Rajat Kumar Thakur

Delhi, India
8810544717 rajatlovescloud@gmail.com linkedin.com/in/rajat-kumar-thakur github.com/rajat-kumar-thakur

Education

Indian Institute of Information Technology Vadodara **November 2022 – June 2026**
Bachelor of Technology in Computer Science and Engineering (CGPA: 9.46) *Gujarat, India*

Experience

Software Engineer Intern **September 2025 – Present**
Cableteque Corporation *California, USA*
- Engineered a scalable search agent for advanced querying using MCP servers to optimize and achieve 97.6% accuracy.
- Collaborated to developing, and deploying Agentic models in production, and collaborating with a 5-member team.

Summer Intern **May 2025 – July 2025**
IIT Gandhinagar *Gujarat, India*
- Fine-Tuned language models for code generation and execution on resource-constrained devices, achieving inference by StarCoder model in under 3 seconds.
- Optimized small language models for resource-constrained devices, reducing inference latency by 46.3% for TinyLlama.

Projects

Hindi AI Assistant [Link] | *FastAPI, Next.js, OpenCV, LangChain, Langgraph* **Nov 2025**
- Built emotion-aware AI assistant with real-time facial expression detection across 8 emotion states using OpenCV; integrated multi-modal pipeline combining voice, vision, and GPT-5 for contextual Hindi conversations.
- Implemented end-to-end Hindi speech processing with Google Speech API and gTTS, reducing response latency to ¡2s; architected FastAPI backend serving 15+ REST endpoints with conversation state management via LangGraph.

LLM Semantic Query Engine [Link] | *FastAPI, RAG, Gemini API, Pinecone, SQLite* **August 2025**
- Architected and deployed a full-stack PDF Q&A application using FastAPI. Improved performance by implementing efficient retrieval strategies that reduced response time to 2.1 seconds and supported the project's full life cycle.
- Integrated parsing, semantic chunking (1000+ chunks per document), and APIs to deliver structured answers with citations, reasoning, and confidence metrics.

Pose Estimation for Time-Critical Applications [Link] | *Computer Vision* **March 2025**
- Devised a pipeline that is 4x faster for pose estimation in point clouds using surface variation, Harris corner detection, and pose estimation using corner points.
- Achieved a 70% reduction in computation time (from 26.8 s to 7.9 s) compared to state-of-the-art methods.
- Paper accepted into the 11th International Conference on Pattern Recognition and Machine Intelligence (PREMI) '25.

Relevant Coursework

- Data Structures
- Software Engineering
- Image Processing
- Database Management
- System Software
- Data Analytics
- Algorithms
- Machine Learning
- Cryptography
- Computer Networks
- Artificial Intelligence
- Operating Systems
- Computer Architecture
- OOPS
- 5G Communication
- Parrallel Computing

Technical Skills

Languages: C, C++, Python, R, JavaScript
Databases: SQL, MongoDB, PostgreSQL
Developer Tools: VS Code, Google Cloud Platform, ArchiMate, StarUML, Jupyter Notebook, TensorBoard, Docker
Technologies/Frameworks: Next.js, React, Node, Linux, Git, Postman, Express, Langgraph, Tailwind, LangChain
Cloud Platforms: Google Cloud Platform, AWS, Vertex AI
Libraries & Tools: NumPy, Pandas, PyTorch, GitHub Actions, Hugging Face Transformers, FAISS, Sentence-Transformers
Generative AI / LLM Skills: Prompt Engineering, RAG Pipelines, Fine-tuning & Instruction Tuning, LLM Evaluation, Agentic Workflows, Vector Databases, Retrieval-Augmented Generation, Open-Source LLMs (Llama 3, Mistral, Gemma)

Extracurricular

- Mentored over 190 students as a **Teaching Assitant** in developing skills in C and designing scalable software solutions
- Secured **2nd place** in the Battle Bots and Sentience events at a Fest for designing a robotic hand and a combat robot.
- Led an 8-member team as **General Secretary of the Academics Committee** to plan and execute academic events.
- Attained mastery of complex algorithms through completion of 500+ challenging **LeetCode** problems.LeetCode
- Earned an NVIDIA Certificate of Competency in evaluation and light customization of LLMs.Certificate
- Completed a comprehensive 12-hour online training in deep learning offered by NVIDIA.Certificate