Dhruv Parmar

AI/ML Engineer

(314) 299-0784 | dhruvprmr0108@gmail.com | San Francisco, CA | https://www.linkedin.com/in/dhruvprmr

Summary

AI/ML Engineer with 4+ years of experience building production search, retrieval, and healthcare ML systems. Specialized in Python, PyTorch, hybrid retrieval, ranking, RAG, and LLM evaluation. Experienced in improving relevance, latency, model performance, and infrastructure efficiency through rigorous experimentation. Proven ability to deploy reliable, scalable ML systems using AWS, Kubernetes, and distributed infrastructure.

Technical Skills

Programming Languages: Python, SQL, Rust, C++, TypeScript, JSON
Machine Learning & Deep Learning: Artificial Intelligence (AI), Machine Learning, Deep Learning, PyTorch, Scikit-learn, Supervised Learning, Feature Engineering, Model Optimization, Model Serving, INT8 Quantization, CUDA, Drift Detection
Generative AI & LLM Systems: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Dense Embeddings, Semantic Search, LLM-Based Grading, Human Feedback, Prompt Evaluation, Evidence Retrieval, Passage Selection
Information Retrieval & Search: Hybrid Retrieval, Vector Search, Lexical Search, Learning to Rank, Cross-Encoder Reranking, Candidate Filtering, Relevance Scoring, Search Indexing, Web-Scale Search, Crawl Freshness Modeling, Ranking Evaluation
ML Evaluation & Experimentation: Offline Evaluation, A/B Testing, Benchmarking, Relevance Metrics, SimpleQA, Error Analysis, Model Validation, Statistical Analysis, Performance and Quality Evaluation, Data Analysis, Model Auditing
Data Engineering & Healthcare AI: Apache Airflow, Databricks, Kedro, ETL/ELT Pipelines, Batch and Streaming Processing, Clinical and Claims Data, REST APIs, OAuth 2.0, Bulk APIs, Data Quality Validation, Great Expectations, FHIR Interoperability, Workflow Orchestration, Data Pipeline Development, API Integration
MLOps & Model Governance: MLflow, Databricks Feature Store, FeatureVault, NannyML, Evidently AI, Experiment Tracking, Model Versioning, Model Monitoring, Automated Testing, Rollback Strategies, AI Governance, Continuous Improvement
Cloud, Infrastructure & Distributed Systems: AWS, Kubernetes, Docker, GitLab CI/CD, Microservices, Distributed Systems, Autoscaling, Observability, API Monitoring, OpenAPI, Scalable System Design, Cloud-Native Architecture
Soft Skills: Cross-Functional Collaboration, System Design, Research-to-Production Execution, Performance Optimization, Ownership, Technical Leadership, Decision-Making, Prioritization, Mentorship, Stakeholder Communication

Professional Experience

AI/ML Engineer
Perplexity | San Francisco, CA — May 2025 – Present

- Improved Python-based hybrid retrieval and ranking pipelines supporting 200 million daily queries, contributing to 93.0% overall SimpleQA accuracy through more accurate candidate selection and evidence ranking.
- Reduced p50 search latency by 30.2% through staged candidate filtering, embedding-based scoring, selective cross-encoder reranking, and systematic performance optimization across high-volume production workloads.
- Reduced embedding storage requirements by 75% through native INT8 quantization, validating retrieval quality against full-precision baselines before deployment across high-volume semantic search workloads.
- Built retrieval and ranking models using Python, PyTorch, dense embeddings, lexical signals, and cross-encoders to improve document- and passage-level evidence selection for production RAG pipelines.
- Developed ML-driven crawl-prioritization models using Python to predict URL importance and update frequency, helping allocate reindexing capacity across a web index exceeding 200 billion URLs.
- Built offline evaluation pipelines using Python, SQL, retrieval benchmarks, LLM-based graders, human judgments, A/B tests, and relevance metrics to establish release criteria for ranking improvements.
- Deployed distributed retrieval services on AWS and Kubernetes, implementing autoscaling, health checks, fault isolation, load balancing, and observability for reliable low-latency request processing at global scale.
- Engineered AWS-based batch and streaming indexing pipelines using Python, transforming crawled documents into lexical, semantic, passage-level, and metadata representations for production retrieval.
- Optimized GPU model-serving infrastructure using Kubernetes, CUDA, Rust, and C++, improving embedding generation and cross-encoder inference under demanding production latency and throughput requirements.
- Strengthened production delivery through CI/CD, containerized deployments, automated testing, versioned indexes, rollback procedures, API monitoring, and REST interfaces documented with OpenAPI specifications.

Machine Learning Engineer
HCL Tech | India Aug 2021 – Jun 2024

- Built Python-based healthcare ML pipelines that generated patient-risk insights for care-management workflows used in product deployments reporting 22% lower hospital readmission rates.
- Automated data validation and feature processing across more than 6,000 healthcare quality rules, improving the reliability of analytics used in deployments reporting a 10% increase in quality-gap closure.
- Standardized model deployment, monitoring, and versioning workflows, reducing production rollout time from 3 weeks to 3 days across client environments.
- Built Python and SQL transformation pipelines for clinical, claims, pharmacy, laboratory, and social-determinants data supporting a healthcare platform used by 37,000+ providers across 1,000+ care settings.
- Implemented OAuth-secured HL7 FHIR and REST API integrations using standardized resources, JSON payloads, and bulk APIs for bidirectional exchange of patient and clinical data with EHR systems.
- Created reusable ML workflows with Kedro, MLflow, Databricks Feature Store, FeatureVault, and Great Expectations for governed feature engineering, experiment tracking, data validation, and reproducible training.
- Deployed healthcare processing workloads with Databricks in isolated AWS-hosted environments, following client requirements for regional data residency, encryption, masking, networking, and access control.
- Orchestrated Python data and model pipelines with Apache Airflow and automated containerized releases through GitLab CI/CD for repeatable deployments across multiple AWS client environments.
- Built AWS serverless APIs that securely exposed patient-risk insights, quality measures, and ML outputs to downstream clinical applications and point-of-care workflows.
- Implemented production monitoring with NannyML and Evidently AI for drift, data quality, model performance, and pipeline health, with threshold-based Slack alerts supporting rapid operational response.

Certifications

- AWS Certified AI Practitioner
- NVIDIA Generative AI LLMs Associate
- Databricks Machine Learning Associate

Projects

Master of Science

Game Node – AI-Powered Gaming Dashboard

- Engineered a personalized game recommendation engine using Steam Web API data—including playtime, achievements, and favorites—with 24-hour caching to improve response time and reduce redundant API calls.
- Developed an AI-powered complaint resolution system that generated contextual recommendations and supported human-escalation workflows, reducing manual triage effort.

Computer Vision Reconstruction Pipeline

- Engineered an end-to-end 3D computer vision pipeline using Python, COLMAP, and Open3D to convert monocular smartphone videos into high-fidelity, simulation-ready 3D assets.
- Implemented automated quality-control and point-cloud processing workflows to identify and resolve visual and geometric defects, improving reconstruction reliability for downstream simulations.

Education

Master of Science in Computer Science
Saint Louis University

Bachelor of Science, Computer Applications
The Maharaja Sayajirao University of Baroda