Himanshu Kumar

+91-7763855036 | himanshukumariiitn@gmail.com | LinkedIn | Google Scholar | GitHub | Portfolio

Education

Indian Institute of Information Technology, Nagpur 2022 – 2026
B.Tech in Computer Science and Engineering — CGPA: 8.41/10

Research & Publications

When Gujarati Meets English: Toward Robust Translation of Code-Mixed Low-Resource Indian Language LM4UC, AAAI 2026
- Constructed a 30K Gujlish–English parallel corpus using BPCC and LLM-based code-mixed generation (GPT-4o) with human validation, enabling translation for Romanized Gujarati code-mixed text into English.
- Fine-tuned NLLB-200 on the dataset and introduced Gujlish evaluation benchmarks (adapted from XNLI and IN22), achieving 1.5–2× improvements in BLEU and ChrF++ over Google Translate with strong human preference.

NanoVLM: How Small Can Vision Language Models Be and Still Generate Coherent Text? arXiv preprint
- Built NanoVLM (mini/base/large), parameter-efficient VLMs generating coherent captions at up to 10× smaller scale than standard VLMs.
- Curated ShortDesc/LongDesc caption datasets and proposed evaluation axes (creativity, consistency, semantic coherence) to quantify model size vs. quality trade-offs.

Work Experience

AI Research Intern – Vizuara AI Labs Jan 2026 – March 2026
- Investigated implicit demographic encoding in vision models using a multi-scale interpretability pipeline combining linear probing, filter-level correlation analysis, and representation geometry across ResNet and ViT architectures.
- Identified distributed subspace encoding of sensitive attributes and performed targeted filter ablations, reducing gender probing accuracy from 85% to 43% while preserving task performance and revealing causal bottlenecks.

AI Intern – WSAI, IIT Madras May 2025 – July 2025
- Developed an ambient soundscape music generation pipeline using MusicGen, training on a curated 10-hour dataset and building a web platform to collect structured human preference feedback.
- Applied Reinforcement Learning from Human Feedback (RLHF) to align generated audio with human aesthetic and perceptual preferences using reward modeling and preference-based optimization.

ML Engineering Intern – ProCohat Technologies June 2024 – Aug 2024
- Architected a production MultiPDF RAG system with semantic chunking, embedding-based retrieval, and vector similarity search across 100+ documents, achieving 85% retrieval accuracy with Supabase pgvector storage.
- Deployed end-to-end ML inference pipeline serving 50+ daily queries with sub-2s latency using FastAPI, achieving measurable improvement in document retrieval quality over keyword-based baselines.

Projects

Multimodal RAG over HuggingFace Courses GitHub
- Built a full-stack multimodal RAG system over 8 HuggingFace courses (2,200+ chunks) using semantic chunking, BGE-small text embeddings and CLIP image embeddings indexed in a Qdrant vector database with named vectors.
- Implemented streaming inference with FastAPI (SSE),and LLM generation(Gemini 2.5 Flash with Llama 3.3 70B fallback); deployed backend on HuggingFace Spaces and Next.js frontend on Vercel.

Healthcare Appointment Scheduling Multi-Agent System GitHub
- Built a LangGraph-based multi-agent workflow for healthcare scheduling where agents coordinate intent parsing, doctor availability retrieval, and appointment booking through a shared state graph and tool-based orchestration.
- Implemented real-time slot validation and conflict resolution with database-backed scheduling logic, integrating asynchronous email/SMS notifications to handle booking confirmations, updates, and cancellations.

Sankshipt: Multilingual News Summarization GitHub
- Developed a transformer-based multilingual news summarization system supporting 10 Indian languages, aggregating articles from heterogeneous sources and generating summaries using cross-lingual normalization and preprocessing.
- Implemented language detection, content cleaning, and entity-aware summarization to preserve key events, names, and contextual information while producing short summaries suitable for multilingual news consumption.

Multi-Label Sentiment Analysis GitHub
- Built a 9-label multi-label sentiment classification model using DistilBERT to capture overlapping emotional categories in user-generated text, enabling simultaneous prediction of multiple sentiment labels.
- Addressed severe label imbalance using weighted binary cross-entropy and threshold tuning, achieving 88% accuracy while improving detection of minority sentiment classes.

Technical Skills & Coursework

Languages: Python, C/C++, SQL
ML/AI: NLP, Computer Vision, Multimodal Learning, Generative AI, LLMs, Fine-Tuning (LoRA/PEFT), RLHF, RAG, Vector Search, Prompt Engineering, TTS/STT
Frameworks: PyTorch, TensorFlow, HuggingFace Transformers, Sentence-Transformers, LangChain, LangGraph, scikit-learn, Pandas, NumPy
Tools & Infrastructure: Docker, Git, FastAPI, Flask, Qdrant, Supabase, AWS, Vercel, HuggingFace Spaces, Weights & Biases
Core Coursework: Linear Algebra, Probability & Statistics, Discrete Mathematics and Graph Theory, DSA, DBMS, Operating Systems, CUDA Programming, Computer Networks, Machine Learning, NLP, Reinforcement Learning

Achievements

2nd Runner-up — AI Hackathon, Jagriti, IIT BHU (700+ teams) — Built Multilabel Sentiment Analysis System
2nd Runner-up — Enigma CTF, Codefest, IIT BHU (1600+ teams)