

Mihir Khatri

Vasai, India · khatrimihir672@gmail.com · +91 9226570903
[GitHub](#) · [Portfolio](#) · [LinkedIn](#) · [Kaggle](#) · [ORCID: 0009-0004-9166-6891](#)

PROFESSIONAL SUMMARY

Undergraduate Computer Science student focused on machine learning systems, GPU-aware inference optimisation, and LLM efficiency research. Published independent researcher with hands-on experience in CUDA kernel development, transformer architectures, agentic multi-agent systems, and production-grade ML pipelines. Interests span supervised learning, deep neural networks, sequential models, dimensionality reduction, and generative AI with a bias toward measurable, deployment-oriented results.

EDUCATION

Sardar Patel Institute of Technology, Mumbai University

Mumbai, India

B.Tech in Computer Science and Engineering

Expected Aug 2028

- GPA: 9.35 / 10.0
- Relevant Coursework: Linear Algebra, Probability & Statistics, Data Structures & Algorithms, Discrete Mathematics, Operating Systems, Computer Networks

PUBLICATION

Ro-SVD: Position-Aware SVD Compression of KV Caches in GQA Models | *Zenodo, May 2026* — [code](#)

- Proposed online activation-level KV cache compression using position-aware truncated SVD on pre-RoPE key tensors in Grouped-Query Attention architectures.
- Evaluated across 1,602 samples on all 16 LongBench subsets using Qwen2.5-7B-Instruct with 4-bit NF4 quantisation on dual NVIDIA T4 GPUs.
- Achieved 73.91% win rate on ROUGE-L against uncompressed baseline while reducing KV-cache memory by 35.1%.
- Designed adaptive staircase rank allocation with per-head energy thresholding to preserve quality in high-entropy attention heads.
- Implemented custom CUDA de-rotation kernel with cuBLAS SGEMM batched low-rank projection; derived randomised SVD optimisation path reducing complexity from $O(\text{Thd}^2)$ to $O(\text{Thr} \log r)$.

INDUSTRY EXPERIENCE

Daten & Wissen Pvt. Ltd.

Thane, India

AI/ML Intern

2025

- Fine-tuned PaddleOCR models for license plate detection and recognition on a dataset of 2,000 training and 1,000 test images (cumulative across localiser and recogniser stages), achieving ~88% end-to-end recognition accuracy after 1.5 days of full training.
- Improved robustness under noisy real-world conditions via geometry correction, adaptive filtering, and contrast normalisation.
- Evaluated segmentation architectures including U-Net, SAM, and ResNet backbones for production-oriented latency-accuracy trade-offs.
- Conducted Optuna hyperparameter optimisation sweeps; analysed convergence stability across training runs.
- Built end-to-end document understanding pipelines using PyTorch and TensorFlow.

RESEARCH & PROJECTS

ZENSO — Neural Engagement Analytics Platform | *LangGraph, Multi-Agent Systems* · 2026 (Ongoing)

- Leading ML development on a neural engagement analytics platform that models viewer engagement from video content.
- Designed a multi-agent product architecture in LangGraph and validated model efficiency trade-offs through structured ablation experimentation.

SkillFlow AI — Agentic Peer Skill-Bartering Platform | *Built for JPMC Agastya Foundation* · 2025–2026 — [code](#) · [live](#)

- Designed and deployed an agentic AI platform enabling students to barter skills peer-to-peer, matching learners and mentors based on complementary skill profiles.
- Built as part of volunteer work mentoring school students through the JPMC Agastya Foundation NGO initiative.
- Implemented autonomous agent logic to parse user-submitted skill profiles, generate compatible matches, and guide users through the bartering workflow without manual moderation.

Automated Form-Filling Agent | *LangGraph, Selenium, LLM Web Search — FAR AWAY 2026 Hackathon · 2026* — [code](#)

- Built an agentic system that ingests a user's profile once and autonomously fills out recurring application forms going forward, removing repetitive manual re-entry.
- LLM agent searches and parses the target form/website structure to infer required fields and constraints specific to that application.
- Dynamically generates Selenium automation scripts to autofill the identified form end-to-end based on the user's stored profile data.

Transformer Encoder-Decoder from Scratch | *PyTorch · 2025*

- Implemented multi-head self-attention, causal masking, positional encoding, and autoregressive decoding from first principles in PyTorch.
- Studied gradient flow, training stability, and architectural behaviour compared with RNN/LSTM baselines; developed intuition for attention complexity and sequence modelling trade-offs.

LLM Inference Optimisation: KV Cache via Low-Rank Approximation | *PyTorch, CUDA, Triton · 2026*

- Built end-to-end pipeline for inverse RoPE de-rotation, truncated SVD, and progressive online re-compression of KV caches.
- Ran 7-configuration evaluation sweep over 1,602 LongBench samples tracking ROUGE-L, perplexity, latency, and GPU memory usage.
- Quantified 29.5% latency overhead of Python-loop SVD; explored Triton-based batched randomised SVD to reduce this cost.
- Demonstrated compatibility between activation-level compression and 4-bit NF4 quantised inference.

OCR + LLM Reconstruction Pipeline | *Multimodal NLU · 2026*

- Built multimodal document reconstruction pipeline combining image preprocessing, OCR, and LLM-based reasoning.
- Explored latency-quality trade-offs between local quantised inference and API-based models; analysed hallucination behaviour across prompting strategies.

Raga Classification — Audio Deep Learning | *PyTorch · 2025*

- Built a 68-class audio classification system for Indian classical ragas using mel-spectrogram and MFCC representations; achieved 68% classification accuracy.
- Studied challenges of non-Western musical feature extraction and class imbalance in fine-grained audio taxonomies.

YOLO-Based Rockfall Detection — “Mineguard” | *Computer Vision · 2025* — [code](#)

- Trained YOLO object detection model for safety-critical rockfall detection on limited annotated datasets.
- Analysed precision-recall trade-offs under varying IoU thresholds; explored data augmentation strategies to improve generalisation.

CERTIFICATIONS & CONTINUED LEARNING

- Hugging Face — Agents Course, Unit 1 Fundamentals Certificate (LangGraph, smolagents, LlamaIndex) · 2026
- Hugging Face — NLP Course Certificate · 2026
- Google Cloud Skills Boost — Agent Fundamentals · 2026
- Google Cloud Skills Boost — Use Machine Learning APIs on Google Cloud (skill badge) · 2026
- NVIDIA Deep Learning Institute — Agentic AI Explained · 2026

TECHNICAL SKILLS

Languages: Python, C, C++, Java, SQL, CUDA C, JavaScript

ML / DL: PyTorch, TensorFlow, Hugging Face Transformers, PEFT/LoRA, BitsAndBytes, PaddleOCR, YOLO, Ollama

Agentic Systems: LangGraph, multi-agent orchestration, Selenium browser automation, LLM tool-calling

Systems: CUDA kernel development, cuBLAS, Triton, GPU profiling, model quantisation, memory-constrained inference

Tools: Git, GitHub, Kaggle, Jupyter, VS Code, Optuna, LongBench, Google Colab, Docker