

Bharat Kaurav

bharat.kaurav.prof@gmail.com • +91-62687-60066 • GitHub • Portfolio • LinkedIn

PROFESSIONAL SUMMARY

Software and AI Engineer with track record of architecting high-scale enterprise platforms and optimizing generative AI pipelines, including a 5x latency reduction for clinical diagnostics and \$300k+ ARR growth for YC-backed startups

WORK EXPERIENCE

• AI Engineer - Contractor

Qure.ai – Research & Optimization Team

Dec 2025 – Present

Bangalore, India (Remote)

- **Redesigned** a sequential 5-stage AI pipeline into an **orchestrated parallel architecture**, implementing **model tiering** for generating Logic Trees to reduce recommendation generation latency from **26s to <5s** and remove LLM API costs
- **Developed** a high-scale **evaluation framework** to validate clinical accuracy (Lung-RADS, Fleischner) across **10k+ test cases**, establishing a repeatable automated testing loop to measure and maintain **100%+ pass rates** during iterations

• Software Engineer - Contractor

Foundation Industries (YC X25) – Full Stack & DevOps Team

Dec 2025 – Mar 2026

San Francisco, USA (Remote)

- **Architected** an enterprise manufacturing platform with **\$300k ARR**, high-performance **React Three Fiber** 3D CAD viewer, robust **multi-file upload pipeline** via **Vercel Blob** for 50MB+ datasets, integrating **Stripe**, **Supabase** for automated workflows
- **Engineered** high-availability infrastructure, reducing bundle sizes by implementing **dynamic imports**, **lazy loading** for heavy libraries; streamlined operations by building automated notification layer using **React Email**, **Resend**, and **Slack Webhooks**

• Software Engineer

Metzev Inc.

May 2025 – Oct 2025

Georgia, USA (Remote)

- **Spearheaded the end-to-end development** of an EV battery platform for **15k+ users**, **multiple enterprise clients**, achieving **\$50k ARR** by architecting a scalable **Stripe/Supabase** infrastructure

• MLOps Engineer

unstudio.ai – Research & Backend Development Team

Nov 2024 – Apr 2025

Gurgaon, India (Remote)

- Deployed a **FLUX personalization pipeline** using **LoRA**, **ControlNets** and **Redis-backed distributed system**, achieving **35% higher image quality**, **60% faster training/inference**, for European furniture clients

• Machine Learning Consultant

Styldod Inc. – Research & Optimization Team

Jun 2024 – Oct 2024

Bangalore, India (Remote)

- Spearheaded diffusion model acceleration via **OneDiff & StableFast** and developed an automated CV pipeline, achieving **50% faster inference**, a **30% GPU reduction**, and a **75% decrease** in manual effort across **300k+ images**

• Projects & Collaborations

- **Virtual Try-On Optimization (Alle, Dec 2024)**: Directed 5-engineer team optimizing LoRA-based FLUX system, delivering **2.5x faster startup** and **98.7 PSNR** on A40/H100 GPUs
- **LLM-SQL Generation (Attentions.ai, Apr 2024)**: Led 7-member team fine-tuning LLaMA, Gemma, Qwen & Phi models for enterprise SQL tasks, achieving **~90% accuracy** on complex benchmarks

PROJECTS

• Dynamic Agentic RAG system

13th Inter IIT-Tech Meet, IIT Bombay

Dec 2024

GitHub

- Architected production-grade **multi-pipeline RAG system** with **3 specialized architectures** (HybridFL, FLARE, DRAGIN) supporting both legal & financial domains with **real-time knowledge graph updates** via Pathway
- Achieved **15-20% performance gains** over GPT-4 and open-source LLMs on custom evaluation datasets; implemented **Docker-containerized microservices** with separate vector stores and **Streamlit production UI**

• Behavior Simulation Challenge

12th Inter IIT-Tech Meet, IIT Madras

Nov 2023

GitHub

- **Bronze Medal (Rank 3/22+ IITs)** - Engineered dual-task ML pipeline: tweet engagement prediction model achieving **96% accuracy** and content generation system using fine-tuned **LLaMA architecture**
- Developed scalable **multimodal data processing pipeline** handling **1M+ heterogeneous records** (text, image, audio, video) with optimized feature extraction and model training workflows

OPEN SOURCE CONTRIBUTIONS

- **EchoMimicV2 & Float**: Developed notebooks to standardize initial model setup and workflow testing
- **PixArt Sigma**: Debugged and fixed errors in the LoRA training script, enabling fine-tuning of models

PUBLICATIONS

- **B. Kaurav, S. S. Dar, A. Jain, C. S. Raghaw, M. Z. U. Rehman, N. Kumar**. “An explainable deep neural network with frequency-aware channel and spatial refinement for flood prediction in sustainable cities.” *Sustainable Cities and Society*, Vol. 130, 2025. <https://doi.org/10.1016/j.scs.2025.106480>
- **Under Review**: Forest-fire detection using a modified YOLO architecture; Landslide area segmentation with remote sensing imagery from Sentinel-2 satellite

TECHNICAL SKILLS

- **Languages & Tools**: Python, C/C++, SQL, Cursor, Claude Code, Shell Scripting
- **ML & Generative AI**: PyTorch, HuggingFace Transformers, Diffusers, LoRA, BitsAndBytes, LangChain/LlamaIndex
- **Web & Backend**: React, Next.js, FastAPI, Flask, Django, PostgreSQL, Stripe, Vercel
- **MLOps & Infrastructure**: Docker, AWS (EC2/S3), GitHub Actions, Redis, Supabase

EDUCATION & ACHIEVEMENTS

- **B.Tech in Computer Science and Engineering**, Indian Institute of Technology Indore
GPA: **8.66/10 (Current)** **Oct 2022 – May 2026**
- Inter IIT Tech Meet 12.0 Bronze Medal; JEE Advanced rank 1083 (out of 150k+ applicants); JEE Mains rank 5995 (out of 1M+ applicants); among top 300 candidates for IPhO India team