VINAYAK SAHU

Diving Into AI/ML

vinayak1672006@gmail.com • GitHub • LinkedIn • **PORTFOLIO**

Contact: +918858838007

Technical Skills

- Languages & Frameworks: Python, SQL, Shell Scripting, FastAPI, Git/GitHub, Gradio
- ML & AI Libraries: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, SentenceTransformers, CLIP, FAISS, YOLOv8, LangChain, LangGraph, LlamaIndex
- Big Data & Query Engines: Hadoop, HiveQL, HDFS, Cloudera VM
- Data Analytics & Visualization: Pandas, NumPy, Matplotlib, Tableau, Power BI, Advanced Excel

Projects

PrepGraph - RAG-Based Course Chatbot | Personal Project [GitHub]

- Deployed a **Retrieval-Augmented Generation (RAG) chatbot** combining **6 subjects' syllabi + 200+ PYQs** into one knowledge base for exam prep.
- **Implemented** document loaders (**PyPDF, PPT**), **chunking (300+50 overlap)**, and a **1k-character retrieval cap**, **reduced context size and LLM token usage by ~40%**.
- **Impact:** adopted by student groups (30+ active users) during semester exams; answered 200+ queries with lower latency & cost.

Stack: LangGraph, Groq API (LLaMA3-8B), FAISS, SentenceTransformer, Gradio.

Linux Command Copilot - Offline AI Assistant for Shell Automation (Under Active Development) **[GitHub]**

- Automated system administration tasks using a terminal AI assistant that converts natural language queries into Linux shell commands; this eliminated 75% of manual command lookups.
- Fine-tuned **Phi-2** with **Low-Rank Adaptation (LoRA)** on a custom dataset of size 150, improving command accuracy by **35% in benchmark evaluations**.
- Deployed inside a VirtualBox Linux VM for **fully offline execution with no cloud dependency**.

Stack: Hugging Face Transformers, Python, Shell Scripting, Low-resource LLM Inference, Prompt Engineering.

Big Data Analytics on MovieLens – *Hive + Hadoop Query Engine Project [GitHub]*

- Designed and implemented a **scalable ETL pipeline** for large-scale movie datasets using Hadoop & Hive.
- Created external Hive tables and optimized queries with **partitioning & bucketing**, reducing query execution time by **~40%**.
- Extracted insights on genres, ratings, and user behavior; reproducible in **Cloudera VM**.

Stack: Hadoop, Hive, HiveQL, HDFS, Cloudera VM, Linux CLI, Big Data Analytics

Reel2Retail – AI Fashion Video-to-Product Matching [GitHub]

- Built an AI system to **detect fashion items in social media reels** and match them with products in a catalog.
- Reduced false positives with frame-differencing & async caching, cutting latency by **~30%**.
- Achieved **~85% retrieval accuracy** on 100+ frames from the reel samples, while reducing latency via async caching and frame differencing.

Stack: YOLOv8, OpenAI CLIP, FAISS, OpenCV, Scikit-learn, NLP, Asynchronous Programming, GPU Inference

Achievements

First Prize – KIIT FED Hackathon: Built and integrated an intelligent chatbot with *MoneyMate*, a finance tracker, enhancing user interaction and financial management.

Education

Kalinga Institute of Industrial Technology (KIIT)
BTech in Computer Science | 2023 – 2027

Certifications

IBM – Courses on Supervised ML (Classification & Regression), Unsupervised ML, and Exploratory Data Analysis (Completed Jan – Mar 2025)