Himanshu Sharma

Software Engineer — AI Agent Platforms & Full-Stack Development

Gurugram, India · +91 6267386284 · himanshusharmamig196@gmail.com · linkedin.com/in/himanshu-sharma-ab7101313 · github.com/HIMpcgithub3000

SUMMARY

Final-year B.Tech CS student focused on Applied AI and LLM Engineering — building AI products that solve real business problems, not just demos. Shipped Nexus, a solo full-stack platform for directing and monitoring Claude CLI agents in production, plus RAG applications, AI agent/workflow automation platforms, document-intelligence systems, and AI-powered developer tools across Next.js/TypeScript/Tailwind and Python/FastAPI. Skilled at directing AI coding agents, owning ambiguous specs end-to-end, and building production-ready systems and UI for asynchronous, probabilistic systems. Open to AI Engineering, Applied AI, Generative AI, LLM Engineering, AI Product Engineering, and Software Engineering roles.

EXPERIENCE

Blostem — SDE Intern — Jun 2026 – Present

- Architected and shipped **Nexus**, a multi-tenant AI agent orchestration platform, in close collaboration with senior engineers and product.
- Integrated SigNoz error logs, Bitbucket source, and Google Chat alerts for automated root-cause analysis — compressing 3-hour manual error traces into 10 minutes.
- Built a Python/FastAPI LLM-usage dashboard across 3 OpenRouter models for 50+ engineers, cutting usage audits from 4 hrs/week to 15 min (94% reduction).

UptoSkills — Frontend Developer Intern — Jun 2025 – May 2026

- Shipped 12 production React/TypeScript components for an HRMS platform used by 100+ employees.
- Increased user engagement 40% through rapid UI iteration with stakeholders.

PitchX (ACIC – BMU Incubator) — Full Stack Intern — Jan 2025 – Apr 2025

- Delivered 8 product features across 3 modules in a 14-week cycle; owned 5 backend endpoints end-to-end.
- Reduced ticket turnaround from 5 to 2 days (60% improvement); authored release documentation.

PROJECTS

Nexus — Multi-Tenant AI Agent Orchestration & Monitoring Platform — Apr 2026 – Present

Node.js · Express · TypeScript · PostgreSQL · Prisma · Socket.IO · React · Tailwind · Claude CLI · MCP

- Production control plane for directing/monitoring AI agents (Claude CLI subprocesses) in real time; integrates SigNoz, Bitbucket, and Google Chat for automated root-cause analysis — 45+ service modules, ~18K LOC, 175+ automated tests.
- Board/Task/Approval workflows for human review of agent decisions before irreversible actions, with a full audit trail across 40+ Prisma models, 50+ REST endpoints, and 25+ WebSocket event types tracking session state, token usage, and live errors.

NextFlow — LLM & Media Workflow Builder — Jan 2026 – Present

Next.js 16 · React Flow · TypeScript · Trigger.dev · Gemini · FFmpeg · Transloadit · Prisma · PostgreSQL — github.com/HIMpcgithub3000/nextflow

- Visual DAG workflow builder: 6 node types, typed ports, cycle prevention, undo/redo, JSON import/export, debounced autosave, per-user persistence.
- Topological executor running text, media-upload, Gemini-vision, image-crop, and video-frame-extraction nodes via Trigger.dev workers; run history in PostgreSQL/Prisma via Clerk auth and Zod validation.

Vernacular FD Advisor — 15-Language Voice & RAG Assistant — May 2025 – Present

Next.js 15 · React 19 · TypeScript · Python · FastAPI · FAISS · BGE-M3 · BM25 · Ollama · PostgreSQL — github.com/HIMpcgithub3000/Vernacular-FD-Advisor

- Citation-grounded conversational assistant across 15 Indian languages: Web Speech STT/TTS, language-specific UI state, source drawers, retrieval telemetry, evidence highlighting.
- Python RAG backend: FAISS, BGE-M3 embeddings, BM25, Reciprocal Rank Fusion, MMR, cross-encoder reranking, plus intent routing, compliance redaction, deterministic financial math, and 35-case multilingual evaluation tracking.

MediMind — Healthcare & Insurance RAG Platform · NHCX Pipeline — Hackathon Finalist

- MediMind: evidence-grounded QA for healthcare/insurance — every answer traceable to source documents (github.com/HIMpcgithub3000/medimind).
- NHCX (Finalist): autonomous GenAI pipeline converting unstructured insurance PDFs into NHCX FHIR R4-compliant structured data, 100+ policy documents processed on fully local infrastructure (github.com/HIMpcgithub3000/NHCx).

SKILLS

Languages: Python, TypeScript, JavaScript, SQL, C++

AI Agents & Tooling: Claude CLI, Claude Code, Codex CLI, GitHub Copilot, Cursor, MCP, agent harness design, agent orchestration, prompt engineering, ML fundamentals (scikit-learn, PyTorch, TensorFlow, MLflow, predictive modeling)

RAG & LLMs: FAISS, ChromaDB, BGE-M3, BM25, RRF, MMR, cross-encoder reranking, LangChain, Ollama, Gemini, grounded generation, citations

Front-End: React, Next.js, Vite, TypeScript, HTML, CSS, Tailwind, Zustand, React Query, React Flow, responsive UI, dashboards

Back-End: Node.js, Express, FastAPI, Flask, REST APIs, WebSockets (Socket.IO), PostgreSQL, Prisma, Zod, queues/workers, Trigger.dev, Git

Integrations: Gmail, Google Drive (OAuth), Google Chat, Bitbucket API, SigNoz, SMTP

EDUCATION

BML Munjal University Aug 2023 – May 2027

B.Tech, Computer Science & Engineering — CGPA 6.67/10

Coursework: ML, NLP, data structures, software engineering, predictive modeling

CERTIFICATIONS

- Oracle Cloud Infrastructure 2025 Certified AI Foundations Associate
- Goldman Sachs — Software Engineering Job Simulation
- Work with Components in Figma