

Ankur Singh

✉ mleankursingh@gmail.com | 🌐 Ankur-singh.github.io | 📄 github.com/Ankur-singh

ML Engineer with 5+ YoE architecting and delivering production-grade AI/ML solutions. Versatile full-stack developer with deep expertise in GenAI. Proven track record of **leading ML teams**, **establishing engineering best practices**, **data pipelines**, and **driving rapid, iterative development** of high-impact ML features and products in **fast-paced environments**. **Kaggle competitions** and **hackathon winner**, and active **open-source contributor**

TECHNOLOGY

Languages: Python, SQL, JavaScript | **Data:** Ray, Spark, Postgres, MongoDB, BigQuery, AirFlow | **ML/AI:** Scikit, Pandas, Numpy, PyTorch, HF Transformers, OAI Triton | **Backend:** FastAPI, Django, REST, gRPC | **Cloud:** AWS, GCP | **DevOps:** Docker, K8s, CI/CD | **LLMs:** Post-Training, LangChain, LlamaIndex, SmolAgents, MCP, vLLM

EXPERIENCE

AI Solutions Engineer, Intel – Santa Clara, CA May 2023 – July 2025

- **Accelerate GenAI model performance by up to 3x** through profiling and optimizing PyTorch workloads – including distributed training, **PeFT**, **quantization**, and **inference** – on Intel hardware (CPU & XPU)
- **Architected** and developed **modular, microservices-based LLM applications** (Q&A Chatbot, RAG, WebSearch agent, Video Search), with **containerized deployment** via Docker Compose and K8s

Graduate Research Asst., SJSU Research Foundation – San Jose, CA Sep 2022 – May 2024

- **Enabled real-time edge inference** by optimizing YOLOv8 for object detection and segmentation on NVIDIA Jetson (latency \approx 150 ms), supporting autonomous driving use-case
- **Pioneered a Decentralized Traffic Flow Prediction system** using **Federated Learning**, improving privacy preserving forecasting across nodes, published in *IEEE* 📄

Machine Learning (ML) Lead, Zoop.One – Pune, India Sep 2021 – July 2022

- **Saved \$1M+ annually** by leading the design and deployment of 4 core ML microservices housing 20+ CV models; scaled system to serve over **2M monthly requests** through a modular, cloud-native architecture
- **Delivered industry-leading OCR pipeline** (7+ CV models) for ID card parsing with **6× faster response time** (\approx 1s) and superior layout-aware accuracy—beating commercial APIs in both latency and precision
- **Built a real-time liveliness detection service** with face recognition, eye tracking, and spoof detection; achieved **sub-200ms latency**, strengthening anti-fraud protection across customer onboarding workflows

CoFounder and CEO, AI Adventures – Pune, India Aug 2018 – Sep 2021

- **Delivered automation solutions saving 100+ manual hours/week** through diverse ML projects, including **image-based product search**, **receipt digitization**, and **visual defect inspection** for enterprise clients

OPEN SOURCE & ACHIEVEMENTS

- Contributed to high-impact, PyTorch native ML libraries including **torch tune** (LLM post-training) and **fast.ai**
- Authored Python packages **Colab-everything** (50K+ downloads) & **torchserve-client**, simplifying workflows for ML prototyping and model inference
- **Grand Prize Winner**, AI for Social Good **hackathon @ Intel Innovation 2022**
- **3x Kaggle Bronze Medal** – *Shopee: Price Match Guarantee*, *Global Wheat Detection*, & *MoA* competition

PROJECTS

- Developed **PocketPlus AI**, a cross-platform bookmark catalog with LLM-driven summarization, labeling, semantic search via embeddings, and multi-resource chat in collaborative workspaces
- Built **Snap Jobs**, leveraging LLM-based extraction to index 20K+ jobs and fine-tuned LLaMA2-7B for personalized, targeted resume generation

EDUCATION

San Jose State University, MS in Software Engineering Aug 2022 – May 2024 (3.9/4.0)

College of Engineering Pune (COEP), BS in Information Technology Aug 2014 – May 2018 (7.7/10)