Bharat Kaurav

bharatkaurav2211@gmail.com • +91-62687-60066 • GitHub • Portfolio • LinkedIn

Professional Summary

Backend & ML Engineer skilled in building scalable web systems and deploying optimized ML models in production, with hands-on experience in cloud infrastructure, distributed systems, and generative AI.

Work Experience

- **Software Engineer** *May 2025 – Present*
  Metzev – Backend Development & Deployment Team Remote – Georgia, USA
  - Developed a scalable EV battery management platform using **Next.js** and **Supabase**, with real-time data handling and optimizations like **code splitting, lazy loading**, and **database indexing** to boost performance.
  - **Scaled to 15,000+ users**, across **1,000+ stores in USA**, secured **5 major clients**, and ensured platform reliability through performance-focused design and robust backend architecture.
- **Machine Learning Engineer** *Nov 2024 – Apr 2025*
  unstudio.ai – Research & Backend Development Team Remote – Gurgaon, India
  - Engineered a scalable, personalized product training and visualization pipeline for **FLUX** using LoRA fine-tuning, improving image quality and experience for **2000+ users globally**.
  - Built a **distributed training system** using Redis Pub/Sub, with robust error handling, logging, and automated CI/CD via GitHub Actions.
- **Machine Learning Intern** *Jun 2024 – Oct 2024*
  Styldod Inc. – Research & Optimization Team Remote – Bangalore, India
  - Optimized diffusion model inference using Onediffusion and Stablefast, achieving **50% faster inference** and **30% GPU usage reduction**.
  - Developed a scalable pipeline for image collection, quality control, and labeling; tested on **300k+ images**.
- **Additional Projects & Collaborations**
  - **Virtual Try-On Optimization (Alle, Dec 2024)**: Led team of 5 to optimize LoRA-based try-on system for FLUX, achieving **2.5× faster boot-up** and **98.7 PSNR** on A40/H100 GPUs.
  - **LLM-SQL Generation (Attentions.ai, Apr 2024)**: Led team of 5–7 to fine-tune LLaMA, Gemma, Qwen & Phi models for complex SQL query tasks; achieved **90% accuracy**.

Projects

- **Sobra: Soybean leaf disease classification and explanation** *Apr 2025*
  - Implemented a **multimodal AI system** for soybean disease classification using vision-language models and attention-based reasoning; achieved **95% accuracy** on real-world data.
  - Integrated a **RAG pipeline** to suggest symptom-based remedies from damaged leaf images, enabling explainable treatment recommendations.
- **Dynamic Agentic RAG system** *Dec 2024*
  13th Inter IIT-Tech Meet, IIT Bombay
  - Designed a dynamic agentic **RAG system** using **Pathway**, featuring real-time ingestion, tool calling, and modular tool integration.
  - Evaluated on self-curated legal & finance datasets; outperformed closed and open-source LLMs in knowledge-intensive tasks.
- **Behavior Simulation Challenge** *Nov 2023*
  12th Inter IIT-Tech Meet, IIT Madras
  - Built a tweet prediction model with **96% accuracy**; fine-tuned **LLaMA** for engagement optimization tasks.
  - Processed **1M+ multimodal records** (text, image, audio, video) for training and evaluation.

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

Technical Skills

- **Languages & Core:** Python, C/C++, SQL, JavaScript/TypeScript
- **ML Frameworks:** PyTorch, Transformers, BitsandBytes, DINOv2, Diffusers
- **Web Development:** React, Next.js, FastAPI, Flask, Django
- **MLOps & Cloud:** Docker, AWS, GCP, GitHub CI/CD, Redis

Education & Achievements

- **B.Tech in Computer Science and Engineering**, Indian Institute of Technology Indore
  CGPA: **8.44 (Current)** **2022 – 2026 (Expected)**
- **Awards:** Inter IIT Tech Meet 12.0 – Bronze (IIT Madras), IITI SoC – Gold (CyberSecurity), JEE Adv – AIR 1083, JEE Mains – AIR 5995