MONTE CARLO

EVALUATION GUIDE

Excellence with Data + AI Observability

A Comprehensive Guide to Data + AI Observability Solutions and RFP Template

Table of Contents

I. Introduction

Data + AI observability has been one of the hottest emerging data engineering technologies of the last several years. And with data quality and reliability becoming a central topic in the data product and AI conversation, that trajectory isn't changing any time soon.

Benefits of data + AI observability include:

- Increasing data + AI trust and adoption;
- Mitigating the risks of bad data + AI;
- Boosting data product and AI agent ROI;
- And reducing time and resource costs associated with data + AI reliability

And as the category grows, so too do the alternatives. Following Monte Carlo's creation of the data + AI observability category in 2019, a myriad of alternative data + AI observability tools have entered the market at various levels of maturity.

So, with so much choice, how does a data + AI leader decide? In this post, we'll share the must-have features and benefits based on both analyst perspective and our own experience building the category, as well as a sample RFP template you can use to support your own evaluation process.

Let's get started!

II. The Key Components of Data + AI Observability

What's data + AI observability again?

Data + AI Observability is a comprehensive approach to observing the health and reliability of your data, system, code and AI models end-to-end. It closes the loop between data inputs and AI agent outputs—ensuring trust, scale, and business impact.

The core features of data + AI observability tools were originally defined across five pillars that included four types of data anomalies (freshness, volume, schema, quality) and data lineage.

Since then the industry and category have exploded. Now Gartner and others define these platforms mainly by the workflows they accelerate:

- **Detect**- Proactively alerted to issues with data and AI outputs.
- **Triage**- Get the right alert to the right team with the right urgency level.
- **Resolve**- Uncover the root cause of an issue so it can be resolved quickly.
- **Measure**- Report and surface quality and operational response metrics.

A particular emphasis is placed not just on detecting issues, but providing end-to-end visibility across the data + AI lifecycle to enable fast troubleshooting and issue resolution.

But more on that later. For now, let's take a look at each of the critical dimensions impact the efficacy, usability, and longevity of a data + AI observability tool.

Data + AI observability evaluation criteria

Evaluation criteria should be focused on features that drive business value, rather than a loosely compiled list of features submitted by users using their current legacy tools as reference.

Most evaluation criteria should be driving a proven benefit realized by most organizations that have adopted a data + AI observability platform such as:

- Reclaimed time for data personnel
- Avoided losses from data + AI downtime
- Improved AI/ML models
- Improved collaboration from data trust
- Reduced storage + compute cost

Here are the key evaluation criteria categories we feel are the most critical to driving these benefits:

- Enterprise readiness
- End-to-end coverage
- Seamless incident management
- Integrated data lineage
- Comprehensive root cause analysis
- Quick time-to-value
- Agent observability

So, now that we've got those points in mind, let's dive into each of these criteria categories in a bit more detail.

Enterprise readiness

The world of data + AI is always evolving. That's why you need an experienced strategic advisor. Any vendors can promise the world, but what can they actually deliver?

Will that team of 12 people in a garage still be around in a year to deliver on key feature requests? These are important questions to answer through customer reference calls to understand a solution's overall maturity.

Some key areas to evaluate for enterprise readiness include:

- **Security–** SOC II certification? Robust role based access controls?
- **Architecture–** Do they have multiple deployment options for the level of control over the connection? How does it impact data warehouse/lakehouse performance?
- **Usability–** Is an alert just pretty or will it actually save you time—like bulk update incidents or being able to deploy monitors-as-code.
- **Scalability–** What are their largest deployments? Has this organization proven its ability to grow alongside its customers? Other key features here include supporting domains, reporting, change logging, and more.
- **Support–** Data + AI observability isn't just a technology, it's an operational process. The maturity of the vendor's customer success organization can impact your own level of success.
- **Innovation history and roadmap–** The data world changes rapidly and as we enter the AI era, you need a partner that has a history of first-to-market innovation. Fast followers are often anything but, with comparative features shipped 6 months to a year later. (That's 25 in chief data officer years.) Cloud-native solutions often have an advantage here.

End-to-end coverage

The true power of data + AI observability lies in its ability to integrate across modern data platform layers to create end-to-end visibility into your critical pipelines.

For years, data testing–whether it was hardcoded, dbt tests, or some other type of unit test–was the primary mechanism to catch bad data.

While still relevant in the right context, the problem with data testing as a complete practice is that you couldn't possibly write a test for every single way your data could break. No matter how well you know your pipelines, unknown unknowns will still be a fact of life. And even if you could identify every potential break (which you can't), you certainly wouldn't be able to scale your testing to account for each one as your environment grew.

Data + AI observability tools should offer both broad anomaly detection across all the tables once they have been added to your selected schemas, as well as deep monitoring for issues inherent in the data itself. **This also includes the outputs from your AI agents.**

A strong observability tool will also integrate widely across your data + AI platform, from ingestion to consumption, and enable quick time-to-value through simple plug and play integrations.

Be sure to verify that your chosen solution offers tooling integrations for each of the layers you'll need to monitor in order to validate the quality of your data products, as well as integrations into existing workflows with tools like Slack, Microsoft Teams, Jira, and GitHub.

Incident management

Most data + AI teams we talk to initially have a detection focused mind-set as it relates to reliability, likely formed from their experience with unit testing.

The beauty of data + AI observability is that not only can you catch more meaningful incidents, but the best solutions will also include features that improve and accelerate your ability to manage incidents.

Reliability issues are inevitable and having tools to mitigate the impact of those issues provides tremendous value. A few areas to consider include:

- **Impact analysis —** How do you know if an incident is critical and requires prioritizing? Easy—you look at the impact. Data + AI observability tools that provide automated column-level lineage out-of-the-box will also often provide an impact radius dashboard to illustrate how far a quality issue has extended from its root. This can help engineers understand at a glance how many teams or products have been impacted by a particular issue and who needs to be kept informed to updates in status.

- **Internal team collaboration —** Once an alert has triggered there needs to be a process for assigning and potentially transferring ownership surrounding the incident. This may involve integrating with external ticket management solutions like JIRA or ServiceNow, or some teams may choose to manage the incident lifecycle within the data + AI observability tool itself. Either way, it's helpful to have the flexibility to do both.

- **Proactive communication with data consumers —** When consumers use bad data to make decisions, you lose trust. Data + AI observability solutions should have means for proactively communicating with data consumers the current health of particular datasets or data products.

Integrated data lineage

Lineage is a dependency map that allows you to visualize the flow of data through your pipelines and simplify root cause analysis and remediation.

While a variety of tools will provide lineage mapping at the table level, fewer extend that lineage across your key systems where data is produced, processed, and consumed.

It's also important that your data lineage and data incident detection features work as an integrated solution within the same platform. A key reason for this is that lineage grouped alerting not only reduces alert fatigue, but helps tell a more cohesive story when an event impacts multiple tables.

Rather than getting 12 jumbled chapters that may be part of one or two stories, you are getting an alert with the full story and table of contents.

Automated root-cause analysis

What is your standard root cause analysis process? Does it feel disjointed hopping across multiple tools? How long does it take to resolve an issue?

Data can go bad in a lot of ways. AI even more so. A comprehensive data + AI observability tool should help you identify if the root cause is an issue with the data, system, code, or model.

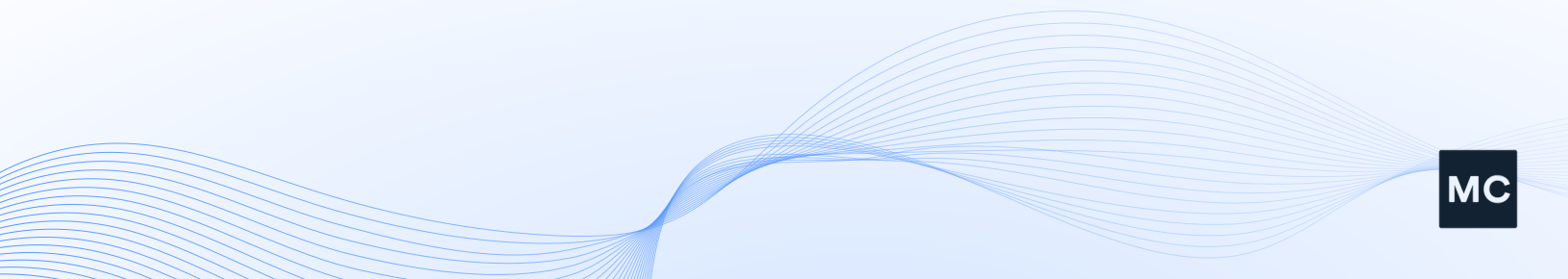
For example, the data can be bad from the source. If an application went buggy and you started seeing an abnormally low sales price from orders in New York, that would be considered a data issue.

Alternatively, a data environment is made up of a panoply of irreducibly complex systems that all need to work in tandem to deliver valuable data products for your downstream consumers. Sometimes the issue is hidden within this web of dependencies. If you had an Airflow job that caused your data to fail, the real culprit wouldn't be the data but a system issue.

Or if a bad dbt model or data warehouse query change ultimately broke the data product downstream, that would be considered a code issue.

AI agents have their own data, systems, and code dependencies. The context provided by the data could be inaccurate or missing, DAGs can fail, and small changes to prompts, agent orchestration, or model version can make a big difference on the final output.

Advanced tools will even have agentic troubleshooting workflows to automatically and accurately identify the root cause of an issue in minutes so your team can reduce the time to resolution.



Quick Time To Value

Data + AI observability is intended to reduce work—not add more.

If a data + AI observability tool is providing the right integrations and automated monitors for your environment out-of-the-box, it will be quick to implement and deliver near immediate time-to-value for data teams.

A data + AI observability solution that requires more than an hour to get set up and more than a couple of days to start delivering value is unlikely to deliver the data quality efficiencies that a growing data organization would require to scale data quality long-term.

Agent observability

Building valuable generative AI or agent applications requires reliable first party data. Many agents are designed to query the data warehouse or lakehouse to get the context they need--if the data isn't reliable and accurate they will appear to hallucinate.

But even with perfectly accurate context, agents can still produce outputs unfit for use. It could be a missing citation, wrong tone of voice, or a simple but devastating hallucination.

Data + AI observability tools need to be able to monitor AI outputs using both LLM-as-judge and deterministic monitors along with features allow monitoring to scale in production such as data sampling, trace debugging, span filtering and more.

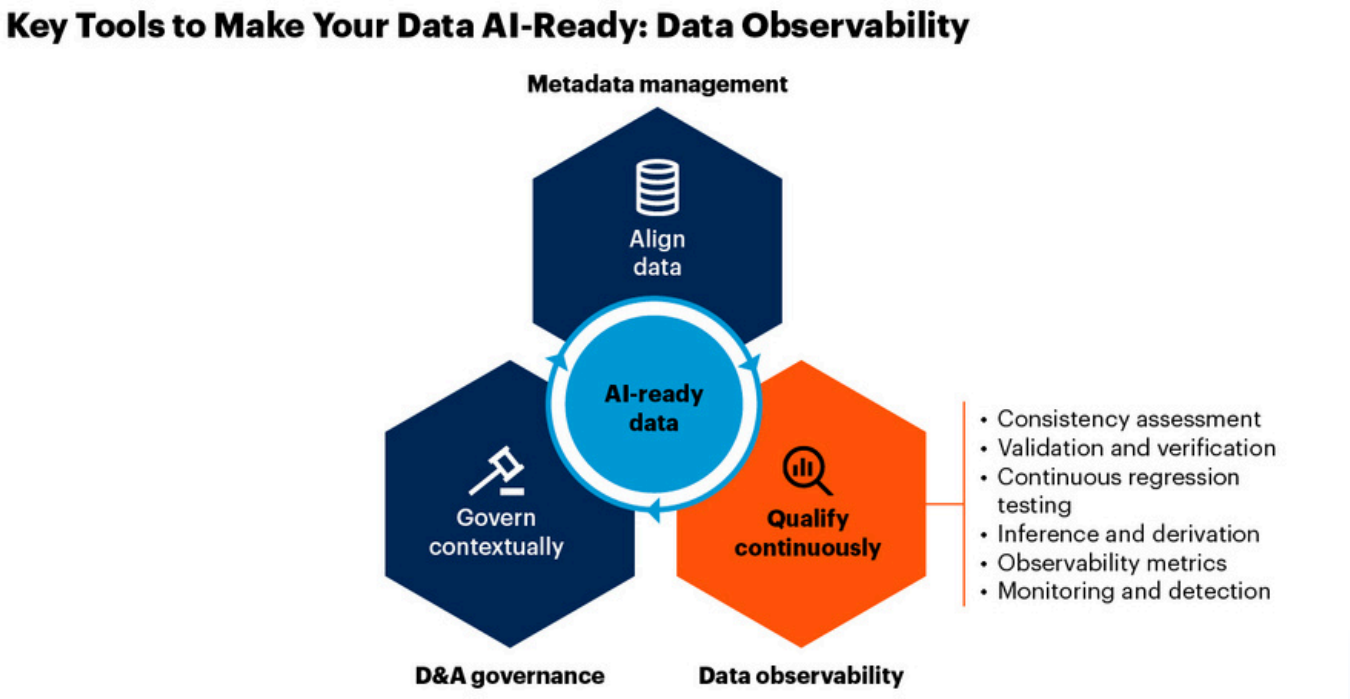
III. Analyst Perspective

Gartner

While Gartner hasn't produced a data + AI observability magic quadrant or report ranking data + AI observability vendors, they have placed it on the 2025 Data Management Hype Cycle with a "HIGH" benefit rating.

They say data and analytics leaders should, "Explore the data + AI observability tools available in the market by investigating their features, upfront setup, deployment models and possible constraints. Also consider how it fits to overall data ecosystems and how it interoperates with the existing tools."

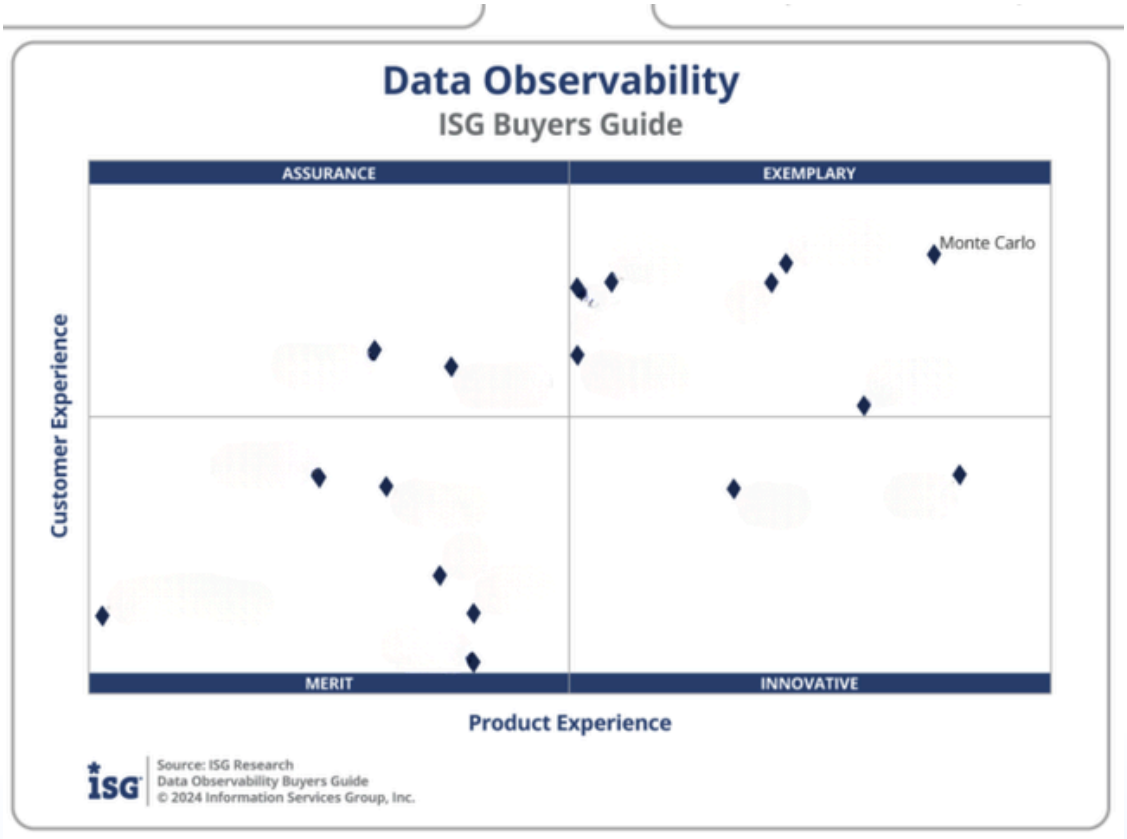
They have also referenced it as one of three core technologies for achieving "AI-Ready data."



ISG Buyers Guide

The ISG Buyers Guide really hit the nail on the head saying that the best data + AI observability solutions go beyond detection to focus on resolution, prevention and other workflows:

“The research largely focuses on how vendors apply data + AI observability and the specific processes where some specialize, such as the detection of data reliability issues, compared to resolution and prevention. **Vendors that have more breadth and depth and support the entire set of needs fared better than others.** Vendors who specialize in the detection of data reliability issues did not perform as well as the others.”



G2 Crowd

G2 was one of the earliest non-vendor resources to put together a credible list of data + AI observability vendors and a definition for the category. They say:

To qualify for inclusion in the G2 Crowd data + AI observability category, a product must:

- Proactively monitor, alert, track, log, compare, and analyze data for any errors or issues across the entire data stack
- Monitor data at rest and data in motion, and does not require data extraction from current storage location
- Connect to an existing stack without any need to write code or modify data pipelines

Vendors are evaluated by verified users of the product across a list of organizational and product specific capabilities including:

- Quality of support
- Ease of admin
- Ease of use
- Integrations
- Alerting
- Monitoring
- Product direction
- Automation
- Single pane view

IV. RFP Template

Data + AI Observability Request For Proposals

Section	Key Capabilities	Criteria
Company Details	Vendor experience	What is your experience in the industry?What references and case studies can you provide from similar projects?
	Innovation and roadmap	What is your vision for the future of data + AI observability? What are your planned enhancements and new features?
Security	Deployment options	Is your product delivered as a Software-as-a-service (SaaS) offering? Can you deploy the agent in our AWS/GCP/Azure instance if desired?
	System Integration and data handling	Can you provide proof of SOC2 Type II certification? What options do you provide for authentication?Do you have role based access controls?Do you provide an API for retrieving security events to import into a SIEM?Are third-party penetration tests available?What data is exported from our environment? Is it encrypted?
Configuration and Management	APIs	What functionality is available via API? Is the API REST?
	CLI & SDK	Is a command line interface available to simplify API interactions? Are any SDKs available, e.g. for use in Python scripts or Data Science Notebooks?
	Airflow Operator	Can monitoring and alerting be configured from within your Airflow jobs, without breaking a workflow?
	Monitors and notifications as code	Can monitoring and alert routing be defined in a file placed under version control? Can multiple monitors be created at once using abstractions, or must monitoring be instantiated for each table and field one by one?
	Performance	How will your product impact our data warehouse/lake/lakehouse performance and compute costs?

RFP Continued

Section	Key Capabilities	Criteria
Integrations	Warehouse /Lake /Lakehouse	What solutions do you currently support? What are your planned enhancements and new features?
	Other databases	See above.
	BI tools	See above.
	Integration, transformatio, orchestration	See above.
	Collaboration tools	See above.
	Data catalogs	See above.
	Code repositories	See above.
Support & Use	Product support	Do you provide web-based self-support resources? Do you offer 24/7 support (24 hour SLA) What kind of support is provided during major releases? Are your releases backwards compatible? Do you charge additional fees for providing product support?
	Usability and onboarding	What training, onboarding, and ongoing support is available? How are deployment and product best practices shared amongst users? How is the time required to execute tasks minimized by the UI and key workflows? Is administration and management of the platform and its capabilities low-code or no-code?
Pricing Structure	Licensing model	What is the basis of licenses for the product? Is on-demand/usage based pricing available?

RFP Continued

Section	Key Capabilities	Criteria
Monitoring	AI-Powered Monitors & Agents	What kinds of intelligent features allow for finding anomalies without the need for manual input? Does the platform leverage AI to define and recommend relevant monitoring? To what extent is anomaly detection manually configurable? Can anomaly detection be automated to cover tables upon creation based on schema, database, domain, tag, and name?
	Data quality monitoring	Does the product automatically detect freshness, schema, and row count anomalies? Does the product provide monitoring for percents and/or counts of null or missing values? Uniqueness, duplicates? Min, max, average, stddev, variance, skew, percentiles? Does the tool offer dimension tracking capability? Does the platform offer pre-built data validations where thresholds be defined and deployed without code?
	Data consistency	Does the product support custom business logic that involves more than one table? More than one database?
	Custom monitoring	In addition to automated tests and checks, what ability does the user have to create custom checks and tests? Can these tests be scheduled and/or ran manually? Are sample SQL templates provided within the tool for defining a custom monitor? Can custom SQL rules be generated with AI ?
	Preventative actions	Does the tool have circuit breakers which can stop pipelines when data does not meet a set of quality thresholds? Can the tool show the impact of a pull request or dbt model change on downstream assets such as specific dashboards?
	Data profiling	How does the service profile data shape statistically? (what stats are provided/considered for various data types)
Query usage and performance monitoring	Cost and runtime optimization	How does the service monitor query usage and efficiency for the objects under observation? What sort of feedback or recommendations does the tool provide? How is the warehouse monitored for both usage and storage efficiency?

RFP Continued

Section	Key Capabilities	Criteria
Incident Management	AI Workflows	Are there AI-powered workflows or MCP Server for identifying and remediating common alert /monitoring trends and gaps?
	Alerting and ownership	What alert channels are supported? What guidance or initial information do alerts provide? How is ownership of incidents and data assets tracked ?
	Lineage and impact analysis	Can the product infer the priority of detected issues by factors such as asset usage? Can users understand the point of origin upstream and impact downstream using data lineage? Can users understand cross system dependencies with syncs, dashboards, topics, transformation models, DAGs, and more all displayed on one linage graph?
	Asset metadata /self-service data discovery	Is key metadata from across the modem data stack, including object tags, surfaced in one pane of glass for each table or data asset? Does it include asset type, schema, table ID, users, usage, logs, and other relevant documentation?
	Consumer notification channels	What service/ticket management/workflow integrations exist? Will they enhance collaboration across the data + AI team? How can data consumers be proactively alerted to incidents or the current health status of a dataset?
	Incident tracking	Does the product document incidents, alert feedback, users involved, and resolution notes as reference for future issues?
	SLA support	Can SLAs or other alert grouping methods be defined? Can alerts be directed to different teams? How flexible is the grouping system?

RFP Continued

Section	Key Capabilities	Criteria
Resolution/ Root Cause Analysis	Insights & AI Troubleshooting	Does the tool detect insights to facilitate in the discovery of the root cause of a particular data incident? Does the tool use AI to accelerate root cause analysis and troubleshooting?
	Data-level RCA	Can the product automatically surface row-level previews of data that likely contributed to a detected anomaly for use in debugging? Does the tool provide segmentation analysis capabilities to further hone in on specific segments where the data is anomalous?
	System-level RCA	Does the product integrate broadly across our data systems to surface alerts behind a single pane of glass?
	Code-level RCA	Does the product correlate changes to query code (no code transformation tools) to specific incidents? Does the tool monitor query performance and execution?
Reporting	Data health trends reporting	Does the product offer pre-built dashboards showing data health metrics over time at the organizational, domain, and data product level?
	Data health status	Does the product offer pre-built dashboards showing the current health of tables or data products?
	Data operational response reporting	Does the product offer pre-built dashboards showing the operational metrics (time to response, time to fixed) of different teams?

RFP Continued

Section	Key Capabilities	Criteria
Agent Observability	Unified Platform	Can the platform monitor for quality/reliability issues across both data and AI outputs? Within the same interface? Does the solution support role-based access control so that only authorized users can view sensitive traces or telemetry? Is there support for integration with existing communication tools (e.g., Slack, Teams, email) and incident management tools (e.g., PagerDuty, ServiceNow, Jira) for alert routing?
	Tracing	Can telemetry remain within the our own data infrastructure (warehouse, lakehouse, or lake) to maintain compliance, governance, and audibility? Is telemetry for agent operations (e.g. prompts, context, completions, tokens) captured & structured based on industry standards? What models and frameworks are supported?
	Evaluations & Monitoring	Can monitors be deployed at the trace (run) and span levels? Can anomalies or regressions in agent behavior be detected (via anomaly detection) without manual threshold tuning? Can anomalies or regressions in agent operational metrics (cost, latency, etc) be monitored? Can they be detected (via anomaly detection) without manual threshold tuning? Does the platform provide pre-built LLM-as-Judge evaluation monitors to assess agent output quality (e.g. relevance, prompt adherence, clarity, task completion)? Can they be customized? Does the platform provide pre-built SQL evaluation monitors (or “checks”) to assess agent output quality (e.g. output length, format, banned words, etc.)? Can they be easily customized? Can users define custom criteria and evaluation monitors? Can users evaluate the effectiveness of their monitors (custom or otherwise) by testing them on historical traces? Can the system support sampling strategies to scale monitoring across large volumes of agent runs? Can criteria be selected to filter spans for monitoring or set to monitor specific segments? Do evaluations provide a change log and run history?

V. The Future of Data + AI Observability

What's next for data + AI observability

There's one critical feature that we didn't mention earlier, that plays a huge role in the long-term viability of a data + AI observability solution.

And that's category leadership.

Like any piece of enterprise software, you aren't just making a decision for the here and now—you're making a bet on the future as well. When you choose a data + AI observability solution, you're making a statement about the vision of that company and how closely it aligns to your own long-term goals. "Will this partner make the right decisions to continue to provide my organization with adequate data quality coverage in the future?"

Particularly as AI proliferates, having a solution that will innovate when and how you need it is equally as important as what that platform offers today.

Monte Carlo been named a proven category leader by the likes of G2, Gartner, ISG, and the industry at large; and has also become the de facto leader for AI reliability as well.

There's no question that AI is a data product. And with a mission to power data quality for your most critical data products, Monte Carlo is committed to helping you deliver the most reliable and valuable AI products for your stakeholders.

Just starting your data + AI observability journey?

Contact our team today to find out how Monte Carlo can help your team save time, reduce costs, and maximize your data resources with our category-creating data + AI observability solution.

Check out more helpful resources on data and AI trends and best practices, including:

- Data Downtime Blog: Get fresh tips, how-tos, and expert advice on all things data.
- Data + AI observability Product Tour: Check out this self-guided screenshot tour showing just how a data + AI observability platform works.
- Data + AI Observability Value Calculator: Enter in a few specifics about your data environment and see the benfits provided by data + AI Observability. Produced by Forrester.