

AKSHIT PARETA

Forward Deployed AI Engineer | LLMs, RAG, Multi-Agent Systems & Production AI Deployment

Bundi, Rajasthan, India • akshitpareta01@gmail.com • +91-9680810188

akshitpareta.com • github.com/akshitpareta • linkedin.com/in/akshitpareta

SUMMARY

Forward Deployed AI Engineer who builds and ships — not demos, but production systems businesses run on. 4+ years deploying AI across 50+ client engagements in 8 industries, owning the full discovery-to-deployment lifecycle: scoping, POC, pilot, production, post-launch iteration. Shipped production RAG platforms, 50-agent orchestration systems, and on-premise LLM clusters on self-hosted GPUs. Commerce-educated, self-taught; core strength is scoping ambiguous business problems and shipping systems clients actually use.

TECHNICAL SKILLS

Languages: Python, JavaScript, TypeScript, SQL

AI/ML: LLMs, RAG, Multi-Agent Orchestration, Embeddings, Prompt Engineering, QLoRA Fine-Tuning, Quantization (GGUF/AWQ), Tool/Function Calling, LangChain, LangGraph, Vector Search, Evals (Golden Sets, LLM-as-Judge, Regression)

Inference & Full-Stack: vLLM, Ollama, HuggingFace, ComfyUI, LangSmith, Langfuse; React, Next.js, Node.js, FastAPI, Flutter, Tailwind

Cloud/Infra & Data: AWS, Docker, Terraform, Linux, Nginx, CI/CD, System Design, Multi-Tenant RBAC, Data Isolation; PostgreSQL, MongoDB, ChromaDB, Qdrant, Redis, SQLite

Deployment Lifecycle: Discovery, Scoping, POC, Pilot, Production, Stakeholder Onboarding, Post-Launch Iteration

Tools: Claude API, GPT API, Gemini API, N8N, Zapier, Git, Figma

PROFESSIONAL EXPERIENCE

Engineering Lead — AI Systems

Gosotek Technology Pvt Ltd, Jaipur • Jul 2025 – Present (promoted to lead in 4 months)

- Hands-on owner of all AI architecture and production deployments while leading a 20-person delivery team (3 devs, 3 designers, 2 ops, 1 AI) — shipped 28 production websites and 2 mobile apps.
- Deployed a multi-layer RAG legal-research platform (Gemini Embeddings + ChromaDB) for a 6-founder law firm — ran discovery calls, designed retrieval, shipped to production with WhatsApp plugin + analytics dashboard; saved 15–30 hrs/week per attorney, +50% case capacity.
- Deployed a multi-tenant RAG ops platform for a conglomerate of 6–8 companies — per-company data isolation, multi-tenant RBAC, shared memory; onboarded founders + management and iterated daily.
- Architected OpenSalesAI (LangGraph multi-agent) for an FMCG distributor across 50–500 retail outlets — supervisor routes to 5 specialist agents over Qdrant + LangChain RAG with 3-tier LLM fallback.
- Built a golden-set eval harness (340 cases, LLM-as-judge, LangSmith) — drove citation precision 71% to 89% via hybrid search while tracking hallucination rate across prompt/model changes.
- Scaled MyKiranaBuddy to 5,000+ users and Rs.50L revenue at 70–80% MoM growth — cut logistics 40% via route optimization, +60% upsell with in-house field-sales app (GPS, dynamic commissions, pay-later EMI).
- Operate a self-hosted inference cluster (2× RTX 5090, 64GB VRAM) serving quantized Qwen/GLM via vLLM with QLoRA pipelines — eliminated ~Rs.40K/month in cloud GPU cost.

Freelance Software Engineer & AI Consultant

Self-Employed | India & US Clients • 2021 – Jul 2025

- One-person forward-deployed engineering team for 50+ clients across 8 industries — owned discovery to deploy; delivered 28 production sites + 40+ workflow automations; advised US startups on GTM/pitch decks.

KEY DEPLOYMENTS

Legal LLM — On-Premise Multi-Tenant RAG + WhatsApp & ERP

Client: Multinational Law Firm • 4× RTX 5090 • 1TB NVMe • Frappe ERP • Production

- Privacy-first legal AI on client's own 4× RTX 5090 cluster — data isolated on-client, surfaced in attorneys' daily WhatsApp via @mention; RAG agent connected to Frappe ERP for operational actions (case status, doc retrieval, scheduling) — saved 15–30 hrs/week, +50% case capacity.

ReVideoz — Social Media Intelligence & Content Recreation

Client: Marketing Agency • Gemini API • Apify • Parallel Agents • Production

- Multi-platform intelligence across Instagram/X/TikTok/YouTube/Reddit/Facebook; parallel scraping engine (8–13 concurrent agents, multi-provider key rotation, multimodal Gemini on 3-tier cost model) — 95% content-similarity recreation via Higgsfield/Kling.

Garment Manufacturing — Full AI Digital Transformation

Client: Clothing Manufacturer • Frappe ERP • 11Labs Voice • Claude Code • Production

- Full digital transformation for a non-tech manufacturer — WhatsApp + in-app chatbot, 11Labs voice agent, Frappe ERP with NL AI assistant, Claude Code staff training; cut transcription errors 30–40% to 2–3% via Gemini transcription layer; multi-language role-based access.

Finance Client — Voice AI Onboarding & Fraud Prevention

Client: Financial Services • Grok API Voice • GPS Scoring • Production

- Voice onboarding (GPS scoring, voice-to-voice) for Tier 3–4 users — cut onboarding 15–20 min to 2–3 min, fake profiles 30–40% to 5–10%; Grok voice-auditing agent with 2FA cut fraud 80% and scaled loan disbursement Rs.10L to Rs.20–30L/month.

EDUCATION & CERTIFICATIONS & LANGUAGES

M.Com, Business Administration — Kota University (2023) • B.Com, Business Management — Kota University (2021). Software engineering: self-taught — production AI systems, multi-agent orchestration, GPU fine-tuning.

Certs: IBM Product Management • Zapier Expert • Google Analytics. Languages: English (Professional) • Hindi (Native).