Sundaram Kumar Jha

Delhi-NCR, India | +91 6204446866 | jhasundarm@gmail.com | GitHub | Portfolio

Software engineer who turns complex, open-ended problems into dependable, production-grade software across the full stack. Specializes in AI infrastructure and developer-facing tooling, with end-to-end ownership spanning backend services, APIs, and the interfaces and pipelines that ship them. Pairs a research-minded approach with a fast, pragmatic delivery style and a consistent focus on reliability, security, and measurable engineering impact.

EDUCATION

Manav Rachna International Institute of Research and Studies — **Delhi, India**
Bachelor of Technology in Computer Science | *Graduation: June 2025* — CGPA: 7.2/10.0

TECHNICAL SKILLS

- **Programming Languages:** Python, TypeScript, JavaScript, Golang, HTML/CSS
- **AI Technologies:** Memory Systems, Agentic Workflows, AI Gateways, MCPs, RAG Systems, Skills
- **Tools and Frameworks:** Next.js, React.js, Docker, Kubernetes, GitHub Actions, PostgreSQL, Supabase

PROFESSIONAL EXPERIENCE

Aden Hive (Multi-Agent Harness for Production AI) — **Aug 2025 – Present**

- Shipped OpenRouter LLM-provider support across macOS, Linux, and Windows, adding model-level readiness validation at setup and fixing response parsing so agent tool calls execute correctly instead of rendering as raw text.
- Hardened agent command execution against prompt- and shell-injection by building a shared command-sanitizer module enforced across both execution paths, backed by comprehensive safe/blocked command test coverage.

OpenSre (AI SRE Agents) — **Aug 2025 – Present**

- Re-architected integration handling around a shared catalog centralizing normalization, environment loading, and service-precedence resolution, eliminating duplicated logic across the resolve and verification nodes.
- Strengthened automated root-cause analysis for RDS storage exhaustion by mapping Grafana metrics into structured CloudWatch-style evidence and surfacing FreeStorageSpace and WriteIOPS summaries in the diagnosis prompt, with regression tests.

Traceroot AI (Observability for Agents) — **Aug 2025 – Present**

- Added CrewAI to the getting-started onboarding flow and made the manual setup path reflect the selected integration, enabling one-click trace instrumentation for multi-agent CrewAI workflows.
- Made dev and prod workflows reliable across Windows, macOS, and Linux via a cross-platform Python task runner and a Docker-backed fallback for ClickHouse (goose) database migrations.

Synapsis Medical Technologies — **Feb 2025 – Jul 2025**

- Developed and integrated real-time communication features for a logistics platform, enabling live data synchronization between drivers and administrators to optimize oil-industry workflows.
- Engineered a responsive, full-stack web application using TypeScript, Next.js, and Tailwind CSS, featuring distinct portals for drivers and administrators to streamline operational management.

CloudDrove — **Oct 2024 – Feb 2025**

- Architected “Smurf,” a unified DevOps CLI consolidating Docker, Helm, and Terraform commands to cut multi-platform deployment time by 40%, with secure multi-cloud registry integration (ECR, GCR, ACR) accelerating image management by 60% with zero security incidents.
- Engineered an AWS Cost Optimization Calculator integrated with RDS and S3, delivering average annual client savings of $8,000 and boosting lead conversion by 35%; optimized real-time backends and CI/CD pipelines, cutting shipping time by 20%.

PROJECTS

Voice AI Memory System | GitHub

- Engineered a hands-free voice console with TypeScript and Next.js that pairs browser-native speech recognition (custom silence-detection VAD) with sentence-by-sentence streaming text-to-speech and barge-in interruption for natural, continuous conversation with Anthropic Claude.
- Built a real-time memory-graph engine on the Supermemory API that persists dialogue for cross-session recall and optimistically renders interactive document-to-memory nodes on a draggable, zoomable 2D canvas as responses stream.

Context Compaction System | GitHub

- Engineered a context-window manager in TypeScript that tracks live token usage against a budget and auto-compacts conversation history once a configurable 70% threshold is crossed, keeping long-running LLM chats within limits without losing continuity.
- Implemented rolling, LLM-driven summarization that condenses full history into a compact summary, persists it to a long-term history log, and re-seeds the session — preserving cross-turn memory while sharply reducing token consumption.

System Design Architect | GitHub

- Engineered an AI-driven system-design assistant with TypeScript and Next.js that orchestrates four specialized agents — researcher, orchestrator, validator, and canvas generator — to clarify requirements, run Exa-powered web research, and validate architecture plans before generation.
- Built an interactive React Flow canvas that converts LLM-produced design plans into editable node-and-edge architecture diagrams with PNG export and iterative regeneration, improving explainability and system-level reasoning.

ACHIEVEMENTS & CERTIFICATIONS

- **Published Researcher:** Co-authored the research paper “Fuzzy Logic Method to Solve Traffic Congestion,” published in the IEEE AutoCom 2024 conference proceedings.
- **Open Source Contributor:** 50+ merged pull requests across multiple open-source organizations — including Aden Hive, OpenSRE, and Traceroot AI — spanning multi-agent orchestration, AI SRE root-cause analysis, and agent observability tooling.