Scaling Conversational AI in Production: MLOps Strategies for Contact Center Transformation and Operational Excellence

Transforming contact centers through production-ready conversational AI systems that combine human expertise with machine intelligence while maintaining operational reliability.

By:- Manoj Kumar Vunnava

Godaddy

Conf42 MLOps

Agenda

1 **Contact Center AI Challenges**

The unique operational demands of conversational AI in high-volume customer environments

2 **MLOps Framework for Conversational AI**

Building robust training pipelines, monitoring and deployment strategies

3 **Technical Implementation**

Architecture, CI/CD pipelines, feature stores, and model serving patterns

4 **Measuring Success**

Balancing technical metrics with business KPIs and continual improvement

Contact Center AI: A Unique MLOps Challenge

Scale Requirements

Millions of daily interactions requiring sub-second response times

Human-in-the-Loop Complexity

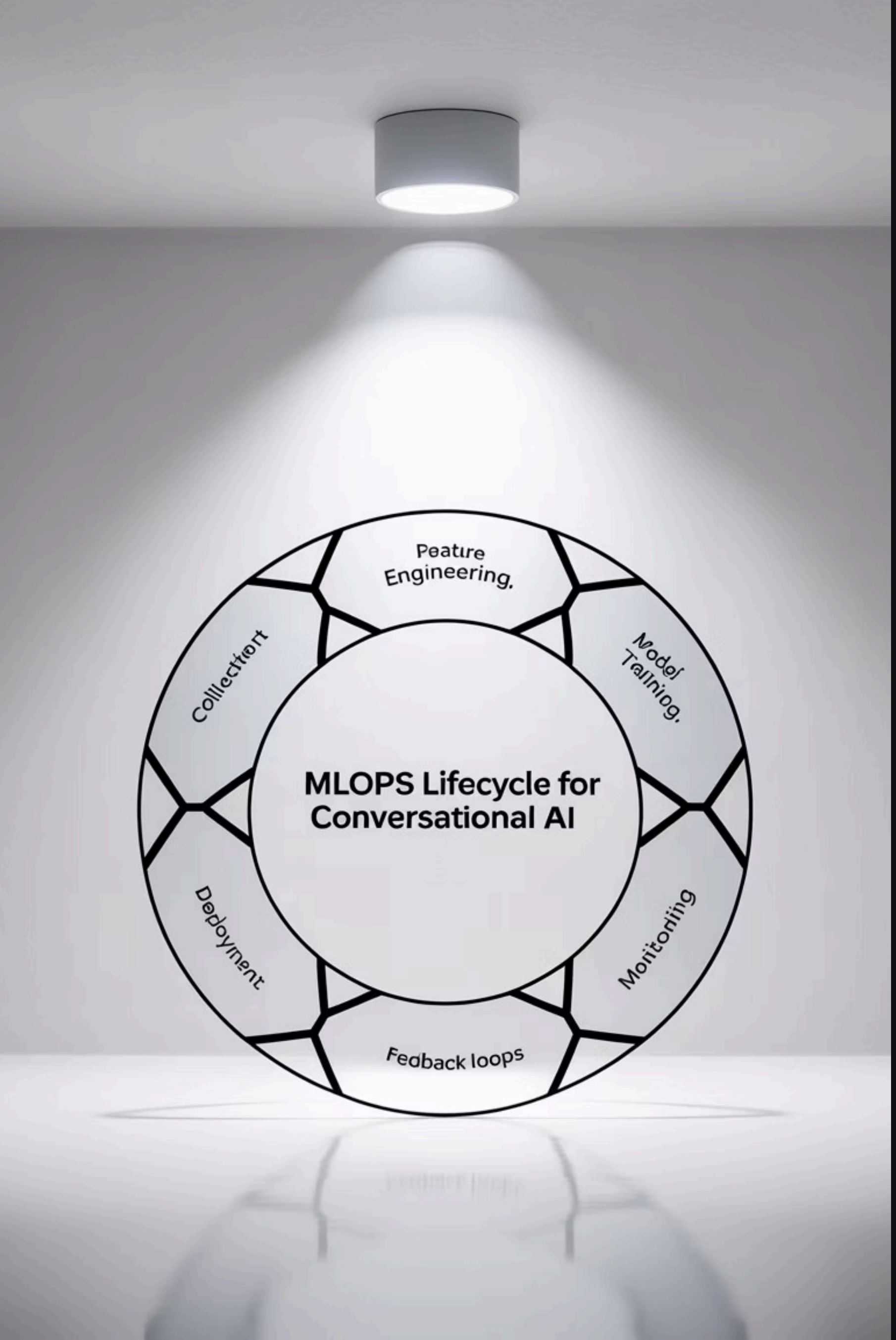
Models must seamlessly integrate with human agents

Cost of Errors

Poor predictions directly impact customer satisfaction and business metrics

Domain Complexity

Diverse language patterns and specialized terminology across industries



A Comprehensive MLOps Framework for Conversational AI

Feature Engineering for Conversational AI

Contextual Features

- Conversation history embeddings
- Session-level metadata
- User profile information

Temporal Features

- Response timing patterns
- Service level indicators
- Historical interaction frequency

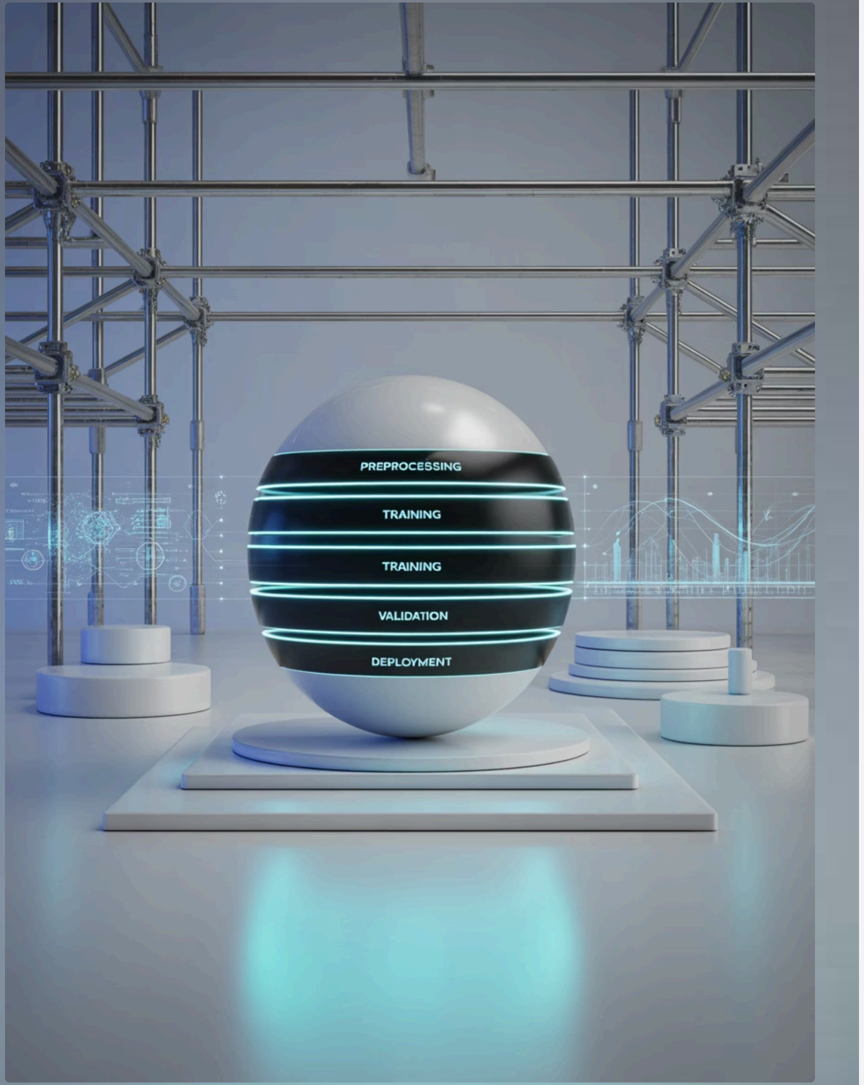
Language Features

- Domain-specific embeddings
- Sentiment analysis scores
- Intent classification signals

Feature Store Requirements

Implement a robust feature store that provides:

- Feature versioning and lineage tracking
- Online/offline consistency
- Low-latency feature retrieval (<100ms)
- Support for both batch and streaming features



Robust Training Pipelines for Conversational AI

Data Preparation

- Automated data quality checks
- Synthetic data generation
- Bias detection routines

Training Infrastructure

- Containerized training jobs
- Experiment tracking
- Hyperparameter optimization

Validation Gates

- Automated test suite
- Fairness metrics
- Business KPI simulations

A/B Testing Framework for Contact Center AI

1

Traffic Allocation

Granular control over which interactions use candidate models, with progressive rollout patterns

2

Shadow Mode Testing

New models run in parallel with production models, recording predictions without affecting customers

3

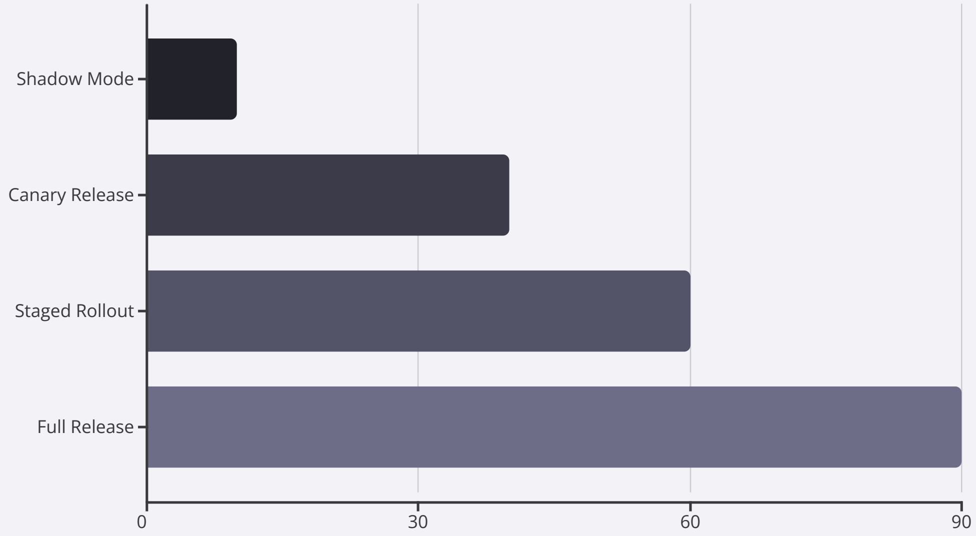
Agent Feedback Loop

Human agents provide rapid feedback on model quality during controlled rollouts

4

Automatic Rollback

Real-time monitoring triggers instant rollbacks if performance degrades beyond thresholds



Testing strategies vary in risk level. Start with the lowest risk approach and progressively increase exposure.

Model Monitoring Strategies for Contact Center AI

1 Technical Metrics

- Inference latency (p50, p95, p99)
- Request throughput
- Model confidence scores
- Error rates and exceptions

2 Data Quality

- Feature drift detection
- Input distribution monitoring
- Missing value rates
- Outlier detection

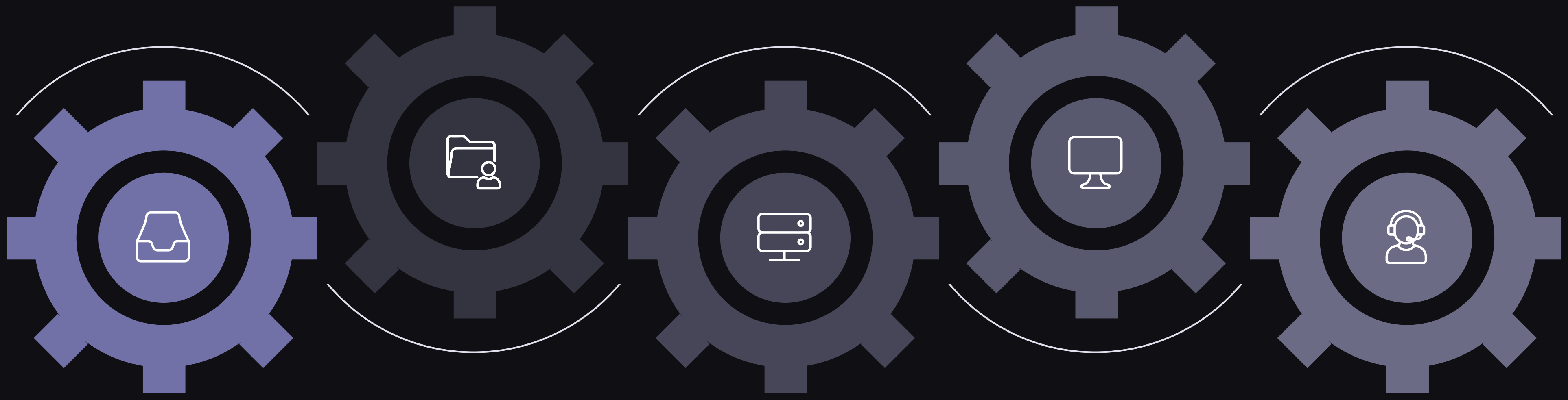
3 Model Performance

- Accuracy & precision metrics
- Agent override rates
- Conversation length changes
- Automatic ground truth collection

4 Business Impact

- Resolution time
- Customer satisfaction scores
- Agent efficiency metrics
- Automation success rate

Architectural Patterns for Serving Conversational AI



Feature Store

Store precomputed features and context

Monitoring

Metrics, logs, and alerting

Data Ingestion

Collect calls, chats, and events

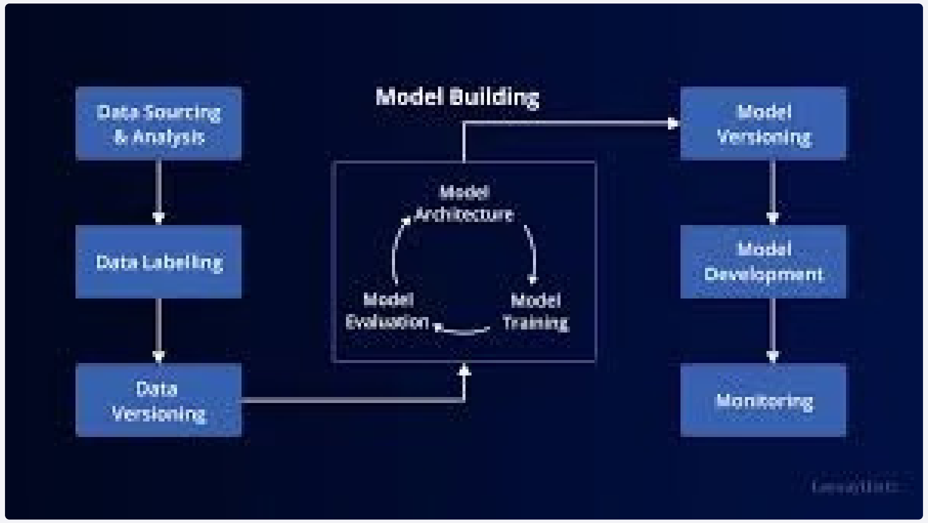
Model Serving

Real-time inference and routing

Agent Workstation

Integrated UI and suggestions

CI/CD Pipelines for Conversational AI Models



> "A robust CI/CD pipeline reduces model deployment time from weeks to hours while improving reliability and traceability."

Version Control

Model code, configs, and feature definitions in Git with branch protection

Automated Tests

Unit tests, integration tests, and model-specific performance tests triggered on commit

Model Registry

Centralized artifact storage with versioning, signatures, and metadata

Deployment Automation

Containerized model serving with gradual traffic shifting and automated canary analysis

Data Drift Detection Systems

Common Drift Patterns in Contact Centers

Seasonal Patterns

Holiday periods, product launches, and promotional campaigns cause predictable shifts in customer inquiry patterns.

Emergent Issues

Service outages, product defects, or viral social media posts create sudden spikes in specific inquiry types.

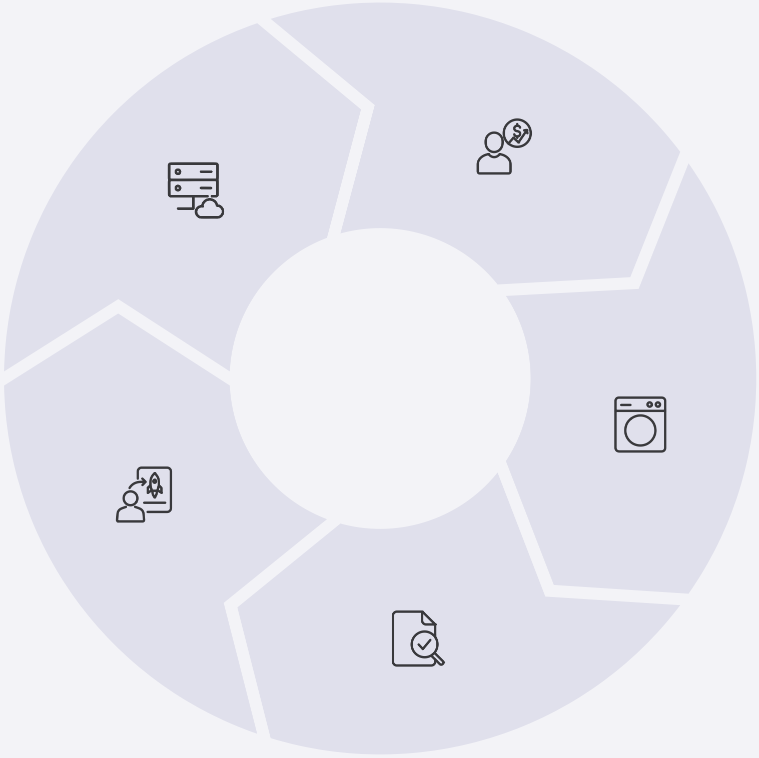
Gradual Evolution

Customer language patterns, product terminology, and issue types evolve slowly over time.

Implementation Approaches

- Statistical distance metrics (KL divergence, Jensen-Shannon) for numerical features
- Embedding space monitoring for text inputs
- Topic modeling to detect emerging conversation clusters
- Anomaly detection on model confidence scores
- Agent override frequency monitoring by segment

Automated Model Retraining Workflows



Data Collection

Agent-verified interactions are automatically tagged and stored

Drift Analysis

Statistical monitoring identifies when retraining is needed

Model Training

Automated training jobs with version control and experiment tracking

Validation

Multi-stage testing against historical and shadow mode data

Deployment

Gradual rollout with automated performance monitoring

Key Takeaways: MLOps for Contact Center AI

Robust Feature Engineering

Invest in a high-performance feature store that handles both contextual and temporal features with low latency

Progressive Deployment

Implement shadow mode testing and canary releases to manage risk during model updates

Multi-level Monitoring

Monitor technical metrics, data quality, model performance, and business KPIs with automated alerts

Continuous Improvement

Establish feedback loops with automated retraining workflows triggered by drift detection

"Successful conversational AI doesn't replace human agents—it enhances their capabilities by handling routine tasks and providing intelligent assistance."

Thank You