

HIMANSH RAJ

Applied AI Engineer

LLM Gateways • Agent Orchestration • Inference Systems

iamthehimansh@gmail.com | himansh.in | github.com/iamthehimansh | linkedin.com/in/iamthehimansh | Noida, India

SUMMARY

AI engineer who ships LLM infrastructure. Author of **Maya OS** ([npm i -g maya-os](https://npm.im)): a local-first AI operating system with an OpenAI/Anthropic-compatible model gateway, multi-agent orchestration over real PTYs, and local GGUF inference on Apple Silicon. At Colate, built the multi-provider gateway and agent runtimes behind Cana, serving live traffic on AWS/Kubernetes: **6.05B cached input tokens, 53% input-token reuse** and **35% lower inference latency**. Comfortable from llama.cpp and RAM guardrails at the bottom to FastAPI services and dashboards at the top.

SHIPPED PRODUCTS

Maya OS: local-first AI operating system [npm: maya-os](https://npm.im) • maya.himansh.in • TypeScript • MIT

- **Model gateway** exposing OpenAI- and Anthropic-compatible APIs, routing one request surface across cloud providers *and* local GGUF models (llama.cpp/Metal) with **RAM guardrails** that degrade instead of crashing the host.
- **Agent orchestration over real PTYs**. Claude Code, Codex and `maya` code run as genuine terminal processes, coordinated as teams by a persistent loop agent.
- **Message bus** + SQLite-backed contracts for durable multi-agent state, across an **8-package monorepo** (model-runtime, gateway, agent-runtime, bus, orchestrator, core, CLI, dashboard) shipped as an installable CLI + web dashboard.

maya code (`mayac`): the coding agent [@maya-os/code](https://npm.im) • 4 releases

- Provider-agnostic **agent loop** over the Anthropic and OpenAI SDKs, supporting both API-key and subscription auth.
- Full-screen **Ink TUI**, skills system, **MCP** integration, sub-agents, multi-stage workflows, and mid-turn steering of a running agent.

StudyGit: VS Code extension Published on the VS Code Marketplace. Interactive commit playback with character-level diffing.

EXPERIENCE

Applied AI Engineer | Colate

Jul 2025 to Dec 2025

Remote

- Developed a **multi-provider LLM gateway** on a shared request/event representation that translates between OpenAI, Anthropic and Gemini formats, covering streaming, tool calls and multimodal inputs.
- Implemented **prompt caching and token-usage accounting**: the internal dashboard reported **6.05B cached input tokens (53% input-token reuse)** and **35% lower inference latency** from caching and runtime optimisations.
- Built **multi-agent orchestration** for Cana with worker, reviewer and lead roles, dependency-aware delegation, inter-agent messaging and completion handoffs for long-running tasks.
- Engineered a **vector database and semantic code-search engine**, indexing repository-scale corpora for sub-second retrieval.
- Shipped **persistent AWS/Kubernetes runtimes** with session recovery so long-running agent jobs survive restarts without losing context; evaluation harness across **90+ test prompts** caught robustness regressions before release.

Research Intern | MISN Lab, IIT Delhi

May 2026 to Present

Advisor: Dr. Sandeep Kumar

- Building **LLM agents for recommendation** that reason over user context, combining retrieval, planning and tool use to beat collaborative-filtering and embedding-similarity baselines.
- Designed the **evaluation methodology** for agentic recommenders: failure-mode characterisation and robustness analysis under distribution shift.

TECHNICAL SKILLS

LLM & Agents	model gateways • OpenAI/Anthropic-compatible APIs • MCP • multi-agent orchestration • agentic tool use • RAG & dense retrieval • vector databases • embeddings • streaming inference • LLM/SLM fine-tuning • evaluation harnesses
Inference/ML	llama.cpp • GGUF • Metal / Apple Silicon • local inference • PyTorch • Hugging Face • Unsloth • TRL • quantisation • latency profiling
Infrastructure	Python • TypeScript • FastAPI • Node • Docker • Kubernetes • AWS • Redis • SQLite • CI/CD • PTY orchestration • C • C++ • Kotlin

SELECTED ENGINEERING PROJECTS

- **boxctl-deck**. GPU workstation fleet control over SSH with live telemetry. Python, Swift
- **F75**. A ~100k-parameter transformer trained end-to-end from scratch on one RTX 5060 Ti, probing next-token behaviour under constrained capacity. PyTorch
- **LlamaVision-1B**. A 6.3M-parameter projector aligning vision features into a frozen Llama-3.3-1B embedding space. PyTorch, multi-modal
- **BlackTable**. RAG service for resume parsing, job-fit scoring and question generation over a Docling + LLM pipeline. FastAPI
- **NetGuard**. No-root Android firewall intercepting DNS via a local VpnService tunnel; per-app and wildcard blocking. Kotlin
- **H++**. A statically-typed language built on one primitive (the bit); compiler + runtime in C with Assembly. C, Assembly

AWARDS

Winner: HIVE Hackathon, IIT Delhi (2024) • **Winner**: FinAThlon, IIT Delhi/SIB (2023) • **Top 10 Internationally**: HERE Hackathon, IIT Bombay TechFest (2023) • **3rd**: DEXTERIX 4.0, Galgotias (2024) • **1st**: University Coding Challenge, IILM (2023)

RESEARCH & EDUCATION

Research. Infinite-context memory architectures and shallow-attention encoders with learned pooling for dense retrieval: himansh.in/technicalsummary

B.Tech, Computer Science & Engineering (AI & ML), IILM University | 2023 to 2027 | Google Cloud Generative AI, 6 certifications