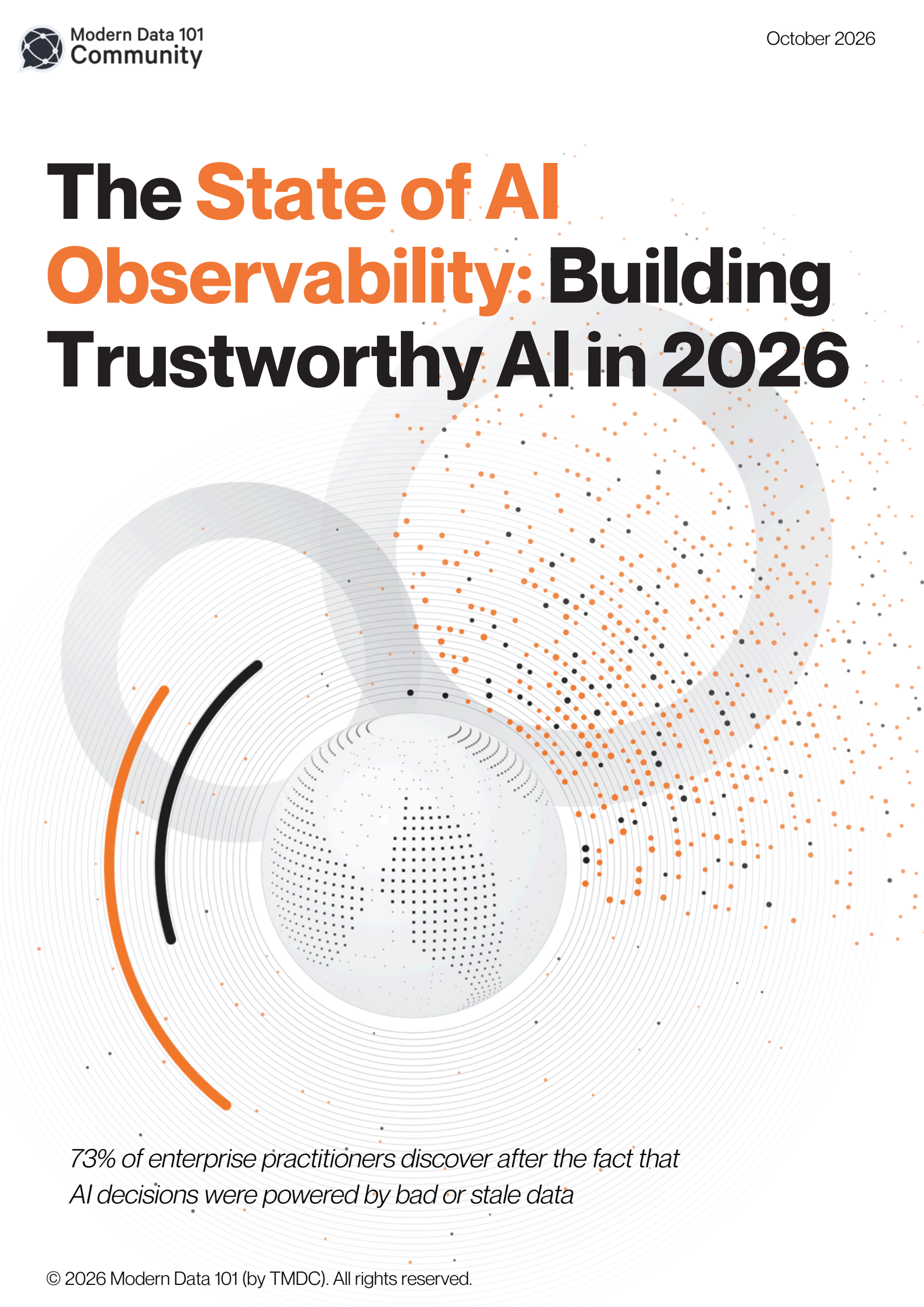


The State of AI Observability: Building Trustworthy AI in 2026



*73% of enterprise practitioners discover after the fact that
AI decisions were powered by bad or stale data*

Table Of Contents

1. Executive Summary
2. What is AI Observability
3. When Healthy Infrastructure Hides Broken AI
4. LLM Monitoring and Observability: The Signals That Matter
5. Data Observability: The Layer Beneath the Model
6. Agent Observability and the Vanishing Checkpoint
7. Trust Latency: What Makes AI Safe in Practice
8. The AI Observability Maturity Curve
9. Sequencing the Build
10. Conclusion
11. References

Executive Summary

From The Modern Data 101 Community

This edition compiles insights from industry leaders, offering direction to practitioners for building safe and reliable AI in 2026 and 2027.

Enterprise AI has a visibility problem that presents itself as a quality problem. Models keep answering, dashboards stay green, and the first sign that something went wrong arrives from a customer, an auditor, or a regulator. This brief looks at how enterprises are closing that gap in 2026, what AI observability has to measure to earn its budget, and why seeing a failure and stopping one have turned out to be separate capabilities that most organisations have funded unevenly.

The adoption curve is steep. LLM observability investment is forecasted to reach 50% of GenAI deployments by 2028, up from 15% today, and 40% of organisations deploying AI to run dedicated AI observability tools by the same year. Observability is moving from optional tooling to expected infrastructure faster than most budgets have adjusted.

The control gap is wider. In a survey of 101 executives in the US and Canada, 59% said they can detect AI failures, while only 23% said they can routinely stop AI from taking the wrong action. Our central argument follows from that gap. Safe AI is an operating capability that lives in the system around the model, and its best single measure is trust latency, the time between an AI system going wrong and the organisation stopping it.

What is AI Observability

AI observability broadly refers to the ability to understand what an AI system is doing, why it is doing it, and whether the result can be trusted, using evidence the system produces while it runs. The practice rests on one discipline: record every meaningful AI action in a form that can be queried later. PwC groups that evidence into four categories: system behaviour, system health, user access and data flow. Gartner adds that this needs dedicated tooling, because drift, bias and flawed LLM logic never appear in infrastructure telemetry.

The term blurs easily, so it helps to separate three questions that sound alike and need different instrumentation.

	Monitoring	Observability	Control (safe AI)
Core question	Is it running?	Is it right, and why?	Can we stop it, and how fast?
Evidence	Latency, uptime, error rate	Traces, groundedness, drift, lineage	Policies, guardrails, approvals, pause authority
Usual owner	Infrastructure and SRE teams	ML platform and data teams	Risk, security and product owners
What it misses	Wrong answers returned with a success status	Failures nobody is authorized to act on	Failures nobody can see

Monitoring, observability and control answer different questions, belong to different teams, and each misses what the others catch | Source: Authors

Why Most AI Tooling Budgets Waste Money?

Most tooling already in place answers the first question. Vendor language blurs all three under one label, so a latency dashboard and an output evaluation engine can both arrive in a proposal as AI observability. Procurement teams that write the question they need answered into the requirement get a cleaner comparison than teams that write the category name.

When Healthy Infrastructure Hides Broken AI

Serious AI failures in production tend to share one trait: the infrastructure looked healthy while they happened. Air Canada's chatbot told a customer he could claim a bereavement fare after travel under a policy that did not exist, and a tribunal ordered the airline to pay damages. Nothing in uptime monitoring would have flagged that answer, because the system responded quickly and without error.

The distance between running and working is wide in the data. Infosys research indicates that only 19% of AI use cases meet all of their business objectives, and a Gartner analyst describes the cause as a visibility gap that makes scaling risky, since most organisations are still working out how to monitor and trust these systems.

56% Intervened, 12% Are Blind: The Reality of Production AI

HFS Research found that 56% of executives had to intervene to correct or stop an AI-driven outcome in the past 12 months, and 12% admitted they would not necessarily know if an AI failure had occurred. The second figure matters more than the first. An organisation that cannot tell whether a failure happened has no baseline to improve from.

Garbage In, Chaos Out: The Root Causes of AI Misbehavior

The most-cited causes were bad or stale data at 56% and broken integrations at 39%, level with unpredictable model behaviour at 39%. Our own survey points the same way. 73% of practitioners report discovering after the fact, sometimes, usually or always, that the data behind a decision was wrong or outdated, and 82% say their analytics tools are mostly not well connected to governance systems.

Infosys research indicates that only 19% of AI use cases meet all of their business objectives

Why Unmonitored AI Use Cases Get Canceled

Read together, these figures describe two blind spots. One sits above the model, where outputs are never graded for accuracy. The other sits below it, where data and integrations change without notifying the systems that depend on them. Observability budgets that cover only the model leave both open.

The commercial cost rarely appears as an incident. Gartner in an instance warned that without explainability and observability foundations, GenAI stays confined to low-risk internal tasks, which caps return on investment. UST makes the same point from the delivery side and places the stall point between prototype and production, where data readiness, governance and ownership are still undefined. In practice, the cost of invisibility is the set of use cases leadership declines to automate because nobody can evidence control.

4

LLM Monitoring and Observability: The Signals That Matter

LLM monitoring and observability work best when designed around the failure modes unique to language models, because standard infrastructure signals say little about them. Three terms get used interchangeably and deserve separation. Monitoring tracks counters such as response time, error rate, token use and cost. Tracing records the causal chain of a single request through retrieval, tools and generation. Observability combines both with evaluation, the step that converts telemetry into a verdict. Without evaluation, logs record that something happened and nothing records whether it was acceptable.

What Gartner says to measure

Gartner observes that traditional observability concentrates on speed and cost, and that priority is shifting toward factual accuracy, logical correctness and sycophancy. Its guidance is to adopt multidimensional LLM observability that tracks latency, drift, token usage and cost, error rates and output quality together, and to run LLM evaluation inside CI/CD pipelines so factual-accuracy and safety checks happen before deployment.

AI Reliability indicators worth a threshold

NASSCOM's guidance on AI reliability names indicators that make the abstract measurable: hallucination rate, groundedness against retrieved sources, response consistency across rephrased queries, latency under load, and the frequency and severity of drift. A number without a threshold and an escalation path is a statistic. A number with both is a control.

Signal layer	Examples	What it tells you	What it cannot tell you
Infrastructure	Latency, throughput, error rate, GPU load	The plumbing works	Whether any answer was right
Model telemetry	Token counts, prompt and response pairs, model and prompt versions, retrieved sources	What was asked, what it cost, what changed	Whether the answer was grounded
Output quality	Groundedness, hallucination rate, relevance, semantic drift, toxicity, user feedback	Whether the answer can be relied on	Why it went wrong upstream

No single signal layer is enough. Infrastructure shows the system is up, telemetry shows what changed, and output quality shows whether answers can be trusted | Source: Authors

Audit-Proofing Your AI: The Logs Regulators Will Demand

Regulation adds a second reader to every log. For high-risk systems, the EU AI Act calls for automatically generated logs and human oversight, and a defensible LLM record names who asked, when, which model answered, what was submitted with sensitive data protected, what came back, which policies applied and which violations fired. Records that capture only metadata cannot answer the question an examiner asks, which is how a specific output was produced.

5

Data Observability: The Layer Beneath the Model

The most expensive AI failures often begin before the model sees a prompt. A retrieval index last refreshed six weeks ago, a schema change that quietly broke an identifier, or two teams computing the same metric differently will each produce fluent answers built on a false premise, and model-level evaluation will blame the model. The evidence supports starting lower in the stack. In our survey, 53% of practitioners said they have no easy way to trace data lineage, and 69% had redone work because of data issues within the last quarter.

Observability that starts at the model produces a diagnosis with no cause.
Observability that starts at the data produces a cause with no symptom.

The useful system joins them through lineage keyed to each individual inference, so a bad answer can be walked back to the dataset version, retrieval index and prompt version that produced it. [Modern Data 101's earlier research](#) describes one reason this matters: training-serving skew, where a model learned from one version of a feature and serves another, degrades decisions silently while every data quality check stays green.

Treat the data product as the unit of observability

A governed data product can expose its own health: a freshness commitment, a named owner, a quality score attached at the source, and change notices when schema or lineage shifts. When those signals exist at the source, AI observability can query them instead of reconstructing them after an incident.

Why this belongs in the observability budget

Gartner warns that without standardised model telemetry, incident resolution for AI applications drags on. We would add that telemetry only shortens resolution when it can be joined to data state. Otherwise, root-cause analysis stops at the model and often reaches the wrong conclusion.

Four signals to inherit from the data layer

- Freshness: when the data was last updated, and whether that lag exceeds what the AI system can tolerate.
- Lineage: where the data came from and what transformed it, so a degraded output traces to a specific event.
- Schema stability: undocumented structural changes are among the most common silent causes of AI failure.
- Quality: completeness, accuracy and uniqueness, scored at the source so every consumer inherits the same trust signal.

53% of practitioners said they have no easy way to trace data lineage, and 69% had redone work because of data issues within the last quarter.

Agent Observability and the Vanishing Checkpoint

Agents turn observability from a reporting function into a safety function, because the review step that used to sit between a model's output and its consequence has disappeared. A single request can pass through several model calls, a retrieval step, multiple tool invocations, and a sub-agent before anything reaches a user. The unit worth observing is therefore the decision path, and the final answer is only its last line. A failure at the third step can surface as nonsense at the eighth, with every step in between logging success.

Scale arrives before inventory

Gartner predicts that by 2028 an average global Fortune 500 enterprise will have over 150,000 agents in use, up from fewer than 15 in 2025, and reports that only 13% of organisations believe they have the right AI agent governance in place. Observability starts with discovery. An agent that has not been inventoried cannot be traced, scoped or retired, which makes sprawl an observability problem before it becomes a governance problem.

Intent has outrun control

Deloitte's survey of 3,235 leaders found 74% expect to use agents at least moderately within two years, while only 21% report a mature governance model for them. McKinsey's 2026 trust survey reports that roughly one-third of organisations reach maturity level three or higher on agentic governance and controls, and that nearly two-thirds cite security and risk as the top barrier to scaling agents. Different samples, same shape: appetite is high, and evidence of control is thin.

Vendors are responding in the product layer. Forrester predicts that half of enterprise ERP vendors will launch autonomous governance modules in 2026, combining explainable AI, automated audit trails and real-time compliance monitoring. The direction of travel is observability moving from a separate tool into the application itself, which will raise the baseline buyers can demand from suppliers.

Where this connects to the earlier volumes

One of Modern data 101's report named agent debt, the gap between the autonomous decisions a system may take and the decisions anyone can trace, own or reverse. While another report described the difference between a policy that states a rule and a runtime control that enforces it. Observability sits between them. It is the instrument that measures agent debt, and it supplies the evidence a runtime control needs before it can act. An organisation that cannot see an agent's decision path cannot service the debt it is running.

7

Trust Latency: What Makes AI Safe in Practice

Safe AI is easiest to reason about as a race between how fast harm can spread and how fast an organisation can respond. We use the term trust latency for the organisation's side of that race: the time between an AI system beginning to behave badly and the organisation stopping it. It splits into three clocks. Detect measures how long before anyone knows. Explain measures how long before the cause is understood well enough to act on. Contain measures how long before the behaviour is stopped or reversed.

Clock	What shortens it	What lengthens it	Question to ask
Detect	Continuous evaluation on live traffic, drift and anomaly alerts, thresholds tied to named owners	Sparse sampling, infrastructure-only signals, alerts with no routing	If output quality degraded tonight, which alert fires and who receives it?
Explain	End-to-end traces, lineage keyed to each inference, versioned prompts and models	Logs without context, separate tools for each team	Can we reconstruct this answer from model version, prompt version, retrieved sources and data freshness?
Contain	Tiered guardrails, rollback paths, scoped agent identity, policy enforcement at runtime	Alerts that need approval before anyone can act, no owner with pause authority	Who can switch this off, and how long does it take?

Every AI failure runs on three clocks. How fast you detect, explain and contain it depends on choices made before launch | Source: Authors

Enterprises have invested unevenly across the three clocks. HFS Research found that 59% of executives can quickly detect when AI is not performing as expected, while 23% can routinely stop it from taking the wrong action, a 36-point drop between seeing and stopping. The asymmetry is predictable. Dashboards are visible, purchasable and easy to demo. Containment is organisational as much as technical, since it needs a named owner with the authority to pause a system at 2 a.m. without convening a committee.

Why trust latency belongs on a board dashboard

Trust latency turns observability spend into a number an executive can interpret without reading a trace. It also scales with autonomy. A summariser that drafts internal notes can tolerate a latency measured in days. An agent that moves money or updates customer records needs one measured in minutes, because its errors reach other systems before a human could review them.

A safe AI system is one whose trust latency stays shorter than the time its errors need to reach a customer, a regulator or another system. Oracle's engineering team describes the same design move as governed execution, in which every consequential action is policy-checked, identity-scoped, observable and auditable. Keyrus frames the management version: model risk management becomes a continuous assurance function instead of a periodic checkpoint.

59% of executives can quickly detect when AI is not performing as expected, while 23% can routinely stop it from taking the wrong action

The AI Observability Maturity Curve

Maturity in observability is best read as how far evidence travels before someone acts on it. Our four-stage model starts at isolated monitoring, where data pipelines and models are watched separately with no shared context. Connected alerting follows, where teams share signals but not lineage. Lineage-aware observability links model behaviour to data causes and makes automated root-cause tracing possible. Active governance sits at the top, where observability feeds the AI control plane and triggers rollback, containment and audit logging without waiting for a person to notice.

Most organisations sit between the second and third stage, and the third is where decision trust becomes achievable. McKinsey's separate maturity survey points to a similar picture, with about one-third of organisations at level three or higher on agentic governance and controls. The ladder below shows what each stage looks like in practice, the proof that you have reached it, and what tends to break next.

How to test a stage claim

Put three questions to every claim: does the signal exist, does it reach a named owner, and does it trigger an action? A team that passes only the first belongs one stage lower than it believes. The usual error runs in one direction.

Infrastructure instrumentation is strong, and coverage of what models say and what agents do is thin, so organisations place themselves a stage above what their output-quality and agent coverage supports. Culture moves the curve as much as tooling does. Data engineers, ML engineers, analysts, risk teams and business owners need shared ownership of the same signals, or the third stage never forms.

Building Reliable AI

Order matters more than completeness here, because each phase produces the inputs the next one needs. A team that buys evaluation tooling before it knows which systems carry the most risk instruments the wrong things thoroughly. The plan below follows the evidence in this brief and keeps every phase tied to a test a sceptical executive could run.

Phase	Focus	What to do	Exit test
Phase 1	Know what you run	Inventory every model, agent and assistant in use, including shadow tools. Tier each by blast radius. Tier 1 is read-only, internal and reversible. Tier 2 faces customers or supports decisions. Tier 3 acts autonomously, touches money or regulated decisions, or calls other agents. Let tier set observability depth, evaluation frequency and containment authority, which keeps spend proportional to exposure.	Can we list every AI system and its tier on one page?
Phase 2	See what they do	Instrument the data layer first: freshness, lineage, schema stability and quality scored at the source. Then move evaluation into production, with thresholds and escalation paths set before launch and the same checks wired into CI/CD as release gates. Treat the governance and control layer as part of the architecture from the start, as Techment's blueprint does, instead of a later addition.	Can we trace a bad answer to its data, prompt and model version?
Phase 3	Be able to stop them	Measure detect, explain and contain times for each tier. Name one owner of the contain clock with authority to pause a system without committee approval. Capture complete, tamper-evident, retention-aligned records of prompts, responses, policies applied and agent decision paths, and adopt open telemetry standards so the evidence survives changes in models and vendors.	Can we show, on request, what the system did and how fast we would stop it?

A phased path to trustworthy AI, with an exit test for each phase that tells you when you're ready to move on | Source: Author.

Conclusion

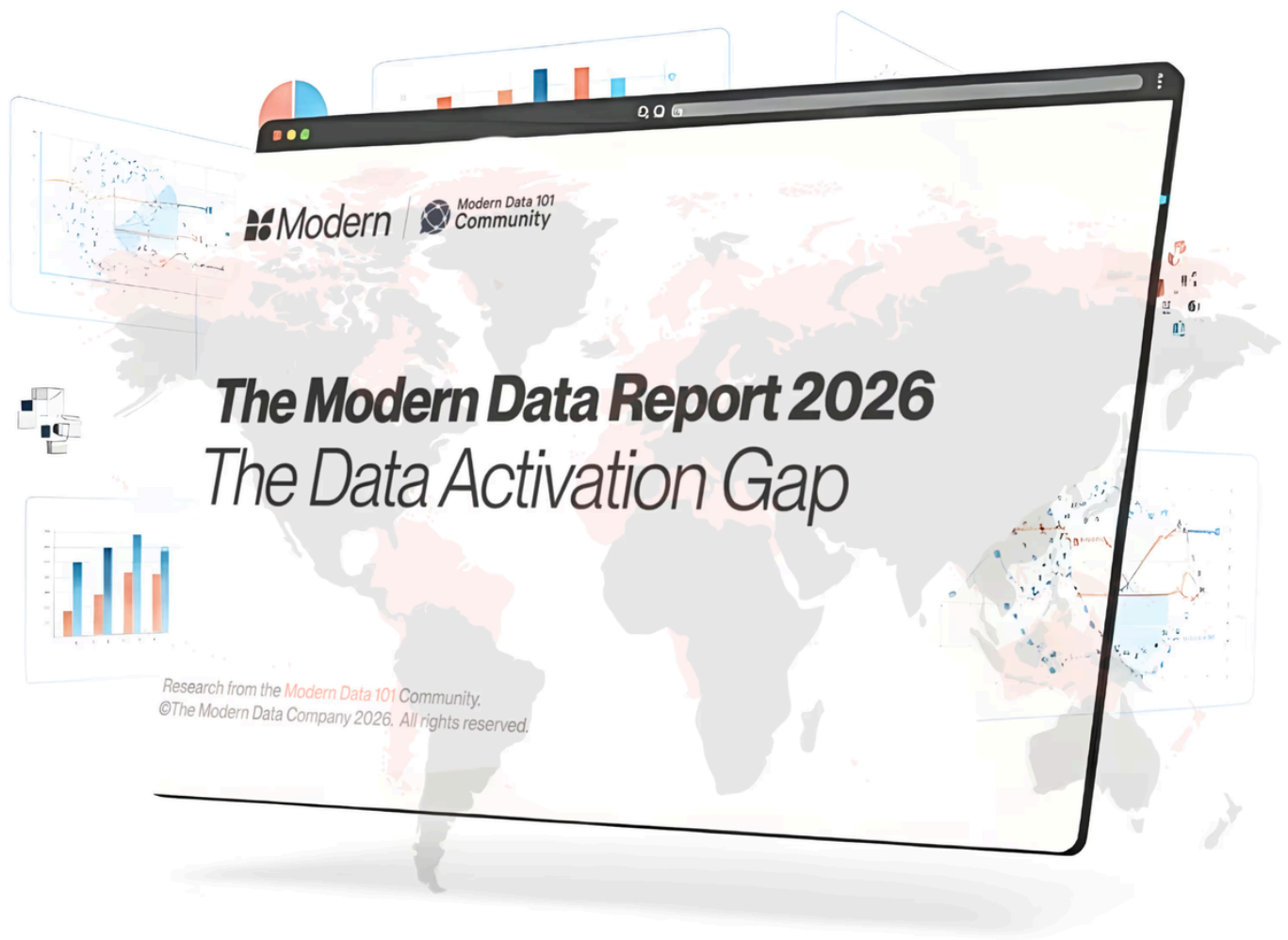
The next twelve months will turn observability data into something enterprises are asked to produce. Customers in regulated sectors will request traces of specific AI decisions, auditors will test whether controls fired when they should have, and regulators will expect records generated at the time rather than reconstructed afterwards. Organisations with short trust latency and complete evidence will sell AI into workflows that others are locked out of, since Gartner's warning that weak foundations confine GenAI to low-risk internal work describes a ceiling that observability raises.

Model choice will matter less to that outcome than most roadmaps assume. The enterprises that earn the right to automate consequential work will be the ones that can show, on request, what their AI did, why it did it, and how quickly they would have stopped it.

References

- [1] Gartner Predicts Explainable AI Will Drive LLM Observability Investments to 50% by 2028. Gartner, March 2026. <https://www.gartner.com/en/newsroom/press-releases/2026-03-30-gartner-predicts-by-2028-explainable-ai-will-drive-llm-observability-investments-to-50-percent-for-secure-genai-deployment>
- [2] Gartner Predicts 40% of Organizations Deploying AI Will Use AI Observability by 2028. Gartner, May 2026. <https://www.gartner.com/en/newsroom/press-releases/2026-05-12-gartner-predicts-40-percent-of-organizations-deploying-ai-will-use-ai-observability-to-monitor-model-performance-by-2028>
- [3] Gartner Identifies Six Steps to Manage AI Agent Sprawl. Gartner, April 2026. <https://www.gartner.com/en/newsroom/press-releases/2026-04-28-gartner-identifies-six-steps-to-manage-artificial-intelligence-agent-sprawl>
- [4] AI Agents Are Scaling Faster Than Their Guardrails. Deloitte Insights, survey of 3,235 leaders, 2026. <https://www.deloitte.com/us/en/insights/topics/emerging-technologies/ai-agents-scaling-faster.html>
- [5] State of AI Trust in 2026: Shifting to the Agentic Era. McKinsey & Company, survey of roughly 500 organizations, March 2026. <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era>
- [6] Predictions 2026: AI Agents and Enterprise Software. Forrester, 2025. <https://www.forrester.com/blogs/predictions-2026-ai-agents-changing-business-models-and-workplace-culture-impact-enterprise-software/>
- [7] Observability: A Critical Ingredient for AI and Agents. PwC US, February 2026. <https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-observability.html>
- [8] Is Your AI Reliable Enough to Be Trusted? HFS Research with TCS, survey of 101 US and Canadian executives, August 2026. <https://www.hfsresearch.com/research/ai-reliable-trusted/>
- [9] Trust but Verify: Why Enterprises Need AI Observability. Infosys Knowledge Institute, August 2026. <https://www.infosys.com/iki/perspectives/trust-verify-enterprises-ai-observability.html>
- [10] AI Observability: The Complete Guide for Enterprise Teams. Liminal, 2026. <https://www.liminal.ai/blog/ai-observability-enterprise-guide>
- [11] AI Reliability: Building Trustworthy AI Beyond Model Accuracy. NASSCOM Community, 2026. <https://community.nasscom.in/communities/ai/ai-reliability-building-trustworthy-ai-beyond-model-accuracy>

- [12] Why AI Observability Will Be Essential for Enterprises by 2026. USDSI, October 2025. <https://www.usdsi.org/data-science-insights/why-ai-observability-will-be-essential-for-enterprises-by-2026>
- [13] How Enterprises Build AI They Can Trust. UST, December 2025. <https://www.ust.com/en/insights/how-enterprises-build-ai-they-can-trust>
- [14] Why AI Trust Will Define Enterprise Leadership in 2026. Mindbreeze, January 2026. <https://www.mindbreeze.com/blog/why-ai-trust-will-define-enterprise-leadership-in-2026>
- [15] AI in 2026: Trustworthy, Safe, and Governed AI Systems. Keyrus, 2026. <https://keyrus.com/uk/en/insights/ai-in-2026-how-to-build-trustworthy-safe-and-governed-ai-systems>
- [16] Enterprise AI Strategy in 2026. Techment, 2026. <https://www.techment.com/blogs/enterprise-ai-strategy-in-2026/>
- [17] Building Trustworthy AI: Governed Execution. Oracle, 2026. <https://blogs.oracle.com/ai-and-datascience/building-trustworthy-ai-at-oracle>
- [18] The Modern Data Report 2026: The Data Activation Gap. Modern Data 101 Community, survey of 540+ practitioners, 2026. <https://www.moderndata101.com/modern-data-report-2026>
- [19] What Is AI Observability? Enterprise Stack and Guide for 2026. Modern Data 101, August 2026. <https://www.moderndata101.com/blogs/blogs-what-is-ai-observability-guide>
- [20] How AI Observability Improves Decision Quality. Modern Data 101, April 2026. <https://www.moderndata101.com/blogs/how-ai-observability-improves-decision-quality>
- [21] Enterprise AI Observability Trends of 2026. Modern Data 101, July 2026. <https://www.moderndata101.com/blogs/top-trends-of-enterprise-ai-observability>
- [22] The Agentic Enterprise (Vol. 03) and The AI Governance Benchmark (Vol. 04). Modern Data 101 Community Whitepapers, 2026.



540+ Respondents, 64 Countries, 29 Industries

[Access the report now](#)



Modern Data 101 Community



Subscribe 
moderndata101.com