

Gagan Jain

AI / LLM Engineer

Jaipur, India (Open to Relocation) | +91 95872 55792 | gagan.jain.se@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

AI Engineer focused on production-grade GenAI systems: agentic AI orchestration, RAG pipelines, LLM fine-tuning (QLoRA/LoRA), MCP tooling, and local-first LLM deployment. Author of NexusAOS, a governance-first, event-sourced AI operating environment written in Rust (12 workspace crates, 981 tests, full CI/CD) with provider-agnostic LLM streaming (OpenAI/Anthropic), policy-enforced agent actions, and terminal/SSH integration. Strong Python + Rust foundation with end-to-end deployment: Docker, GitHub Actions, observability, and AI governance.

TECHNICAL SKILLS

AI / LLM: LLM Integration, RAG Pipelines, Agentic AI Frameworks, Prompt Engineering, Context Engineering, Multi-Agent Systems, Tool-Using Agents, Fine-tuning (QLoRA/LoRA), MCP / Tool Protocols, LangChain / LlamaIndex, Vector Databases, AI Governance

Languages: Python, TypeScript, JavaScript, Rust, C++, C, PHP, HTML/CSS

Databases & DevOps: MySQL, MongoDB, Redis, SQLite, Supabase, Docker, GitHub Actions, Nginx, Unicorn, Prometheus, Linux

LLM Ops: Evaluation & Observability, Local LLM Deployment, OpenAI / Anthropic APIs, Embeddings, Structured Outputs / Function Calling, LiteLLM / Model Routing, Streaming Responses, NLP, Ollama

Backend & APIs: Flask, FastAPI, Node.js, Express, Laravel, Axum (Rust), REST APIs, WebSockets

Frontend & Vision: React, React Native, Next.js, Tailwind CSS, Bootstrap, Monaco Editor, YOLO, OpenCV, Computer Vision

PROJECTS

NexusAOS

AI/AGENTIC

Rust, Tokio, Event Sourcing, LiteLLM, OpenAI, Anthropic, SSH, PTY, TUI, GUI

- Designed a governance-first, event-sourced AI OS in Rust: 12 workspace crates, 981 passing tests, 0 clippy warnings, full GitHub Actions CI/CD
- Built a policy-enforced agent kernel where LLMs propose actions and the kernel validates and records every state change in an append-only audit trail
- Implemented a provider-agnostic LLM layer with OpenAI-compatible and Anthropic streaming, LiteLLM routing, and local-first inference that runs fully offline
- Delivered native terminal emulation (PTY + VT100), SSH multiplexing, a secrets vault, and four interfaces: CLI, TUI, GUI, RPC

[Source Code](#)

[Documentation](#)

nexus-kernel

RUST

Rust, Tokio, OpenAI, Anthropic, SQLite, Ratatui, Iced, SSH, Event Sourcing

- Companion to NexusAOS: shares the same 12-crate Rust workspace and its 981-test suite (tracked under NexusAOS)
- Implemented event-sourced append-only audit trail, policy engine with trust tiers, and provider-swappable model interface
- Integrated OpenAI-compatible and Anthropic streaming with real-time token streaming into TUI/GUI
- Delivered native terminal emulation (PTY + VT100 + Zig parser), SSH multiplexing, and multi-interface CLI/TUI/GUI/RPC

[Source Code](#)

[Documentation](#)

SeshaOS

AI/AGENTIC

Rust, LiteLLM, Gemma, Qwen, Local LLMs, Governance, Event Sourcing, Ubuntu

- Architected specialist model stack: Gemma 4 12B (Planner), Qwen3-Coder 30B (Implementation), Qwen3.5 9B (Vision)
- Designed kernel-centric governance where models propose actions and the kernel validates/records every state change
- Built a LiteLLM-compatible proxy for NVIDIA NIM and other model providers
- Maintained event-sourced architecture with reversible, permissioned actions and offline-first execution

[Source Code](#)

[Documentation](#)

Vyākṛti

COMPILER

Rust, React, TypeScript, Monaco Editor, Zustand, Tailwind CSS, Axum

- Complete compiler pipeline: lexer → parser → type checker → bytecode compiler — all built from scratch in Rust
- Browser-based IDE with React, Monaco Editor, syntax highlighting, autocomplete, and diagnostics
- Rust (axum) backend with compile, REPL, LSP, and file management endpoints via REST + WebSocket
- 123 tests covering the full pipeline, including a self-hosting corpus

[🔗 Source Code](#)
[📖 Documentation](#)

RAG Service GEN AI

Python, FastAPI, ChromaDB, Embeddings, Hybrid Search, RAG, LLM-as-Judge, Docker

- Built a FastAPI RAG service with hybrid retrieval (dense embeddings + BM25 keyword, merged via Reciprocal Rank Fusion) over ChromaDB
- Implemented chunking strategies (recursive, semantic), embedding pipelines, and an /ask endpoint that streams grounded answers with citations
- Shipped LLM-as-judge evaluation harness (faithfulness, answer relevance, correctness) with golden-set CI gate and RAGAS-style reports
- Containerized with Docker Compose; OpenAI-compatible LLM interface with local-first fallback

[🔗 Source Code](#)
[📖 Documentation](#)

LLM Eval Harness GEN AI

Python, LLM-as-Judge, Golden Sets, RAGAS, CLI, CI/CD, Pytest

- Built a golden-set-driven evaluation harness with faithfulness, answer-relevance, and correctness metrics (LLM-as-judge + lexical fallbacks)
- Added offline heuristic scorers so evals run in CI without API keys; JSON + Markdown reports for regression tracking
- Designed golden-set YAML format (question, golden answer, context) and a CLI: `python -m eval_harness --golden-set ...`
- Containerized and CI-ready; used to gate the RAG service pipeline

[🔗 Source Code](#)
[📖 Documentation](#)

AIM — Attendance Information Manager PRODUCTION-READY

Python, Flask, MySQL, Bootstrap, Chart.js, Docker, GitHub Actions

- Production-grade Flask + MySQL platform with Argon2id auth, CSRF protection, brute-force lockout, breached-password scanning, JWT sessions, and strict CSP/HSTS headers
- Deployed Prometheus metrics, structured JSON logging, Chart.js analytics, FullCalendar scheduling, Docker Compose, and GitHub Actions CI/CD; 101 automated pytest tests

[🔗 Source Code](#)
[📖 Documentation](#)

FWRS — Food Waste Reduction System PRODUCTION-READY

Python, PuLP, Folium, Flask, MySQL, Bootstrap

- Built a 3-stage lexicographic LP solver (fairness → priority → cost) to allocate surplus food to NGOs while minimizing waste
- Added expiry-aware routing that penalizes allocations where travel time exceeds food shelf life
- Integrated Folium/Leaflet interactive maps with animated routes, heatmaps, and priority-colored markers
- Shipped a Flask web dashboard plus a desktop Tkinter GUI with charts and CSV export

[🔗 Source Code](#)
[📖 Documentation](#)

EXPERIENCE

Industry 5.0 Industrial Automation Trainee

Aug 2025 – Nov 2025

CodenPlay Robotics

- Built and debugged PLC automation logic (ladder logic, structured text) in CODESYS
- Simulated industrial process workflows in Factory I/O
- Implemented flow-based control/integration in Node-RED
- Covered HMI design, industrial communication, process monitoring, and the debugging→testing→deployment workflow

EDUCATION

Vellore Institute of Technology, Vellore

Graduated 2025

B.Tech in Computer Science and Engineering

CGPA: 7.7 / 10.0 | Relevant Coursework: Data Structures & Algorithms, Operating Systems, Database Systems, Computer Networks, Web Technologies, Artificial Intelligence, Machine Learning