mona

Buyer's Guide for AI & ML Monitoring

Chapter 1: Key Considerations for Evaluating AI/ML Monitoring Solutions

Introduction and Purpose

Monitoring is essential for maintaining the performance, trust, and compliance of AI/ML systems in production. This guide outlines key evaluation criteria for selecting a monitoring solution that fits your operational goals, infrastructure, and business priorities. We'll cover core functionality, integration, scale, pricing, and the importance of long-term vendor partnerships. Let's get into it.

Core Monitoring Capabilities

Model Performance Monitoring

Real-world performance is the only performance that matters. Even well-trained models degrade due to drift, changing conditions, or unforeseen anomalies. Monitoring ensures outputs remain accurate and aligned with benchmarks, whether they're based on accuracy, precision-recall tradeoffs, latency, or business impact. Without it, risks like poor predictions, regulatory issues, or financial loss increase.

The right solution enables visibility into real-time and long-term behavior and allows teams to define and monitor custom KPIs. Look for support for both prebuilt and custom metrics, automatic detection of changes across segments, version comparisons, drift detection, and fairness analysis. A solution should isolate discrepancies, uncover bias, and support adherence to AI fairness guidelines.

Key Considerations

Metrics

What metrics will we need to monitor to detect degradation and anomalies before they impact model performance? (i.e. functionality/quality, system health)

Will our monitoring system allow customizing metrics and parameters to suit our specific needs?

Automatic Tracking

How can this monitoring system automatically track changes in the behavior of monitored metrics in aggregate and in each relevant data segment?

Can we get early notification of outliers, sudden changes, gradual changes, changes across versions, and changes between training and inference?

Bias & Fairness

How would we isolate and report on model behavior discrepancies along pre-defined dimensions?

How would we identify biases?

How would we ensure our system adheres to relevant AI fairness laws and guidelines?

Drift Detection

How will this monitoring system detect data and concept drift?

Pro Tip: Make sure your solution supports both predefined metrics and custom KPIs tailored to business needs.

Data Observability

Accurate models depend on trustworthy data. If training or inference data is missing, inconsistent, or drifting, model outputs will suffer. Data observability tools should identify issues such as missing values, schema shifts, or pipeline failures across both upstream and downstream dependencies. Continuous monitoring reduces the risk of "silent" model failure. Strong solutions reveal where data diverges between training and inference and surface early indicators of degraded performance.

Key Considerations

Data Quality

Can the monitoring system automatically detect and alert us on missing values, anomalies, and errors?

Training vs. Inference

How can we detect discrepancies between training and inference data that would impact model performance?

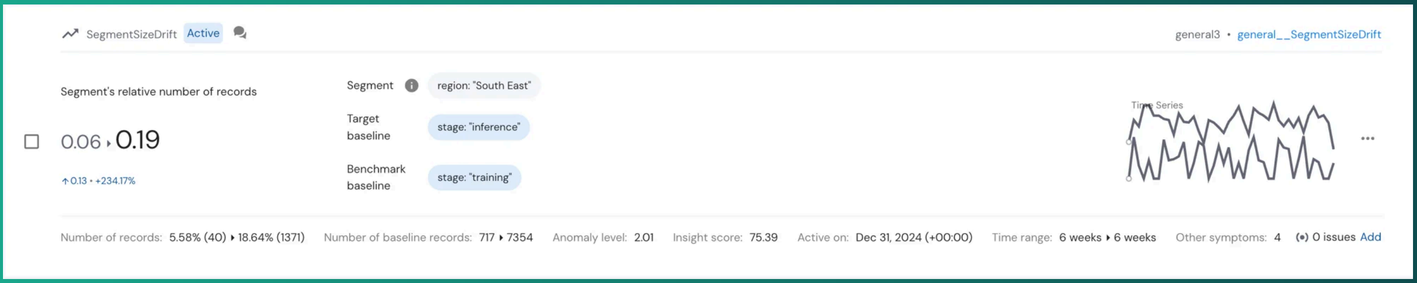
Dependencies

Can the monitoring system understand dependencies?

Can it automatically detect changes in upstream parameters that affect model performance?

Pro Tip: Look for a solution with proactive data quality checks to avoid silent failures in production.

Mona Sample Insight | Detecting Underrepresentation in Training Data Compared to Inference



SegmentSizeDrift Active

general3 • general__SegmentSizeDrift

Segment's relative number of records

□ 0.06 ▸ 0.19

↑0.13 • +234.17%

Segment: region: "South East"

Target baseline: stage: "inference"

Benchmark baseline: stage: "training"

Number of records: 5.58% (40) ▸ 18.64% (1371) Number of baseline records: 717 ▸ 7354 Anomaly level: 2.01 Insight score: 75.39 Active on: Dec 31, 2024 (+00:00) Time range: 6 weeks ▸ 6 weeks Other symptoms: 4 (•) 0 issues Add

Intelligent Alerting and Issue Management

Teams need to catch issues before users do. Alerting systems should go beyond basic thresholds and provide context-aware notifications that distinguish noise from real risk. Integration with tools like Slack or PagerDuty ensures alerts reach the right people quickly. Just as important is having workflows for issue management—being able to log, track, and investigate past alerts so teams can identify patterns, improve root cause analysis, and iterate over time.

Key Considerations

Alerting Intelligence

Can we define custom tests and alerts for performance degradation or data anomalies?

Routing & Workflow

Can the system intelligently route alerts to the right team members who are responsible for troubleshooting?

Does the system support collaboration tools like Slack, Teams, and PagerDuty?

Issue Management

Can the system log, share, archive, and track the handling of insights generated by monitoring tests and alerts?

Signals vs. Noise

Does the system have the logic needed to suppress redundant or superfluous events?

Pro Tip: Choose a system with smart alerting that minimizes noise and avoids alert fatigue.

Visualizations & Investigations

Visualizations help teams diagnose issues faster than raw logs or dashboards. The best monitoring solutions surface insights clearly—highlighting root causes, trends, and segment-specific behaviors—so teams can resolve problems quickly and maintain operational trust.

Key Considerations

Root Cause Analysis

Can the monitoring system generate insights with explanations that improve our ability to diagnose issues?

Can the monitoring system reduce the work it takes tofind the root cause of issues?

Custom Reports & Analytics

Can users create their own dashboards and have the ability to generate and save visualizations and reports?

Investigation Tools

Does the monitoring system support robust investigation tools for deep dive and analytics?

Can it improve Mean Time To Resolution (MTTR) by tying issues to the input data associated with them?

Customizable Dashboards

Does the monitoring UI enable visualizing monito metrics, and reviewing insights and alerts with a degree of personalization?

Pro Tip: Dashboards should be customizable and offer drill-down capabilities to investigate specific incidents quickly.

System and Infrastructure

Deployment and Integrations

Deployment should be fast, flexible, and support your current workflows. The right solution integrates with your ML ecosystem—including model registries, version control, orchestration tools, and cloud providers—and provides robust APIs and SDKs. Whether you prefer SaaS, on-prem, or hybrid, the tool should be easy to install and maintain, with clear versioning and update protocols to reduce friction and risk.

Key Considerations

Installation Options

Do we have the option to deploy on-premises, cloud, or in a hybrid installation?

Stack Compatibility

Can we integrate with common ML and code environments and cloud platforms (AWS, GCP, Azure)?

Patching & Updates

Does the vendor provide adequate frequency of updates, security patches, and versioning support to ensure long-term maintainability?

MLOps Integration

Can we integrate with model registries, version control, and automation tools?

APIs & SDKs

Are the APIs well documented? Does the SDK support customization and automation?

Scale and Performance

As your AI footprint grows, your monitoring platform must handle greater volumes of data and models. It should efficiently process large-scale telemetry without performance tradeoffs and help teams focus on meaningful insights by filtering out noise. The ability to explore historical patterns and detect anomalies across timelines becomes more important at scale, helping reduce operational burden.

Key Considerations

High Performance

Can the monitoring system operate with minimal-to-no performance impact on live inference models?

Scalability

Does the system have the proven scalability to monitor large-scale and high-throughput models?

Pro Tip: Choose a system that balances deep monitoring with minimal overhead to avoid bottlenecks in production.

Security, Compliance, and Governance

In regulated or high-stakes industries, enterprise-grade security and governance are non-negotiable. Monitoring solutions must offer role-based access, audit trails, and data encryption, while also enabling organizations to track explainability, bias, and compliance metrics. Look for solutions aligned with security standards like SOC2 and capable of supporting internal policies and external regulations.

Key Considerations

Data Privacy & Governance

Can the system support encryption, access controls, and compliance with relevant guidelines, e.g., SOC2?

Data Privacy & Governance

Can the system support encryption, access controls, and compliance with relevant guidelines, e.g., SOC2?

Audit Logs & Transparency

Does the system support comprehensive logging of changes and monitoring actions for compliance auditing?

Role-Based Access Control (RBAC)

Does the system support granular permissions to restrict access to sensitive data?

⚡ **Pro Tip:** Security should be a non-negotiable factor—ensure compliance with industry regulations and internal IT policies.

Pricing and Total Cost of Ownership (TCO)

Beyond features, the solution must make financial sense. TCO includes licensing, infrastructure, and the time required to manage and operate the platform. Transparent, usage-based pricing ensures alignment with business value, and operational efficiency can result in significant time and cost savings. An effective monitoring solution helps reduce downtime, avoid costly failures, and improve productivity across teams.

Key Considerations

Pricing Model

Are there pricing options that accommodate different budgets and value points (e.g., pay-as-you-go, enterprise licensing)?

Does the licensing structure provide enough flexibility to support growth and provide cost predictability for budgeting purposes?

Operations

Can we manage the expected level of ongoing maintenance that the solution requires?

Will the vendor-provided support be sufficient to keep us operational?

Infrastructure

Can we project CPU, RAM, and storage requirements over time?

⚡ **Pro Tip:** Pricing structure should be transparent and scalable to accommodate future growth without budget surprises.

Vendor and Relationship Support

Technology is only part of the equation—long-term success depends on the vendor relationship. A strong partner offers responsive support, clear SLAs, and onboarding tailored to both technical and business teams. The vendor should demonstrate a commitment to evolving the platform based on real-world customer needs. Ease of training, ongoing support, and a collaborative roadmap all contribute to successful adoption and long-term impact.

Key Considerations

Service Level Agreements (SLAs)

Are there adequate response times and practices for our DevOps team, the data/AI teams, and the business units?

Onboarding & Training

Will the ease of initial setup be acceptable to our users?

Will training provided by the vendor effectively drive adoption by users?

Product Roadmap

Has the vendor proven a degree of flexibility and ability to accelerate engineering and/or support feature requests?

Do they provide an option for accelerated engineering of specific features?

⚡ **Pro Tip:** Strong vendor support can reduce onboarding time and improve long-term value. Prioritize solutions with dedicated customer success teams.

Keeping AI models performing as expected isn't easy. Models drift, data shifts and unexpected issues pop up—sometimes with serious consequences. But most companies are still trying to tackle AI monitoring on their own, relying on homegrown solutions that weren't built for the complexity of real-world AI.

If you're feeling the pain of unreliable model performance, hidden failures, or compliance headaches, you're not alone. We've worked with teams facing the same challenges, and we know what it takes to build a practical, scalable monitoring strategy. Whether you're looking for best practices, guidance on governance, or just a way to catch issues before they turn into disasters, we're here to help.

To learn about how Mona approaches AI/ML monitoring,
watch our 5-minute intro video below, or take a self-guided tour of the platform.

Watch a Demo

Self-Guided Tour