

Utkarsh Kumar Rai

Cuttack, Odisha | 📞 +91 7848972103 | ✉ utkarshkumarrai2005@gmail.com | 🌐 github | </> leetcode

EDUCATION

ABV Indian Institute of Information Technology and Management, Gwalior

Expected July 2027

B.Tech in Electrical and Electronics Engineering

CGPA: 7.46

Relevant Coursework: Data Structures & Algorithms, Object Oriented Programming, Operating Systems, Database Management Systems, Computer Networks, Computer Organization, Linear Algebra, Probability & Statistics, Machine Learning, Computer Vision

TECHNICAL SKILLS

Languages: Python, C++, C

Backend & Systems: REST APIs, gRPC, TCP/IP Sockets, Distributed Systems, SQL (PostgreSQL / SQLite), Multi-threading, Concurrency

ML / DL: PyTorch, TensorFlow, Scikit-learn, Vision Transformers (ViT), ArcFace, FAISS, LoRA, LLM Fine-tuning, RAG

Computer Vision: OpenCV, Image Processing, Object Detection, Similarity Search

Tools: NumPy, Pandas, Streamlit, HuggingFace, PEFT, Gradio, Git, GitHub, Docker, Jupyter, MLflow

EXPERIENCE

Colloquium Guided Project — ABV-IIITM Gwalior 🌐

May 2026 – July 2026

Under the guidance of Dr. Neelesh Yadav, Department of EEE

Visual Search Engine for E-Commerce Products

Python, PyTorch, FAISS, Streamlit, HuggingFace

- Improved retrieval accuracy by **17 percentage points** over the ResNet50 baseline by fine-tuning **ViT-B/16** with **ArcFace metric learning** on 120,053 product images, achieving **87.17% Recall@10**.
- Reduced search latency by **11×** over brute-force by engineering a **FAISS HNSW ANN index** over 120,053 embeddings, achieving **~0.2 ms** per query with **<3 pp** accuracy loss.
- Optimized end-to-end query response to **~90–150 ms** through a **Streamlit web application** with live image upload, top-K result grid, and per-result similarity scores.
- Deployed the inference pipeline to **HuggingFace Spaces** with CPU-only inference and automated error handling, configuring a git-lfs workflow for 680 MB of model artifacts.
- Presented the system architecture and benchmark results to faculty and peers at the departmental colloquium.

PROJECTS

E-Commerce Product Assistant — LLM Fine-tuning + RAG

🌐 *Python, PyTorch, HuggingFace, PEFT, FAISS, Gradio*

- Fine-tuned **TinyLlama-1.1B-Chat** with **LoRA** ($r=16$, $\sim 1.2\%$ trainable params) on 5,000 e-commerce Q&A pairs on **Google Colab T4 GPU** in under 15 minutes; adapter size ~ 20 MB.
- Built a **RAG pipeline** with a **FAISS index** over 10,000 product descriptions using sentence-transformers, achieving **<50 ms retrieval latency** and **>80%** hit rate on held-out eval.
- Launched an interactive **Gradio chat UI** that grounds LLM responses in retrieved catalog context, removing hallucinated answers; end-to-end latency **~1–3 s** on GPU.

ParaQuery — Distributed SQL Query Engine 🌐

Python, gRPC, SQL, Docker

- Architected a distributed SQL processing engine across **3 parallel worker nodes** via **gRPC streaming**, achieving **1.70×** query speedup over single-node execution and processing **2M+ rows** on local Docker.
- Minimized inter-node data transfer by implementing **predicate pushdown** and **partial aggregation** (MapReduce-style), lowering coordinator-to-worker network traffic.
- Engineered a **fault-recovery** mechanism and an **EXPLAIN endpoint** exposing the full query execution plan, and authored a detailed repository README documenting the architecture, API, and setup.

ACHIEVEMENTS

- Competitive Programming:** Solved **200+ problems** across LeetCode and GeeksforGeeks, spanning arrays, trees, graphs, DP, and greedy techniques.
- JEE Advanced 2023:** Secured All India Rank (AIR) **12,000** (top $\sim 8\%$) among 150,000+ candidates.
- Flipkart Grid 6.0:** Led a 3-member team to clear Level 1 and advance to Level 2 in Flipkart's national-level software development challenge.
- Google Solution Challenge 2025:** Developed and demoed a full-stack application using Google Cloud, Maps API, and Firebase in Google's annual global developer challenge.