

SANDEEP SINGH

AI Engineer | Agent Infrastructure & LLM Systems | Full-Stack Engineer

sandy1711003@gmail.com • linkedin.com/in/sandeep-singh-445058254 • github.com/Sandy-1711 • +91 8953596712

AI Engineer building agentic LLM systems, memory, and RAG — **12 merged PRs** into Mastra's agent runtime (**25k+ stars**), shipped production agents at AiRA, and scaled a campus platform to **10,000+ users**.

Experience

AiRA

Nov 2025 - March 2026

AI Engineer (Agent Orchestration, Memory, RAG & Systems)

Bangalore, India

- Built the flagship **Daily Brief** agent (**Python/FastAPI**) — an **agentic orchestration loop** coordinating Email, Calendar, and **Exa AI** web-search tools via **asyncio** to deliver each user a personalized morning briefing of priority emails, events, and news.
- Designed a self-maintaining **Canvas** agent generating persistent, per-user web apps and documents, with automatic **versioning and data-shape migration** that keeps prior user data compatible across regenerations.
- Architected a multi-tier **memory system** (working / medium / long-term) on **MongoDB** and **Redis** that summarizes each user's conversations and style into a preferences store for personalization and proactive check-ins.
- Built an **LLM evaluation** pipeline with **promptfoo**, benchmarking 8+ models (Anthropic **Claude**, OpenAI, Grok) on tool-calling accuracy, memory use, and latency to standardize model selection.
- Cut routing-**LLM** token usage by **82%** via artifact-based generalized prompting and context deduplication across 3 core prompts, and parallelized the Orchestrator's tool-calling to reduce execution latency.
- Implemented **Langfuse** observability for LLM traces, latency, and cost, cutting time-to-debug for agent failures in production.

IIT Kharagpur

May 2025 - Aug 2025

Research Intern (Multimodal AI)

Kharagpur, India

- Developed **CircuitSynth**, a multimodal reasoning framework, fine-tuning **Qwen models (4B/8B)** with **LoRA/QLoRA** to generate executable Arduino circuit JSONs.
- Achieved **75% accuracy** on complex circuit-generation tasks (vs. near-zero for general-purpose GPT models) via a multi-stage **RAG** pipeline with automated unit-style validation.
- Optimized inference for edge deployment, running quantized **GGUF** models on **Raspberry Pi** devices.

Projects

Samagra — Founding Software Engineer, Indigle Play Store

Sep 2024 - Sep 2025

- Built **Samagra**, a full-stack campus ERP platform grown to **10,000+ active users**, spanning a **React Native** mobile client, a **Node.js/Express REST API**, **MongoDB**, and a **Next.js** web app.
- Engineered **JWT authentication** with **role-based access control**, rate limiting, and **Socket.io** real-time sync across the platform, sustaining **99.9% uptime**.
- Owned the product end-to-end — architecture, **CI/CD**, release, and Play Store distribution — on a **zero-cost** Vercel-hosted backend handling **150k+ monthly API requests** at sub-100ms latency.

Open Source | *TypeScript, Node.js, React, Vue.js*

2025 - Present

- **mastra-ai/mastra (25k+ stars)** [PR list]: **12 merged PRs** across the core agent runtime, streaming, ai-sdk, and Studio — including fixes to **semantic recall** injection, duplicate **stream-event delivery** on observe/replay, and tool-level **background execution** config. Active contributor since June 2026.
- **cal.com (46k+ stars)**: Authored a full **BigBlueButton video integration** (checksum-signed client, encrypted credential storage) plus a generalized install-router fix for non-OAuth apps — **912 LOC, 32 unit tests** (bounty issue).
- **n8n (195k+ stars)**: Authored core infrastructure fixes — a Webhook CSP security bug (**allow-same-origin**) and a cross-platform build-path failure.

Technical Skills

AI/ML & LLM: Claude Code, Hugging Face (PEFT/LoRA), LangChain, Langfuse, LlamaIndex, Mastra, OpenAI/Anthropic/Gemini API, Pinecone, promptfoo, PyTorch, Qdrant

Backend & Ops: asyncio, AWS (Lambda, S3), CI/CD (GitHub Actions), Docker, FastAPI, MongoDB, PostgreSQL, Redis, REST APIs, WebSockets

Frontend & Mobile: Next.js, React Native (Expo), React.js, Shadcn/UI, Tailwind CSS

Languages: JavaScript, Python, SQL, TypeScript

Education

Madan Mohan Malviya University of Technology

Nov 2022 – Aug 2026

B.Tech in Electronics & Communication Engineering (CGPA: 7.32)

Goarakhpur, India