DOSOBE.

Production AI & Observability

From Data Science Lifecycle to Runtime Monitoring of Agentic Systems

DOSOBE Research White Paper — June 2026
Classification: Public

Table of Contents

1. Executive Summary

2. The Data Science Lifecycle for Production AI

3. ML Observability for Agentic Systems

4. Multi-Agent Data Governance

5. MLOps and Infrastructure

6. Recommendations

1. Executive Summary

Deploying AI systems in production — particularly agentic AI systems that act autonomously — requires fundamentally different approaches than developing models in research environments. The gap between a model that achieves 95% accuracy on a benchmark and a system that operates reliably in production across millions of decisions is vast and often underestimated.

This white paper examines the full lifecycle of production AI: from data preparation and model development through deployment, monitoring, and continuous improvement. We focus specifically on the unique challenges introduced by agentic AI systems that traditional MLOps practices do not address.

The insights draw from DOSOBE deployments across financial services, healthcare, and industrial automation.

2. The Data Science Lifecycle for Production AI

The traditional data science lifecycle — business understanding, data acquisition, modelling, evaluation, deployment — was designed for project-based analytics, not production AI systems. For production agentic AI, the lifecycle must be extended with production-specific stages.

2.1 Production-Ready Data Pipeline

Research models typically use curated, static datasets. Production agentic AI systems consume live, streaming data that is noisy, incomplete, and subject to distribution shift. A production-ready data pipeline includes schema validation, anomaly detection, automated quality scoring, and lineage tracking.

2.2 Model Serving Architecture

Agentic AI systems require inference at unpredictable times with low latency. Model serving must support horizontal scaling, model versioning with safe rollback, A/B testing of agent behaviours, and request-level monitoring.

> *A model that works on your laptop is a prototype. A model that works in production is an infrastructure.*

3. ML Observability for Agentic Systems

Traditional observability — metrics, logs, and traces — was designed for deterministic systems where failure modes are known. Agentic AI systems violate every assumption: behaviour is non-deterministic, failure modes are emergent, and internal state is distributed across model weights, context windows, and tool call histories.

3.1 The Three Pillars Extended

Traces: Complete decision chains across agent calls including model inputs/outputs, tool invocations, context states, and confidence scores.

Metrics: Aggregate behaviour monitoring — decision distributions, confidence trends, escalation rates, latency profiles.

Logs: Raw interaction data for forensic analysis.

3.2 Agent-Specific Monitoring

- **Semantic Monitoring**: Tracks whether agent outputs remain within expected semantic boundaries
- **Behavioural Drift Detection**: Identifies shifts in agent decision patterns over time
- **Failure Simulation**: Tests agent behaviour under controlled anomalous conditions

4. Multi-Agent Data Governance

In multi-agent systems, data governance becomes a runtime capability. Each agent may ingest, transform, and pass data to downstream agents. Our framework assigns ownership at ingestion, tracks transformations through a provenance graph, and evaluates quality at each consumption point.

> **Key Principle**
>
> Data quality must be assessed at consumption, not just at ingestion. What was high-quality data at entry may be stale or corrupted by the time an agent acts on it.

5. MLOps and Infrastructure

5.1 Continuous Training and Deployment

Agentic AI systems require continuous model updating as new data and feedback become available. CI/CD pipelines must support model retraining, evaluation against production baselines, and staged rollouts.

5.2 Incident Response

When an agentic AI system behaves incorrectly, the response must be automated and immediate. Our incident response framework includes automated detection, root cause analysis using provenance data, autonomous mitigation, and post-mortem analysis.

6. Recommendations

Capability	Implementation	Priority
Data Pipeline	Schema validation + anomaly detection at ingestion	Critical
Model Serving	Containerised deployment with A/B testing	High
Observability	Extended three pillars + semantic monitoring	High
Data Governance	Provenance graph + quality-at-consumption	Medium
Incident Response	Automated detection + mitigation + post-mortem	High

DOSOBE — dosobe.com

Technology is not the goal. Human performance is.