HIMANSH RAJ

Applied AI Engineer | LLM Infrastructure & Agent Systems

iamthehimansh@gmail.com | +91 6200483104 | Noida, India

himansh.in | github.com/iamthehimansh | linkedin.com/in/iamthehimansh

Education

IILM University, Greater Noida 2023 – Expected 2027

B.Tech, Computer Science & Engineering (Artificial Intelligence & Machine Learning)

Experience

Colate | Applied AI Engineer Jul 2025 – Present

Cana: agent platform spanning desktop, web and mobile | cana.build

- Designed a shared-IR LLM gateway across OpenAI, Anthropic and Gemini; added 108 unit/integration tests for streaming, tool calls and protocol compatibility, plus prompt caching and token accounting.
- Built a Python agent runtime with dependency-aware orchestration, batched tool execution and approvals, integrated with React and Rust/Tauri; added unit and end-to-end regression tests.
- Implemented gVisor-isolated cloud execution and a Redis-backed Kubernetes autoscaler; added persistent workspaces and sequence/ACK event replay to recover missed progress after disconnects.

Indian Institute of Technology Delhi | Research Intern May 2026 – Aug 2026

Machine Intelligence, Signals & Networks Lab | Advisor: Dr. Sandeep Kumar

- Built agentic recommendation pipelines combining retrieval, planning and tool use; developed evaluation methodology for failure modes and robustness.

Open Source Contributions

- **Microsoft Semantic Kernel**: fixed empty tool results becoming `None` during XML prompt round trips; added regression tests. Merged PR #14686.
- **Google ADK**: made Agent Engine tests portable to macOS using GNU coreutils, resolving 13 failing test cases. Integrated upstream via Copybara.
- **AnyPS5**: six merged upstream PRs improving macOS builds, Mach-O signing layout, pooled-buffer handling, sleep timing and graphics tests. Merged contributions.

Selected Projects

Cana Desktop, Web & Mobile (Colate) Product walkthroughs

Shipped desktop/cloud workspace handoff; rebuilt the React Native chat client for iOS/Android with streamed tool activity, remote approvals, attachments and notification deep links.

Maya OS / Mayac | TypeScript, Python, SQLite, MCP Source / Agent

Built local llama.cpp inference, cloud model routing and PTY-backed agent teams; implemented permission-gated tools, persistent sessions and independent checks of generated workspace files.

LlamaVision | PyTorch, Transformers Source

Trained a 6.3M-parameter, two-layer projector connecting frozen CLIP ViT-L/14 and Llama 3.2-1B for image-conditioned text generation; published model and inference code.

Technical Skills

Languages: Python, TypeScript/JavaScript, C/C++, SQL, Rust, Swift

AI & systems: PyTorch, Hugging Face Transformers, MCP, RAG, llama.cpp, React, React Native, FastAPI, Node.js, Docker, Kubernetes, AWS, Redis, PostgreSQL, SQLite, WebSockets, Git, pytest

Awards

Winner: HIVE, IIT Delhi (2024); FinAThon, IIT Delhi (2023).

Top 10 internationally: HERE Hackathon, IIT Bombay TechFest (2023).