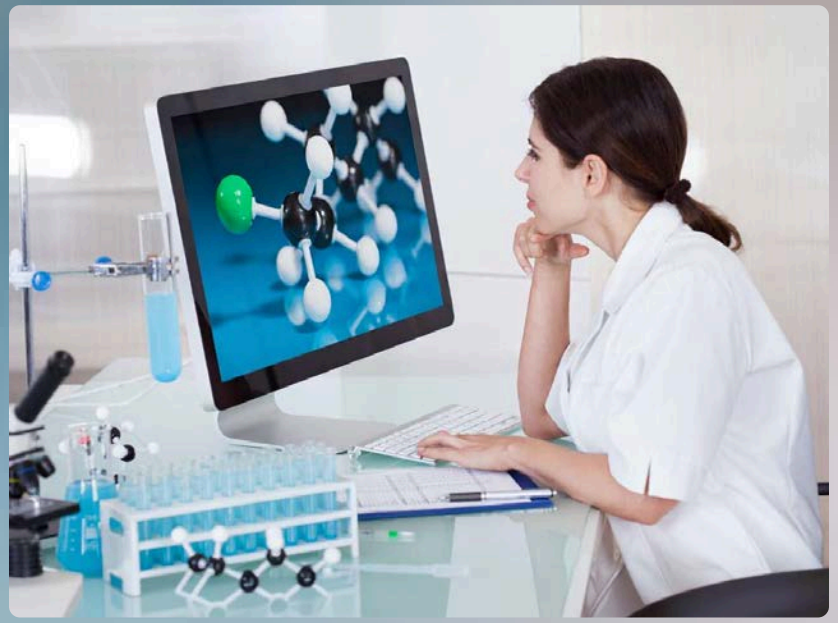


Deploying AI Literature Agents at Scale: MLOps Strategies for Biomedical Research Platforms

By Nishanth Joseph Paulraj

Western Governors University, Utah, USA

Conf42.com MLOps 2025

The Challenge: Information Overload in Biomedical Research

The biomedical research community faces an unprecedented data challenge:

- Over **1 million new publications** annually
- More than **30 million scientific citations** to process
- Traditional search tools failing to connect related research
- Critical insights buried within specialized language
- Researchers spending ~80% of time just finding relevant literature

Our Solution: Intelligent AI Literature Agents

1

Domain-Adapted LLMs

Foundation models specifically trained for biomedical terminology and intricate relationships.

2

Advanced RAG System

Vector-based retrieval enabling semantic understanding across 30+ million research papers.

3

Precision NER Systems

Custom Named Entity Recognition for genes, proteins, and drug interactions, achieving 90%+ accuracy.

4

Robust MLOps Infrastructure

Containerized pipelines, automated retraining, and comprehensive monitoring for continuous improvement.

This system provides sub-second query responses, ensuring rigorous scientific accuracy across complex biomedical domains.

Agenda

01

MLOps Architecture Overview

Unpacking core components and system design principles

02

Biomedical NLP Pipeline

Deep dive into specialized training and deployment workflows

03

Vector Database Scaling

Strategies for efficiently handling millions of research documents

04

Monitoring & Observability

Implementing real-time tracking and robust performance management

05

Continuous Experimentation

Establishing A/B testing frameworks and rigorous scientific validation

06

Deployment Strategies

Unveiling effective infrastructure patterns and practical implementation tips

MLOps Architecture Overview

Our production system architecture is designed for optimal performance, scientific accuracy, and operational efficiency, featuring:

- Containerized microservices for modular scalability
- Asynchronous processing pipelines for efficient batch citation updates
- A real-time inference API layer engineered for sub-second responses
- A versioned model registry with automatic rollback for reliability
- Clear separation of embedding generation from retrieval services
- Dedicated evaluation environments to ensure rigorous scientific validation

Biomedical NLP Pipeline: Domain Adaptation Challenges

Biomedical language presents unique MLOps challenges:

- Highly specialized vocabulary across sub-disciplines
- Entity relationships requiring domain expertise (e.g., gene-protein interactions)
- Contextual meaning that changes across research areas
- Need for continual updates as new discoveries emerge

Traditional NLP pipelines fail in this domain without specialized adaptation techniques.

Named Entity Recognition

90%+ accuracy for genes, proteins, compounds

Relation Extraction

Identifying complex biomedical relationships

Citation Network Analysis

Understanding research lineage and influence

Model Training & Deployment Workflow

Data Ingestion

Automated publication ingestion pipeline processes 5,000+ new papers daily

Training Pipeline

Containerized training with domain-specific fine-tuning and scientific validation

Evaluation Framework

Specialized metrics for biomedical accuracy with domain expert review

Deployment

Automated canary deployments with graduated traffic shifting and rollback triggers

Scaling Vector Databases for 30+ Million Citations

Operational Challenges

- Efficiently generating embeddings for massive document corpora
- Balancing recall against computational costs
- Managing seamless index updates without downtime
- Scaling retrieval mechanisms for concurrent user queries

Our MLOps Solutions

- Asynchronous embedding generation pipelines with batch processing
- Hierarchical indexing strategies with domain-specific sharding
- Read replicas with staged index updates for seamless transitions
- Optimized query caching architecture with semantic similarity matching

This robust approach enabled a 95% reduction in query latency and effortlessly managed a 3x increase in document volume, all without requiring additional infrastructure.

RAG Architecture: Production Implementation

Key Components

- Domain-specific embedding models
- Multi-stage retrieval pipeline
- Citation network enrichment
- Context window optimization

Performance Metrics

- Sub-second query latency (P95)
- Over 90% biomedical entity recognition accuracy
- 85% human-evaluated citation relevance score
- 80% reduction in researcher literature search time

Monitoring & Observability Framework

Model Performance Tracking

- Entity recognition precision/recall by type
- Embedding quality drift detection
- Response latency profiling
- Token usage optimization

Scientific Accuracy Validation

- Citation relevance scoring
- Expert evaluation feedback loops
- Literature coverage metrics
- Factual consistency checks

User Interaction Analysis

- Query pattern monitoring
- Session-based success metrics
- Research journey tracking
- Feedback integration pipeline

Our observability stack integrates MLflow, Prometheus, and custom biomedical validation frameworks, providing both technical and domain-specific insights.

Real-time Monitoring Dashboard

Our custom monitoring solution provides both ML engineers and scientific stakeholders with visibility into:

- Model drift detection with automated alerts
- Citation accuracy by research domain
- System health and performance metrics
- User query patterns and success rates
- Resource utilization and scaling triggers
- A/B test performance comparisons

Continuous Experimentation Methodology

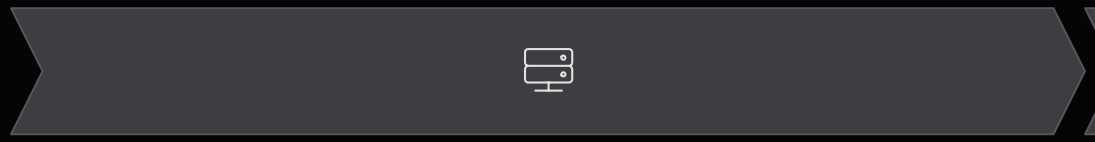
A/B Testing Framework

We've implemented a specialized scientific A/B testing framework that:

- Balances experimentation with scientific reliability
- Enables targeted cohort testing by research domain
- Measures both technical metrics and research outcomes
- Provides statistical confidence for biomedical applications
- Supports multi-variate testing across model components

This approach has yielded an 80% reduction in literature search time while maintaining scientific rigor in results.

Infrastructure Scaling Patterns



Baseline Infrastructure

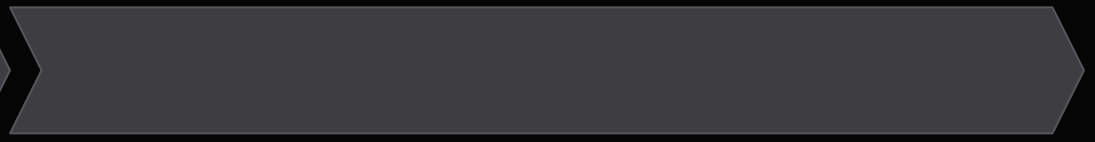
Core components sized for average load with 2x redundancy

Elastic Scaling Layers

Auto-scaling inference endpoints based on query volume

Scheduled Scaling Events

Predictive scaling for known research patterns



Specialized Compute

GPU allocation for embedding generation and inference

Our infrastructure scales efficiently across both batch processing needs (literature updates) and real-time query patterns, maintaining SLAs while optimizing costs.

Key MLOps Learnings for Biomedical AI

Integrating Domain Expertise

MLOps teams must include both ML engineers and domain scientists to build effective biomedical AI systems.

Developing Scientific Validation Pipelines

Standard ML metrics are insufficient; specialized evaluation frameworks are crucial for ensuring scientific accuracy.

Enabling Continuous Corpus Updates

Design for constant knowledge integration as biomedical research evolves, moving beyond static, point-in-time deployments.

Prioritizing Explainability

Biomedical research applications necessitate significantly higher transparency and interpretability standards compared to consumer AI systems.

Thank You !