Pratik Korat

San Jose, CA | +1 (408) 646 5708 | korat.pratik1@gmail.com | LinkedIn | GitHub

SUMMARY

Software Engineer (ML) with 2.5+ years of experience shipping and operating production ML systems, specializing in LLM fine-tuning, inference optimization, and automated evaluation pipelines. Proven record of improving reliability, latency, and model quality through hands-on implementation and iterative experimentation.

EDUCATION

M.S., Software Engineering, San Jose State University, California, **GPA: 3.96/4.00** Jan 2024 – Dec 2025

- Related Coursework: Machine Learning | Deep Learning | Natural Language Processing | Reinforcement Learning

B.E., Information Technology, L.D. College of Engineering, India, **GPA: 3.92 / 4.00** Jul 2018 – Apr 2022

SKILLS & EXPERTISE

Languages: Python, C++, SQL
ML Frameworks: PyTorch, TensorFlow, Keras, JAX, Hugging Face Transformers, LangChain, Scikit-Learn, XGBoost
Inference Optimization: vLLM, SGLang, CUDA, Triton, TensorRT, ONNX, TorchScript, llama.cpp, Model Quantization, PTQ
MLOps: GitHub Actions, Slurm, REST, Docker, AWS, Google Cloud (Vertex AI), Qdrant, Apache Kafka, MLflow, Weights & Biases
Applied ML Domains: Generative AI (LLMs, VLMs, RAG), Natural Language Processing, Computer Vision, Reinforcement Learning, Distributed Training, Time Series Modelling, Federated Learning, Performance Profiling

PROFESSIONAL EXPERIENCE

Machine Learning Research Assistant, Research Foundation, SJSU, California Mar 2025 – Present

- Engineered a high-throughput LLM post-training pipeline using GRPO (reinforcement learning), implementing a dual-reward optimization strategy to reduce hallucinations and improve reasoning and consistency during model evaluation
- Implemented a Canvas-integrated LLM grading service using vLLM, adding rate limiting and request queuing to support 100+ faculty users with stable, low-latency performance
- Delivered an LLM-based grading workflow that automated assignment evaluation, reducing turnaround time by 35% while maintaining 95% scoring consistency across courses.

Teaching Associate, Dept. of Computer Engineering, SJSU, California Aug 2024 – Dec 2025

- Mentored 40+ graduate students on ML/DL systems, debugging training instability, optimizing model architectures, and improving experimental rigor in research and project work

System Engineer - Machine Learning, Tata Consultancy Services Jun 2022 – Oct 2023

- Built and integrated a tabular ML training module within an internal AutoML framework, supporting PyTorch and TensorFlow, automating training and Hyperparameters Optimization workflows, and cutting time-to-production by 37%
- Developed and deployed an unsupervised anomaly detection model (VAE-based) on internal production infrastructure using a containerized workflow, reducing manual audit effort by 17% through automated pattern detection
- Implemented automated evaluation and validation pipelines to support rapid model iteration, standardizing metrics and experiment tracking, and cutting evaluation turnaround time by 40%
- Deployed ML models as FastAPI-based inference services with asynchronous request handling and batching, ensuring low-latency predictions under variable production traffic

PROJECT EXPERIENCE

Scratch LLM Inference System from Scratch | FastAPI, Huggingface, Gemma3, LLMs, Transformers Project Link

- Architected a PyTorch-based LLM inference engine for Gemma-3 (270M) from scratch with KV-cache and batched decode scheduling, exposing an OpenAI-compatible FastAPI /v1/chat/completions API; validated production readiness through regression and load testing, sustaining 32 concurrent requests with 0% errors, 100 tokens/sec, and 6s p95 latency

NanoRouter: Edge-Native Agentic Tool Calling Router | Gemma3 1B, LoRA, Unsloth, Llama.cpp

- Fine-tuned a Small Language Model (SLM) using LoRA for edge/on-device inference, implementing a tool-calling LLM with structured outputs and achieving 93% executable action accuracy under on-device performance constraints

Automated Course Quality Audit Platform | Qwen, RAG, Agentic AI, FastAPI, Prompt Engineering

- Built a production LLM application for Quality Matters (QM) course auditing with Canvas LMS integration, RAG-based evaluation, and Qdrant vector database for embeddings and semantic search, delivering structured scoring, evidence retrieval, and recommendations to 30+ DAUs

Real-Time Company Risk Monitoring Pipeline | Kafka, vLLM, Streaming NLP, SQL, Dashboard

- Architected a Kafka-based real-time NLP pipeline ingesting 1K–5K news articles/day and deployed a vLLM-powered LLM inference service for zero-shot multi-label classification with low latency batched inference

PUBLICATIONS

- A Federated Learning Study on Bias and Fairness in Small Data Medical Applications, pdf **IEEE Big Data Service, 2025**
- FedFAME: Data-Free Contrastive Learning Framework for Federated Semi-Supervised Learning, pdf, **ACM SAC 2023**