HIMANSH RAJ

iamthehimansh@gmail.com • +91-6200483104 • linkedin.com/in/iamthehimansh
github.com/iamthehimansh • himansh.in • F-75 Noida, Sector 27, Uttar Pradesh, 201301

SUMMARY

AI Practitioner and Software Engineer focused on research and development of LLMs/SLMs, multi-agent systems, and learning-based prototypes. Experienced in fine-tuning transformer models (PyTorch, HuggingFace), designing scalable backend services, and shipping production ML pipelines. Particularly interested in studying model behavior, robustness, and failure modes. Proven track record of taking AI products from research prototype to production across freelance, internship, and open-source engagements.

TECHNICAL SKILLS

Languages: Python, C++, C, JavaScript, TypeScript, Dart, SQL, Solidity
ML & AI: PyTorch, HuggingFace Transformers, Unsloth, TRL, LoRA/QLoRA, RLHF, RAG, Vector Embeddings, Fine-tuning (Llama, Mistral), Prompt Engineering, Model Quantization
Agents: LangChain, LangGraph, Multi-agent Orchestration, Tool Use, Function Calling, MCP
Web & Backend: Next.js, React, Three.js, FastAPI, Flask, Django, Node.js, Express, WebSockets, REST APIs
Databases: PostgreSQL, MongoDB, Redis, Firestore, Vector DBs (Pinecone, Chroma, FAISS)
DevOps & Tooling: Docker, Kubernetes, GitHub Actions, CI/CD, Linux, Nginx, Pytest, Jest, Git
Core CS: Data Structures & Algorithms, Operating Systems, Computer Networks, Distributed Systems, System Design

PROFESSIONAL EXPERIENCE

Indian Institute of Technology Delhi, MISN Lab *May 2026 – Present*
Research Intern, Machine Intelligence, Signals & Networks Lab

- Designing LLM-driven agents that produce personalized recommendations by reasoning over user context and preferences, as part of "Agentic Systems for Recommendations" under Dr. Sandeep Kumar.
- Moving beyond static collaborative-filtering and embedding-similarity baselines by building multi-agent pipelines that combine retrieval, planning, and tool use.
- Hardening agent decision-making by developing evaluation methodology for agentic recommenders, including failure-mode characterization and robustness analysis.

Colate *Jul 2025 – Dec 2025*
AI Engineer (Remote, Part-time)

- Delivered sub-second retrieval over millions of code embeddings for a code intelligence platform by architecting a scalable Vector Database and semantic search engine from indexing to query layer.
- Automated full-stack software generation from natural-language specifications by designing multi-agent workflows that coordinate planner, coder, reviewer, and tester agents over shared state.
- Cut inference latency by 35% by profiling latency–accuracy tradeoffs across embedding-storage strategies and rolling out Redis-backed caching on hot paths.
- Surfaced robustness gaps, hallucination patterns, and failure modes across 90+ benchmark prompts by running structured qualitative evaluation and feeding human-in-the-loop signals back into the agent pipeline.

SELECTED PROJECTS

F75 – Small Language Model from Scratch github.com/iamthehimansh/F75
Python, PyTorch, Transformer, Custom Tokenization, GPT-2 style architecture
Demonstrated LLM internals beyond high-level API usage by training a 107K-parameter transformer language model end-to-end on a single GPU, surfacing training dynamics, dropout behavior, and inference characteristics under tight compute constraints.

LlamaVision – Vision-Language Model huggingface.co/iamthehimansh/LlamaVision-llama-3.3-1b
Python, PyTorch, Vision Transformers, Multimodal Alignment
Studied cross-modal representation learning and alignment failures in multimodal LLMs by building a 6.3M-parameter image projector that maps visual features into the text embedding space of a Llama-3 backbone.

3DWebAI – Fine-tuned LLM for Generative 3D Scenes huggingface.co/iamthehimansh/3dAiWeb
Llama-3, Unsloth, TRL, LoRA, React Three Fiber
Generated interactive React Three Fiber 3D web scenes from natural-language prompts by fine-tuning Llama-3 with Unsloth and TRL on a custom 100-repository GitHub instruction dataset.

BlackTable – LLM-based Document Analysis Toolkit github.com/iamthehimansh/Blacktable
Python, FastAPI, OpenAI, RAG, Docling, Vector Embeddings
Shipped a production-ready document-analysis API spanning 4 formats (PDF, Markdown, Word, plain text via Docling) by composing resume scoring, RAG-based Q&A, and structured-extraction pipelines behind a single FastAPI surface.

Doremon – EEG-Controlled Smart Home Automation github.com/iamthehimansh/Doremon
Python, IoT, EEG signal processing, BCI, MQTT
Reached 67% intent-classification accuracy on only 2 hours of hackathon-collected EEG data by building an end-to-end brain-signal-to-smart-device pipeline (signal preprocessing, feature extraction, lightweight classifier).

8085 Microprocessor Simulator github.com/iamthehimansh/8085-simulation
C++, Assembly, Emulation, Systems Programming
Enabled end-to-end 8085 assembly debugging by building a full Intel 8085 emulator with 64KB memory simulation, instruction-level breakpoints, register inspection, and step-through execution.

ONGOING RESEARCH

Infinite Context LLM – Active/Passive Memory Architecture *Jan 2025 – Present*
Designing a memory architecture for LLMs/SLMs with separate active (in-context) and passive (retrievable) memory states, enabling scalable, near-infinite knowledge retention and efficient contextual recall without quadratic attention cost. Exploring how memory consolidation strategies affect downstream task performance and long-horizon reasoning. Research progress documented at himansh.in/blog.

EDUCATION

IILM University, Greater Noida *Apr 2023 – Present*
B.Tech in Computer Science & Engineering (AI & ML Specialization)
Relevant Coursework: Data Structures & Algorithms, Operating Systems, Database Management, Computer Vision, Neural Networks & Deep Learning, Distributed Systems, Computer Networks, Probability & Statistics, Linear Algebra, Discrete Mathematics

ACHIEVEMENTS & AWARDS

- **Winner** – HIVE Hackathon, IIT Delhi (2024) – Built a winning solution among 100+ teams in a 36-hour hackathon focused on innovative software engineering.
- **Winner** – FinAThon Hackathon, IIT Delhi (organized by SIB, 2023) – 1st place in financial technology challenge solving real-world banking problems.
- **3rd Place** – Dexterix Hackathon, Galgotias University (2024) – Recognized for innovation in AI/ML category.
- **Top 10** – IIT Bombay TechFest International (Here's Hackathon, 2023) – Selected among top 10 teams internationally.
- **1st Place** – University Coding Challenge, IILM (2023) – Competitive programming contest covering algorithms and data structures.

CERTIFICATIONS

- **Google Cloud Generative AI Apps** – 6 certificates covering Gemini API, Vertex AI, RAG, agent design, and responsible AI (Jun 2025).
- **Java Programming** – DataFlair (2024).
- **Python Programming & C Programming** – Infosys Springboard (2024).
- **HackerRank Certified:** Node.js, Python, React, REST API Intermediate.

LANGUAGES

Hindi – Native • **English** – Professional Working Proficiency