Rohan Rajendra Patil

AI/ML Engineer

rohanpatil.usa46@gmail.com | +1 (607) 232-2900 | San Francisco, CA | LinkedIn

SUMMARY

- AI/ML Engineer with 5+ years of experience building scalable machine learning infrastructure, data platforms, and production-grade AI systems across large-scale cloud and consumer AI environments.
- Proven expertise in designing batch and streaming pipelines, feature engineering frameworks, and ML workflow orchestration using AWS, Spark, Kafka, Kubernetes, and modern inference stacks.
- Hands-on experience deploying Retrieval-Augmented Generation systems, integrating LLMs, optimizing GPU inference performance, and delivering low-latency AI features in high-concurrency product environments.
- Strong cross-functional collaborator with a track record of translating research and product requirements into reliable, privacy-aware, and observable AI systems that drive measurable business impact.

PROFESSIONAL EXPERIENCE

AI/ML Engineer, Perplexity **06/2024 – Present | CA, USA**

- Architected and deployed Retrieval-Augmented Generation pipelines integrating semantic vector search with contextual web indexing to power grounded summarization within Comet's AI assistant sidebar.
- Integrated on-device lightweight language models for latency-sensitive browsing tasks while dynamically routing complex research queries to cloud-hosted LLM endpoints to maintain sub-second response targets.
- Designed model selection and batching strategies using Triton inference servers and Kubernetes-based GPU nodes, improving throughput efficiency by nearly 25% under high user concurrency.
- Built hybrid retrieval stacks leveraging FAISS and Redis caching to balance recall and precision tradeoffs across active browsing sessions while minimizing redundant web fetch operations.
- Developed passage extraction, ranking, and citation normalization pipelines to enforce source attribution rules, reducing hallucination-prone summaries and increasing factual consistency in A/B evaluations.
- Implemented audio-to-intent pipelines combining speech recognition, intent classification, and NLU components, carefully managing end-to-end latency budgets for conversational browsing workflows.
- Instrumented production-grade evaluation frameworks tracking latency, summary factuality, attribution accuracy, and user satisfaction metrics, enabling data-driven iteration through controlled A/B experiments.
- Built privacy-aware filtering systems incorporating PII detection and secure token management to ensure browser context safety while supporting compliant personalization capabilities.
- Collaborated with frontend, UX, and SRE teams to define telemetry schemas, staged feature rollouts, and observability dashboards using Prometheus and Grafana for production reliability.
- Led rapid 0 to 1 feature experiments for Comet's agentic browsing capabilities, delivering measurable engagement uplift of approximately 18% during initial launch expansion phases.

AI/ML Engineer, Amazon **10/2019 – 06/2023 | India**

- Designed and maintained large-scale batch and streaming data pipelines using AWS S3, Glue, EMR Spark, and Redshift to produce training and inference-ready datasets for forecasting and recommendation systems.
- Built schema-controlled, versioned datasets in Parquet with optimized partitioning strategies, enabling reproducible ML training workflows and improving data retrieval performance by approximately 30%.
- Implemented offline and online feature consistency mechanisms to prevent training-serving skew, ensuring reliable real-time inference inputs across demand forecasting and ranking pipelines.
- Developed feature engineering pipelines generating demand velocity metrics, seasonality indicators, and customer behavioral aggregates consumed by multiple time-series forecasting and optimization models.
- Automated historical backfills and rolling dataset refresh cycles while enforcing strict temporal correctness to eliminate data leakage in multi-horizon forecasting workflows.
- Engineered near-real-time ingestion pipelines using Kafka and Spark Structured Streaming to process operational events such as orders and inventory updates for low-latency feature updates.
- Orchestrated ML data workflows with Step Functions and Airflow, integrating pipelines into SageMaker-based training jobs and reducing data-related model failures by nearly 20%.
- Built model monitoring datasets capturing prediction outputs, feature distributions, and drift signals to support anomaly detection, faster rollbacks, and forecast stability improvements.
- Optimized compute and storage tradeoffs across EMR and Redshift clusters, reducing infrastructure cost by approximately 15% during peak retraining cycles without impacting SLA commitments.
- Partnered closely with Applied Scientists, ML Engineers, and Supply Chain stakeholders to translate evolving forecasting requirements into scalable, production-grade data architectures supporting experimentation at scale.

EDUCATION

Master's in Computer Science **Binghamton, New York**
State University of New York at Binghamton

SKILLS

Programming & Data Languages: Python, SQL, Scala, Java, Bash, Data Structures, Algorithmic Problem Solving, REST API Development.
Machine Learning & Core AI: Supervised Learning, Time Series Forecasting, Ranking Systems, Recommendation Systems, Feature Engineering, Model Evaluation, Statistical Modeling, Experiment Design, A/B Testing.
Generative AI & LLM Engineering: Large Language Models (LLM), Retrieval Augmented Generation (RAG), Prompt Engineering, Embeddings, Context Engineering, Model Selection & Routing, Hallucination Mitigation, LLM Evaluation Frameworks.
Natural Language Processing: Intent Classification, Named Entity Recognition, Conversational AI Systems, Semantic Search, Context Grounding.
Deep Learning & Inference Systems: PyTorch, TensorFlow, Hugging Face Transformers, TorchServe, NVIDIA Triton Inference Server, ONNX.
Search & Vector Infrastructure: FAISS, Pinecone, Hybrid Search, BM25, Vector Indexing, Retrieval Optimization, Redis Caching.
Data Engineering & Streaming: Apache Spark, Spark Structured Streaming, Apache Kafka, Apache Airflow, AWS Glue, ETL Pipelines, Batch & Real Time Processing, Parquet, Data Versioning
Cloud & Platform Engineering: Amazon Web Services (S3, EC2, Redshift, DynamoDB, SageMaker, Lambda), Kubernetes, Docker, Terraform, GPU Compute Infrastructure.
MLOps & Production AI Systems: Model Versioning, Experiment Tracking, ML Workflow Orchestration, Data Drift Detection, Automated Retraining, CI/CD Pipelines, Blue Green Deployment, Observability (Prometheus, Grafana).
Distributed Systems & Scalability: Distributed Systems Design, Low Latency Inference Optimization, Autoscaling, Fault Tolerant Architecture, High Availability Systems

CERTIFICATIONS

AWS Certified Cloud Practitioner
AWS Certified AI Practitioner
AWS Certified Machine Learning Engineer – Associate