MLOps at Scale: Production-Ready AI Systems for HR Transformation

A technical blueprint for deploying generative AI in HR shared service centers while maintaining operational excellence

BY:- Ramprasad Reddy Mittana

NSF International

Conf42 MLOps 2025

Session Overview

1

MLOps Architecture

Real-world pipelines supporting 10M+ monthly HR interactions

2

Performance Monitoring

Drift detection systems & custom HR metrics

3

Data Engineering

Privacy-preserving pipelines & real-time feature engineering

4

Scalable Deployment

Multi-tenant patterns serving 50,000+ employees

5

Governance & Excellence

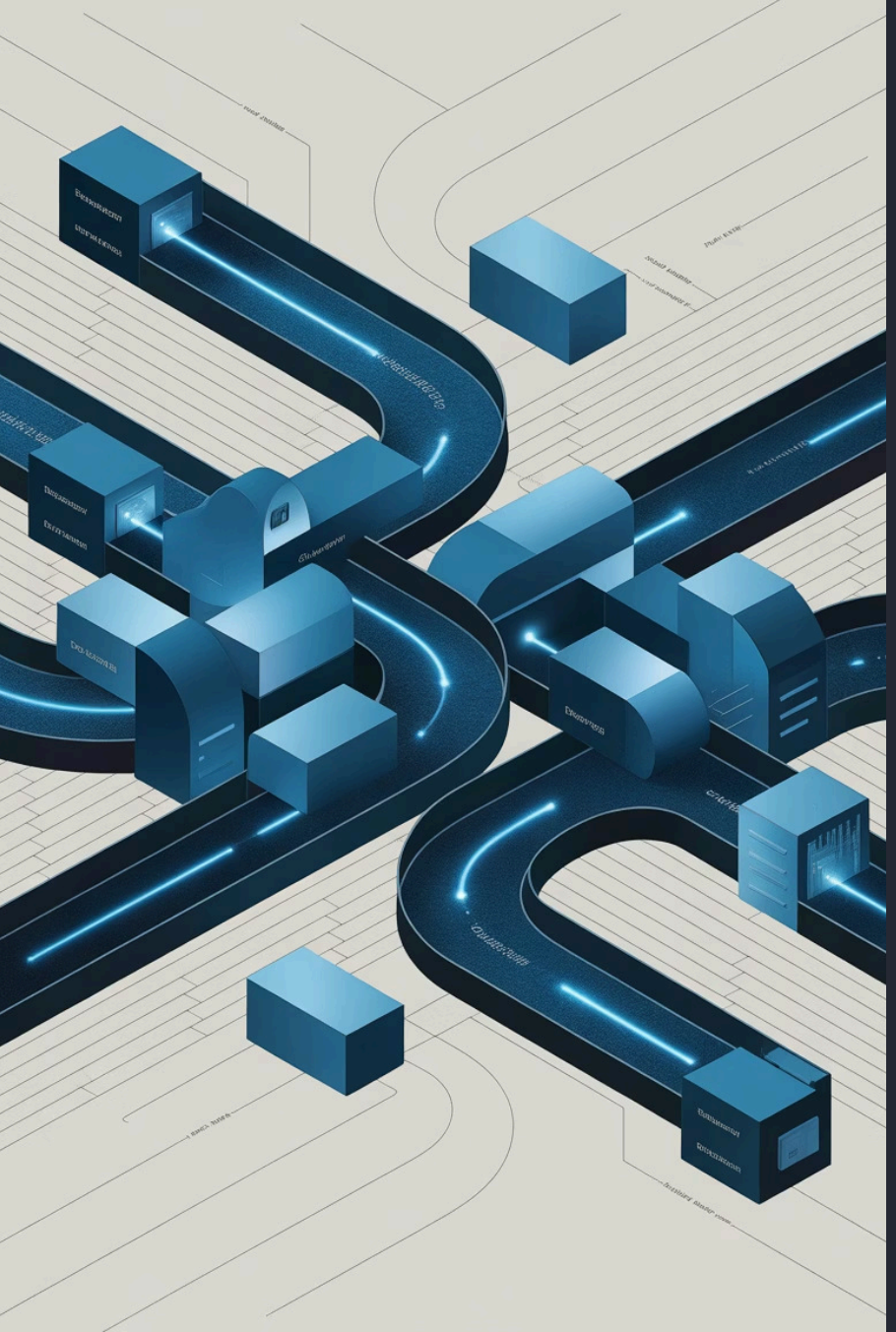
Audit trails, bias detection & compliance monitoring

The HR AI Production Challenge

Enterprise HR shared service centers face unique MLOps challenges:

- Processing sensitive employee data across jurisdictions
- Handling seasonal workforce fluctuations
- Maintaining consistent performance across diverse query types
- Meeting strict compliance requirements in regulated industries
- Scaling to support 10M+ monthly interactions

Traditional MLOps approaches fall short in these high-stakes environments



Production MLOps Architecture

Containerized Model Deployment

Kubernetes-orchestrated containers with GPU acceleration and resource optimization

A/B Testing Framework

Automated canary deployments with statistical significance testing for HR chatbots

Automated Retraining

Scheduled pipelines maintaining 92% accuracy across diverse query types

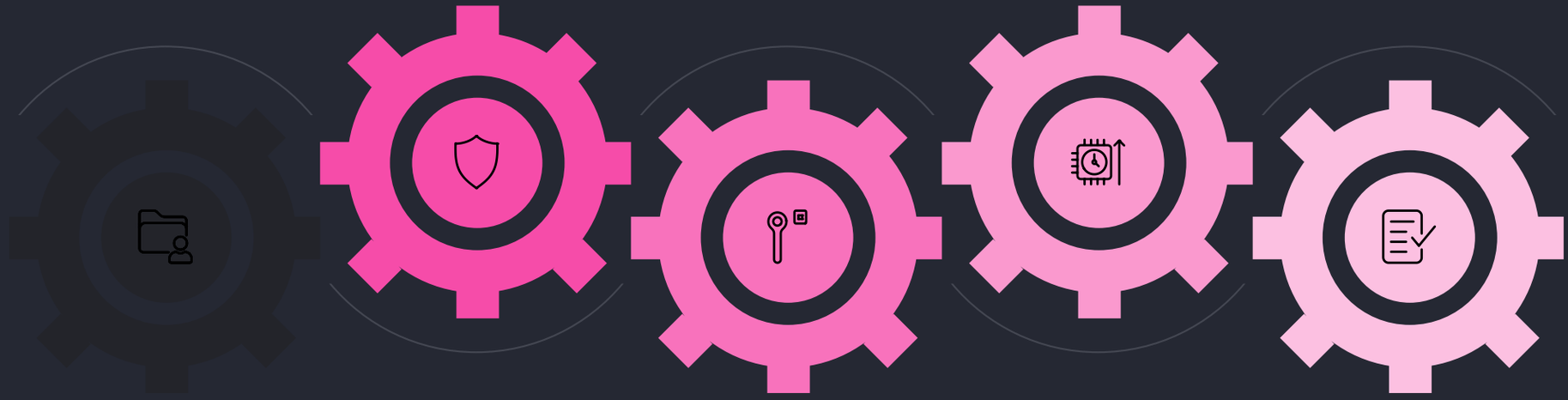
MLOps Pipeline Workflow

Data Validation

Schema, privacy, quality gates

Model Training

Train models with tracked experiments



Data Ingestion

HR, chatbot logs, ticketing systems

Feature Engineering

Create and transform model features

Evaluation & Approval

Assess performance and approve models

Our end-to-end pipeline ensures consistent model quality while enabling rapid iteration

Model Performance & Monitoring

Comprehensive Monitoring Strategy

- Real-time performance dashboards tracking key HR metrics
- Drift detection identifying degradation 73% faster
- Custom metrics for HR-specific applications
- Automated alerting maintaining 99.5% SLA compliance

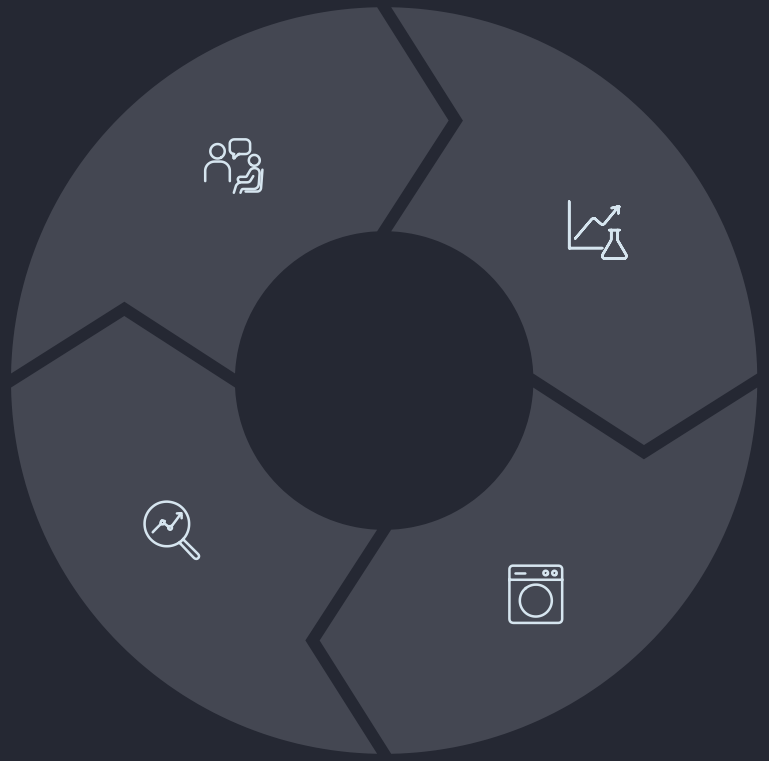
Continuous Feedback & Improvement

User Interactions

Capture HR queries, responses, and satisfaction ratings

Performance Gains

28% accuracy improvement through continuous learning



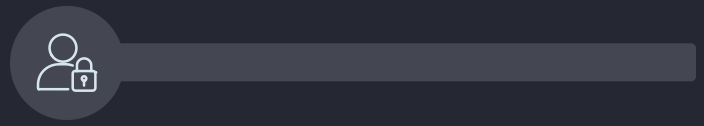
Analysis

Identify patterns in failed interactions

Model Updates

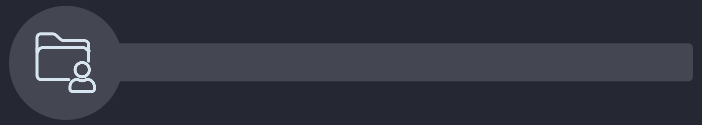
Retrain with supervised examples

Data Engineering for HR AI



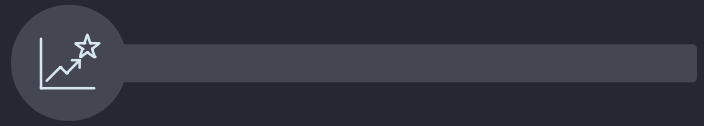
Privacy-Preserving Pipelines

Automated PII detection and tokenization for GDPR/CCPA compliance



Data Quality Validation

Schema enforcement and anomaly detection achieving 99.2% accuracy



Real-Time Feature Engineering

Streaming architecture reducing inference latency by 65%

Technical Implementation: Feature Store

Our HR feature store enables:

- Consistent features across training and inference
- Point-in-time correctness for time-series HR data
- Low-latency serving (<10ms) for real-time interactions
- Feature versioning and lineage tracking
- Automatic feature computation with Spark/Flink

Results: 65% latency reduction and 40% computational cost savings

Scalable Deployment Strategies

Multi-Tenant Architecture

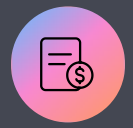
Isolated namespaces with shared infrastructure serving 50,000+ employees

Blue-Green Deployment

Zero-downtime model updates with automated rollback capability

Auto-Scaling Configuration

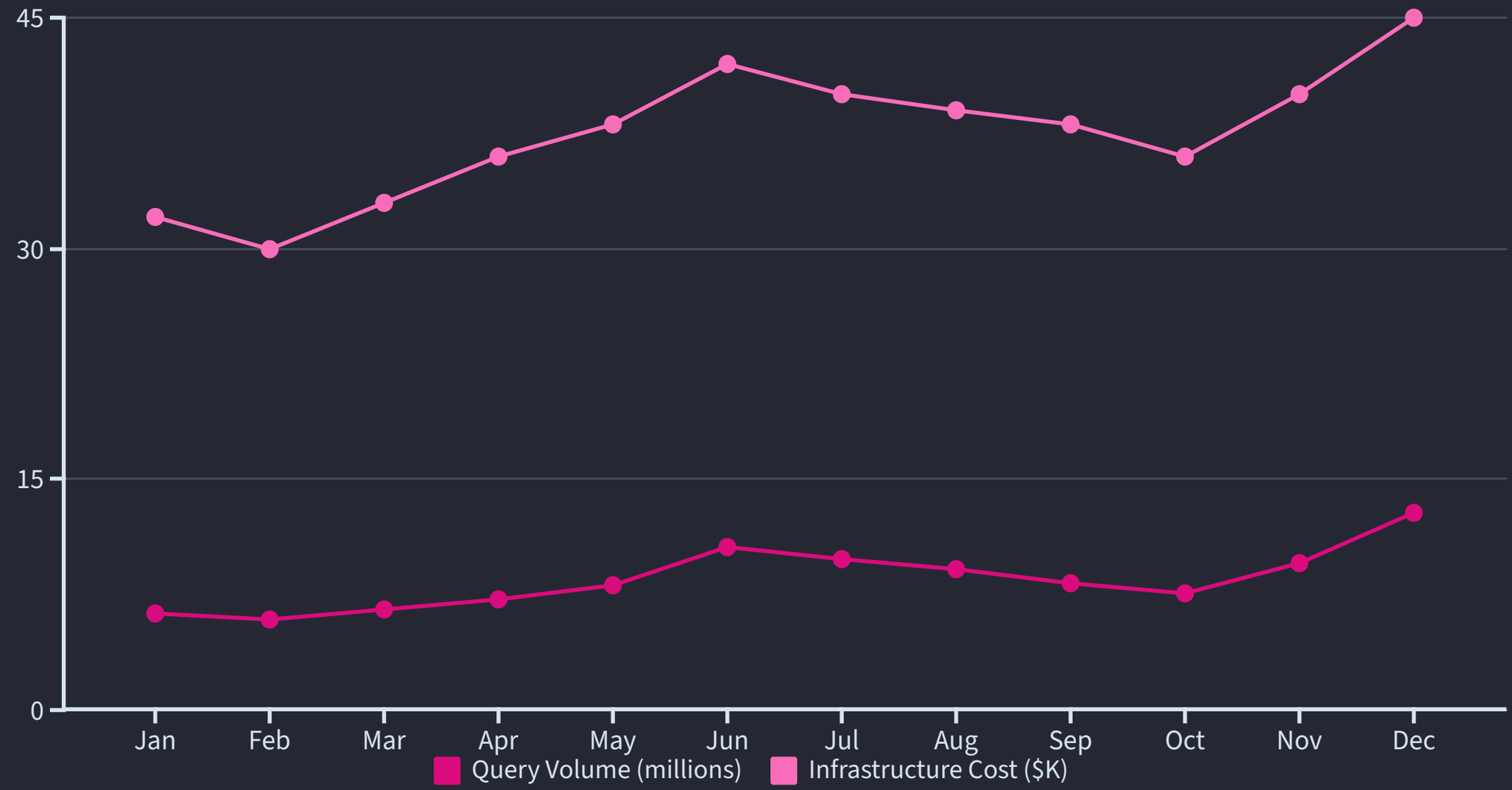
Elastic infrastructure handling 10x traffic spikes during peak periods



Cost Optimization

Spot instances and right-sizing reducing infrastructure costs by 40%

Seasonal Scaling Patterns



Our auto-scaling infrastructure adapts to predictable HR cycles (benefits enrollment, performance reviews, fiscal year-end) while optimizing costs

Operational Excellence & Governance

MLOps Governance Framework

- Comprehensive audit trails for all AI decisions
- Automated bias detection reducing discriminatory outcomes by 89%
- Compliance monitoring across multiple jurisdictions
- Model explainability for HR-specific use cases

Bias Detection & Mitigation

Our automated frameworks have reduced discriminatory outcomes by 89% while maintaining model performance

Key Takeaways

Architecture

Implement containerized deployments with automated retraining pipelines to handle diverse HR query types and seasonal fluctuations

Monitoring

Develop HR-specific metrics and drift detection systems that identify performance degradation 73% faster

Data Engineering

Build privacy-preserving pipelines with real-time feature engineering to reduce latency by 65% while maintaining compliance

Scalability

Design multi-tenant architectures with blue-green deployment and auto-scaling to handle 10x traffic spikes

Governance

Implement comprehensive audit trails and bias detection systems to ensure fair, compliant AI operations

Thank You