AKSHIT PARETA

AI Engineer | LLM Applications, RAG & Multi-Agent Systems | Self-Hosted Inference Infrastructure

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

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

SUMMARY

AI Engineer who ships production LLM systems, not demos. Most candidates stop at API calls — I work below the API layer: fine-tuning Qwen-72B with QLoRA, serving it quantized (INT4) on a self-hosted 2× RTX 5090 cluster via vLLM at 30–40 tokens/sec (p95 first-token latency <1s), and gating every deploy behind a 340-case golden-set eval. Shipped a multi-tenant legal RAG platform (15K docs, citation precision 71% to 89%, hallucination ~2–3%), a LangGraph multi-agent sales system, and a 50-agent orchestration harness. I own the metal, the evals, and the outcomes.

TECHNICAL SKILLS

Models & SDKs: OpenAI, Claude 3.5, Gemini, Qwen, Llama, Kimi, GLM — multi-provider routing, fallback chains

LLM Applications: RAG, Multi-Agent Orchestration, LangChain, LangGraph, Prompt Engineering, Function Calling, Tool Use, Structured Outputs, Streaming

Retrieval: ChromaDB, Qdrant, pgvector, Gemini Embeddings, sentence-transformers, Hybrid Search (BM25 + dense), Cross-encoder Reranking

Fine-tuning & Serving: QLoRA, PEFT, HuggingFace TRL/Accelerate, bitsandbytes, PyTorch, GGUF/AWQ/INT4; vLLM (continuous batching, paged KV cache), Ollama

Evals & Observability: Golden-Set CI, LLM-as-Judge, RAGAS, Citation Grounding, Hallucination Detection, Regression Testing, LangSmith, Langfuse, MLflow

Agents & Backend: Model Context Protocol (MCP), OpenAI Agents SDK, Anthropic Agent SDK, Orchestrator-Worker, ReAct; Python, TypeScript, SQL, FastAPI, Node.js, React, Next.js

Infra & Data: Docker, AWS, Linux, Nginx, GitHub Actions, Prometheus; PostgreSQL, MongoDB, Redis, scikit-learn, XGBoost, Google OR-Tools

PROFESSIONAL EXPERIENCE

Engineering Lead — AI Systems

Gosotek Technology Pvt Ltd, Jaipur • Jul 2025 – Present • Promoted in 4 months

- Shipped a multi-tenant legal RAG platform (Gemini Embeddings + ChromaDB, ~15K docs, ~75K chunks) with hybrid retrieval (BM25 + dense) and cross-encoder reranking — saving 15–30 hrs/week per attorney, +50% case capacity.
- Built the eval that gates every deploy: 340-case golden-set harness, LLM-as-judge cross-validated against human attorney review + LangSmith tracing — citation precision 71% to 89%, hallucination held at ~2–3%.
- Operate a self-hosted inference cluster (2× RTX 5090, 64GB GDDR7) serving quantized Qwen-72B (INT4, ~40GB VRAM) via vLLM — 30–40 tokens/sec, p95 <1s, ~99% uptime, ~Rs.40K/month cloud avoided.
- Fine-tuned Qwen-72B with QLoRA (Transformers, PEFT, TRL, bitsandbytes) on ~2K labeled legal/medical examples — ~10% accuracy lift on domain eval, 2–6 hrs per run.
- Led a 20-person org shipping 28 production sites and 2 mobile apps (5,000+ users, Rs.50L revenue).

Freelance AI Engineer & Consultant

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

- Delivered production AI + automation for 50+ clients across 8 industries; built 40+ N8N/Zapier workflows and 78 reusable templates; advised US startups on AI strategy and product positioning.

FEATURED AI PROJECTS

Legal LLM Platform — Multi-Tenant RAG System

Gemini Embeddings • ChromaDB • Hybrid Search • Cross-encoder Reranking • vLLM

- Multi-layer RAG over ~15K legal docs (~75K chunks), hierarchical chunking by statute/section, served on self-hosted Qwen-72B (INT4) — zero cloud inference cost; 340-case golden-set eval (LLM-as-judge, LangSmith regression gating) drove citation precision 71% to 89%.
- Full per-company data isolation across 6–8 tenants (zero leakage); LLM function calling + real-time web search ground answers in current case law; deployed via web dashboard + WhatsApp plugin.

OpenSalesAI — LangGraph Multi-Agent Sales Platform

LangGraph • Qdrant • FastAPI • Pydantic • Streaming

- LangGraph StateGraph supervisor routes to 5 specialist agents (analytics, coach, collection, order, promo) — in production for FMCG/CPG route-to-market teams; TypeScript dashboard + Python FastAPI service.
- Qdrant vector retrieval, Pydantic structured outputs, streaming, per-agent context budgeting, 3-tier LLM fallback; tracked with MLflow + Prometheus; intent-classifier validated on golden set before each change.

Swarm OS — Multi-Agent Harness

Claude • Gemini • GLM • Kimi • GPT • Model Context Protocol (MCP)

- Hierarchical orchestrator-worker topology coordinating 50+ concurrent sub-agents across 5 frontier models; scoped scratchpad views + per-agent token budgets prevent the cost explosion naive 50-agent sharing causes.

- Max-iteration caps + dependency-cycle detection stop agent loops — runs unattended; self-improving instruction layer cut multi-step task latency ~40% vs single-agent baseline.

Self-Hosted LLM Inference & Training Cluster

2× RTX 5090 (64GB GDDR7) • vLLM • Ollama • HuggingFace QLoRA • ComfyUI

- Tiered stack — vLLM for production inference (continuous batching, paged KV cache, 30–40 tok/s, p95 <1s, ~99% uptime), Ollama for dev, autoscaling router bursts heavy loads to the dual-GPU node; QLoRA pipelines specialize Qwen-72B for legal/medical (~10% lift); ComfyUI media pipelines.

EDUCATION & CERTIFICATIONS & LANGUAGES

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

Certs: IBM Product Management • Zapier Expert • Google Analytics • AWS AI Practitioner (in progress). Languages: English (Professional) • Hindi (Native).