Parag Ekbote

ML Engineer Intern

Pune, India | paragekbote23@gmail.com | 087889 54518 | paragekbote.github.io

linkedin.com/in/parag-ekbote | github.com/ParagEkbote

Summary

ML Engineer Intern specializing in LLM fine-tuning and efficient model inference serving. Contributor to Hugging Face, PyTorch, Dagster and Optuna with 110+ merged PRs across 20+ ML repositories and featured work on the Hugging Face blog. Experienced in deploying optimized AI pipelines on the cloud.

Selected Engineering Outcomes Upto 60% lower inference latency • Up to 2× inference throughput • 50% VRAM Usage

Education

Dr. D.Y. Patil Vidyapeeth, B.Tech in AI & Data Science – Pune, India Aug 2023 – Aug 2027

- GPA: 9.1/10
- Recipient: DPU Merit Scholarship (2023–24), DPU Short-Term Studentship Grant (2024–25)

Selected Projects

Dagster HF Datasets

Open-source library for bringing Hugging Face Datasets into Dagster's asset-based orchestration.

- Implemented end-to-end dataset lifecycle management featuring, multi-step data transformations and direct automated publishing to Hugging Face Hub.
- Built custom Dagster asset decorators for Hugging Face datasets, preserving data lineage and partitioning across data pipelines.
- Engineered high-throughput Parquet IO Manager processing 10,000+ rows/sec, supporting materialized and streaming HF datasets, highlighted in the Dagster Community Showcase.

Efficient Model Serving

Production-grade inference deployment pipelines.

- Reduced image generation latency by 60% for A100 GPUs via dynamic LoRA adapter loading using PyTorch, PEFT and Diffusers, featured in a HF blog.
- Deployed 5 quantized LLM pipelines via 4-bit quantization and sparsity, achieving 2× faster inference and 60% VRAM reduction.
- Built the deployment pipeline integrating pytest validation, Docker and Replicate cutting manual deployment effort from 2 hours to 20 minutes.

Open Source Contributions

- Integrated Hugging Face Trackio experiment tracking across 5 ML projects, including ZenML and LlamaFactory, enabling local-first workflows.
- Cut CI runtime by 50% and contributed 5+ tutorials across 3 major repositories counting Pruna AI, skorch and optuna.
- Contributed to Hugging Face Diffusers and to Pruna AI, spanning documentation improvements and CI/CD.
- Complete contribution history: contributions.md

Experience

Technical Reviewer, **O'Reilly Media** – Pune, India July 2026 – present

- Conducted technical accuracy reviews for upcoming AI titles, validating code implementations across 3 chapters.
- Delivered detailed pre-publication feedback to author teams, refining technical depth and code clarity by 20% across modern AI learning materials.
- Submitted chapter reviews within a 2-week turnaround, reducing author revision cycle time by 15% ahead of publication deadlines

Technical Reviewer, **Manning Publications Co.** – Pune, India Mar 2026 – Apr 2026

- Audited PEFT and Transformers-based fine-tuning implementations across 4 manuscript chapters for technical accuracy and reproducibility.
- Identified Hugging Face API patterns shortcoming and optimized code examples, driving a 17% improvement in example accuracy.
- Delivered manuscript reviews within 3-week turnaround per chapter, shortening author revision cycles by 10%.

AI Research Intern, **Green Pepper** – Pune, India Feb 2024 – Dec 2024

- Delivered ML research briefings across 5 enterprise engagements, directly supporting 3 pilot program launches.
- Distilled 30+ research papers in clinical NLP and medical imaging to evaluate state-of-the-art vision-language model architectures.
- Architected reusable GenAI framework covering RAG, prompt engineering, and guardrails, reducing pitch preparation time from 8 hours to 2 hours.

Skills

Programming: Python, PyTorch, Hugging Face (Transformers, PEFT, TRL, Diffusers), scikit-learn, FastAPI

Systems & MLOps: Docker, GitHub Actions, Linux, Git, Make

Cloud: AWS (S3, Lambda, CloudWatch, Sagemaker), GCP, Azure, Replicate

Data Engineering: NumPy, Pandas, Polars, SQL, Dagster, dbt

Training & Inference: Quantization (bitsandbytes, torchao), Axolotl, Optuna, Trackio, Weights & Biases

Publications

Hyperparameter Optimization for Improved Mine Detection Using Machine Learning: A Methodical Analysis of Classification Techniques Nov 2025

The potent performance of ML models is essential for obtaining intelligent insights from the complexity of modern data. To enhance the performance, we explore the usage of hyperparameter optimization across models. In this study, we analyze several hyperparameter optimization strategies that can enhance the performance of the models. The findings are reproducible and reliable, thereby significantly advancing the academic applications by providing a scalable foundation for model building.

Parag Ekbote, Suraj Maharana, Priyanshi Mulwani, Swati Sharma, Dr. Santosh Kumar, Mily Lal

DOI: 10.1109/InCoWoCo68239.2025.11407267 (IEEE 2nd International Conference for Women in Computing (InCoWoCo))

Invited Talks

1. Learning Software Engineering Through Open Source: A Maintainer's Perspective — Pruna AI (June 2026)

Presented a guest lecture at a leading tech startup on maintainer's perspective on open-source software engineering,

1. Learning Software Engineering Through Open Source — Teknikhögskolan (May 2026)

Anchored a guest lecture at a professional upskilling institution on contributing to large OSS projects and long-term contributor growth.