context Protocol (MCP) servers enabling AI agents to connect with 5+ external tools and APIs, reducing agent integration setup time by 65% compared to manual API wiring.
 - Automated AI microservice deployments via Docker and GitHub Actions CI/CD pipelines on GCP, cutting release cycle time from 3 days to under 6 hours and achieving 99.4% deployment success rate.
 - Led architecture decisions for LLM-powered applications handling 8,000+ document queries per week, maintaining sub-2 second average response time through retrieval pipeline optimization.
 - Collaborated with cross-functional engineering teams to deliver 4 AI products from prototype to production, improving sprint delivery velocity by 28% through AI-assisted development workflows (Cursor, Windsurf).
- Junior Innovation Developer · ThoughtSeed – Bangalore (Hybrid)

Mar 2025 – Jul 2025 · 5 mos

 - Built AI-powered document processing and semantic search pipelines using LangChain, cutting unstructured data retrieval time by 52% compared to prior manual extraction processes.
 - Developed data extraction and enrichment pipelines processing 30,000+ records per week, feeding structured content into downstream AI retrieval and classification systems.
 - Shipped 3 cloud-native features end-to-end API design through CI/CD deployment using GitHub Actions, Docker, and Vercel, reducing time-to-deploy by 45%.
 - Leveraged AI coding tools (Cursor, Rocode) to accelerate feature development cycles by approximately 30% while maintaining team code quality and review standards.
- Full Stack Developer Intern · ThoughtSeed – Remote

Dec 2024 – Mar 2025 · 4 mos

 - Built RESTful backend services with Node.js, Express.js, and PostgreSQL supporting 3 client-facing applications; implemented query indexing that improved database response time by 30%.
 - Containerized 4 services using Docker and configured automated cloud deployment pipelines, reducing environment-related bugs by 40% and standardizing staging-to-production workflows.
 - Designed and documented REST APIs consumed by both web and mobile clients, handling 1,200+ daily requests across internal tools and external integrations.
- AI & Data Engineer Intern · WhatsLegal.ai – Remote (Germany)

Sep 2024 – Dec 2024 · 4 mos

 - Developed a legal document Q&A system using LangChain and semantic retrieval over 100,000+ Indian court judgments, achieving 85% answer relevance score on internal evaluation benchmarks.
 - Built NLP pipelines for structured extraction of case metadata parties, dates, rulings, citations reducing manual legal research time by approximately 60% for the core user workflow.
 - Engineered automated data ingestion and preprocessing workflows that cut document onboarding time from 4 hours to under 30 minutes per batch.

FEATURED PROJECTS

EmployeeAI — Open Source Multi-Agent AI Platform · *LangChain · Agno · MCP · Python · Agent Orchestration* [GitHub →](#)

- Building an open-source AI employee platform with specialized agents for autonomous task execution, workflow orchestration, tool use, and knowledge retrieval architected as a modular digital workforce for businesses and startups.

Tuya Documentation Search System · *Agno · PostgreSQL pgvector · FastAPI · Semantic Search · Vector Databases* [GitHub →](#)

- Multi-agent documentation intelligence system with automated PDF ingestion, chunking, embedding generation, and vector search over large knowledge bases; reduced documentation lookup time by 70% vs manual search in testing. Built interactive FastAPI playground with persistent session management.

AI Voice Agents — Lead Qualification System · *OpenAI · Vapi · Dynamic Call Flows · REST APIs*

- Deployed conversational voice agents for lead qualification and customer interaction, integrating real-time business logic and external APIs; reduced manual outreach effort by 35% across tested call workflows.

Legal Judgment Assistant — Smart India Hackathon 2024 Finalist · *LangChain · Semantic Search · NLP · Python*

- Semantic search platform for 100,000+ Indian legal judgments with natural language querying and intent-based document routing; selected as college-level finalist among 200+ competing teams in the Legal Technology track.

EDUCATION

B.E. Computer Science and Design – Sethu Institute of Technology, Tamil Nadu *2023 – 2027 (Ongoing)*

ACHIEVEMENTS & OPEN SOURCE

- Smart India Hackathon 2024 — College-Level Finalist, Legal Technology track, selected among 200+ teams competing nationally.
- Open-source contributor to Triage AI — AI-powered VS Code developer assistant with modular agents for project management, testing, and security workflows. (GitHub: github.com/Pavun57/Triage-AI)
- Active LLM Security Researcher — studying prompt injection attacks, jailbreak vectors, and production-safe AI agent defense mechanisms for enterprise-grade systems.
- Shipped 5+ production AI systems across legal-tech, enterprise documentation, voice AI, and automation domains within 18 months of professional experience.