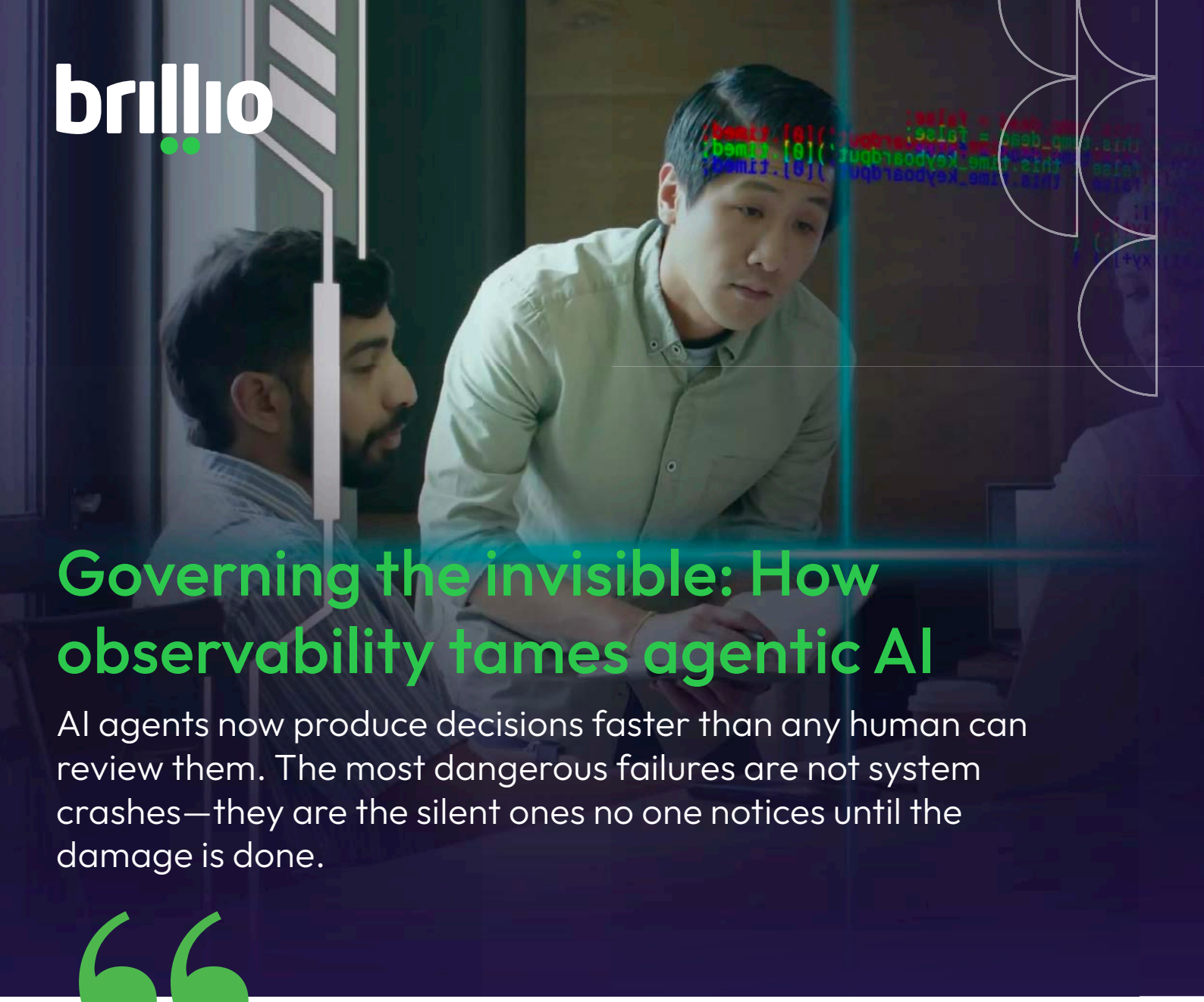


Governing the invisible: How observability tames agentic AI

AI agents now produce decisions faster than any human can review them. The most dangerous failures are not system crashes—they are the silent ones no one notices until the damage is done.

What enterprise leaders need to know about agentic AI observability

- As AI agents gain autonomy, invisible failures like contextual drift, cost overruns, and ungrounded reasoning become the greatest enterprise risk.
- Full-chain observability, from intent to action to outcome, is the baseline requirement for governing what cannot otherwise be seen.
- Our Control Tower methodology provides a structured framework for instrumentation, evaluation, accountability, and intervention without sacrificing agent autonomy.
- A tiered evaluation approach, from LLM-as-a-judge to human-in-the-loop, lets enterprises match monitoring cost to each agent's risk and ROI profile.

Why agentic AI systems fail and why most enterprises cannot see it

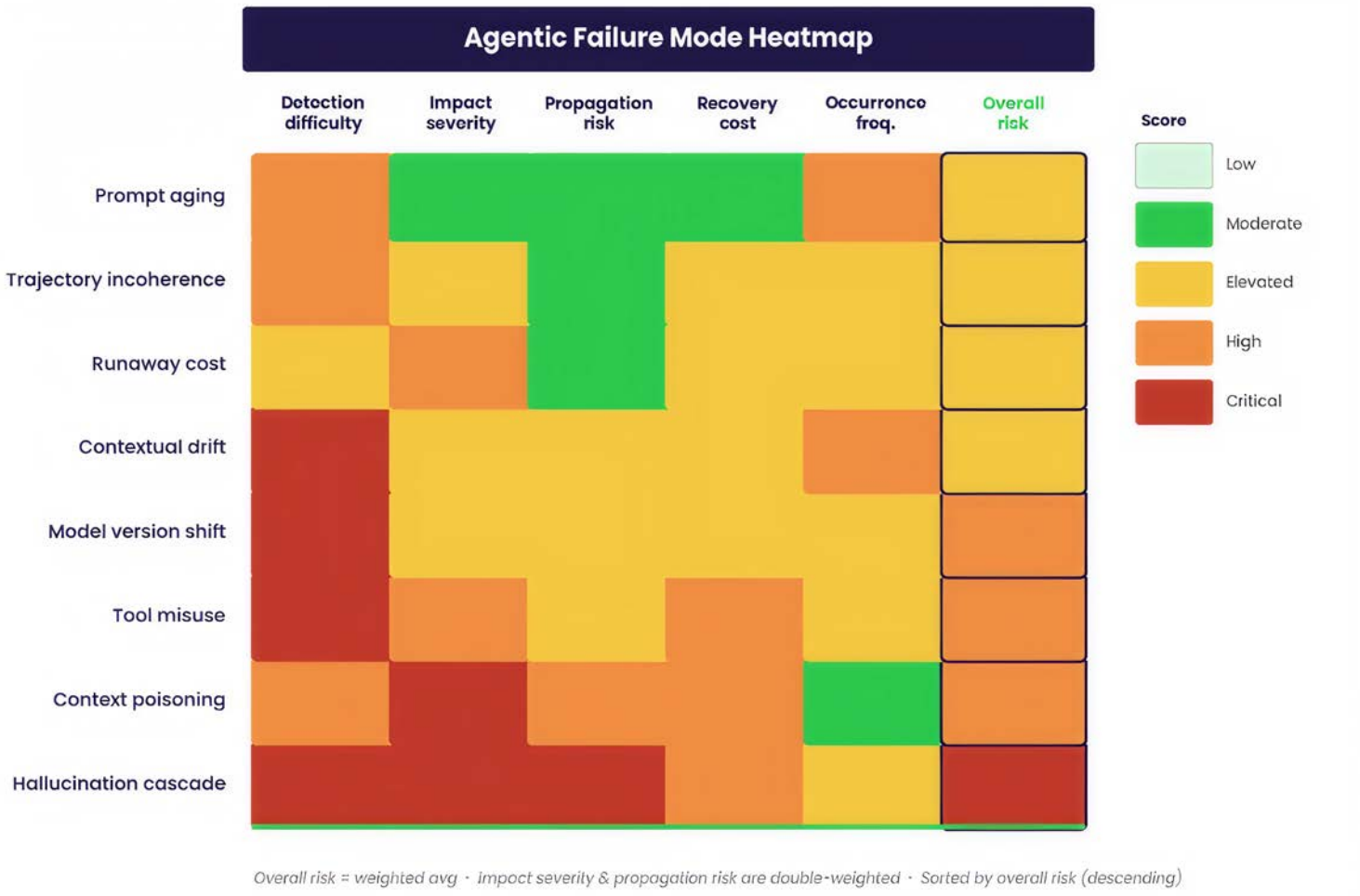
The shift from single large language model (LLM) calls to multi-orchestrated agentic systems happened deceptively quietly. Orchestrators spawned sub-agents. Sub-agents picked up tools. Tools called application programming interfaces. And humans stepped back, occasionally looping in, but less frequently. The surface area of autonomy expanded. Reasoning capability pushed toward frontier intelligence. Agentic systems grew powerful enough to handle tasks that once required entire teams.

Then something unexpected happened. Intelligence and insights began to be generated faster than they could be consumed. The volume of AI-produced output outpaced human capacity to review it. Attention dropped. The loop between AI action and human verification stretched thinner. This is the context in which observability of agentic systems has become not merely useful but essential. When human attention is scarce and agents operate at high velocity, the invisible failures are the most dangerous—silent contextual drift, unchecked cost accumulation, and decisions made on ungrounded reasoning that no one noticed in time.

Observability is the first step toward governance and control. What is not understood cannot be governed. Our Control Tower is built on that foundation: a structured methodology for making agentic systems legible, accountable, and steerable without sacrificing the autonomy that makes them valuable in the first place.

The failure modes already on record

The failure modes of agentic systems are already documented. In 2025 alone, hallucination in professional deliverables, agents misusing tools beyond their intended scope, ungrounded outputs presented as authoritative, and systems that broke down under operational scale all made headlines. The consequences were concrete: financial penalties, legal rulings, and reputational damage.



Five guiding principles that make agentic observability operational

Five principles shape the Control Tower methodology. These aren't aspirational statements—they are the operating constraints that make the framework function.

1. Continuous quality and drift control

Agent failures are silent: gradual drift in an agent's context, a prompt that has aged past its useful life, a model update that subtly shifted behavior. Left undetected, contextual drift erodes agent reputation and degrades customer experience in ways that are difficult to trace back to their source.

2. End-to-end observability across the full agent chain

When observability covers only a slice of the agentic interaction—the tool call but not the reasoning, the output but not the trajectory—it increases the risk of decisions that appear grounded but aren't. Full-chain visibility, from intent to action to outcome, is the baseline requirement.

3. Predictability in cost and ROI

Agents can generate costs that are difficult to anticipate and easy to ignore until they become a crisis. A three-to-six-month cost estimation tracker, paired with monthly ROI mapping, converts agent economics from a surprise into a managed variable.

4. Trust and adoption through responsible AI

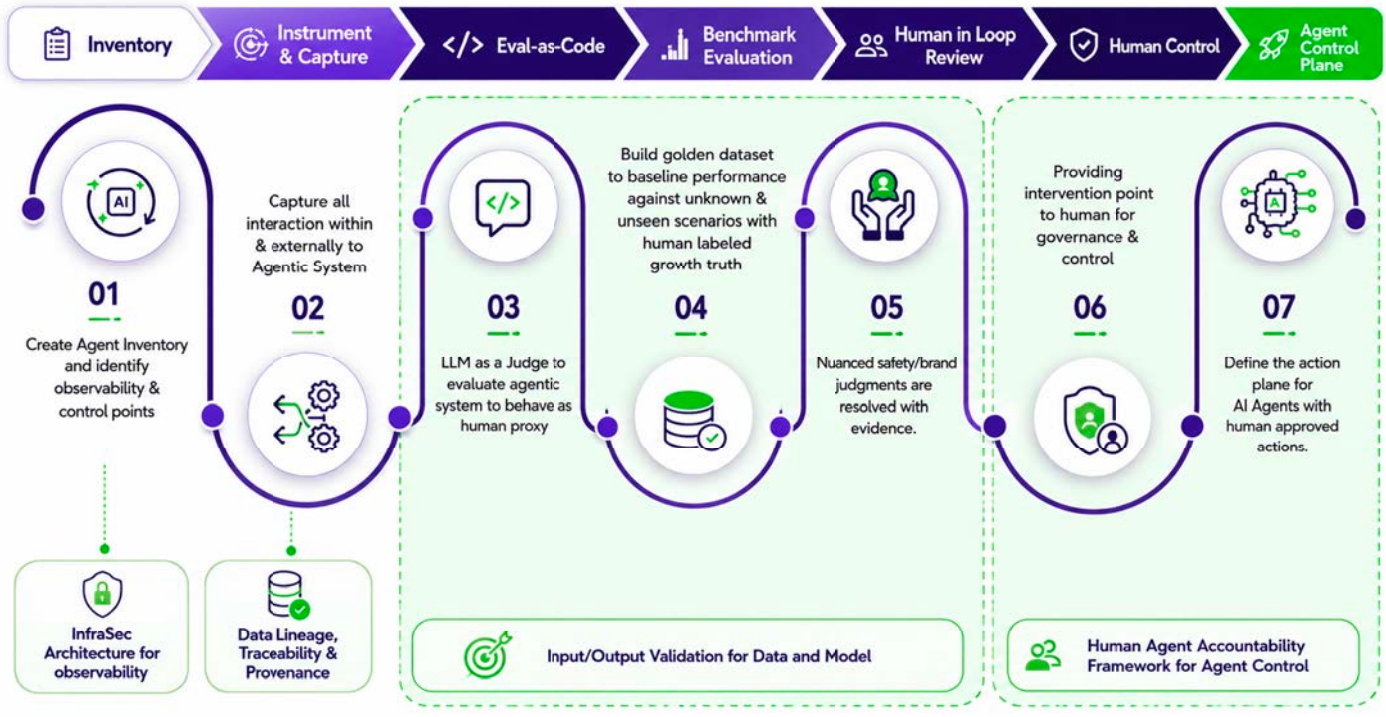
The principles of responsible AI—transparency in how agents reason, explainability of agent decisions, fairness in outputs, and safety guardrails that constrain harmful actions—are the foundation on which enterprise-wide adoption can be built. Without them, agents remain in sandboxes.

5. Centralized governance as a speed breaker, not a barricade

The Control Tower is designed to be a mechanism that enforces a brief, deliberate pause before an agent is granted autonomy—not a barrier that prevents it from operating. The goal is to enable teams to move faster with confidence, not slower with permission slips.

How our Control Tower methodology works

Holistic observability is not an afterthought bolted onto a running system. It is an input signal to AI strategy—designed before the first agent goes to production or is even designed, not diagnosed after the first failure.



Use-case inventory

The Control Tower is built on a concrete inventory of the use cases it governs. This inventory is the reference plane that aligns the observability architecture with the broader InfraSec architecture. Before any instrumentation is deployed, the estate of agents must be cataloged: what each agent does, what tools it can access, what data it touches, and what decisions it is empowered to make.

Instrumentation

Instrumentation captures the raw entities that support every interaction within and external to the agentic system. This includes agent inputs and outputs, tool invocations and their results, latency, cost, model versions, prompt versions, and context state at each step. Instrumentation is the prerequisite for everything that follows.

Evaluations

Evaluations are the core intelligence layer of the Control Tower; the mechanism by which raw instrumentation data is converted into meaningful signal about agent quality, reliability, and safety. Three approaches represent a deliberate hierarchy of cost and reliability:

- **LLM-as-a-judge:** The fastest and most affordable evaluation method. An LLM scores agent outputs against defined criteria. Lower implementation complexity, lower cost, and lower reliability suitable as a first-pass signal or for lower-risk use cases.
- **Benchmarking with a golden dataset:** Agent outputs are evaluated against a curated benchmark dataset annotated by human subject-matter experts. Moderate in cost and complexity to establish but delivers meaningfully higher reliability. The quality of the evaluation is directly tied to the quality of the dataset.
- **Human-in-the-loop:** The most granular and most reliable evaluation approach. Human reviewers assess not just outputs but the agent's full decision trajectory: the reasoning steps, tool selections, and intermediate states that led to the outcome. Higher cost, but the appropriate choice for high-risk, high consequence use cases.

Human intervention trigger

Once an evaluation indicator crosses a defined threshold, it surfaces to the human operator as an intervention point. The design principle here is important: the kill switch belongs to the human, not the system. The Control Tower surfaces the signal and presents the recommendation. The human makes the call: to pause the agent, to investigate, or to let it continue. This preserves accountability without removing autonomy from agents during normal operation.

Agent action plane

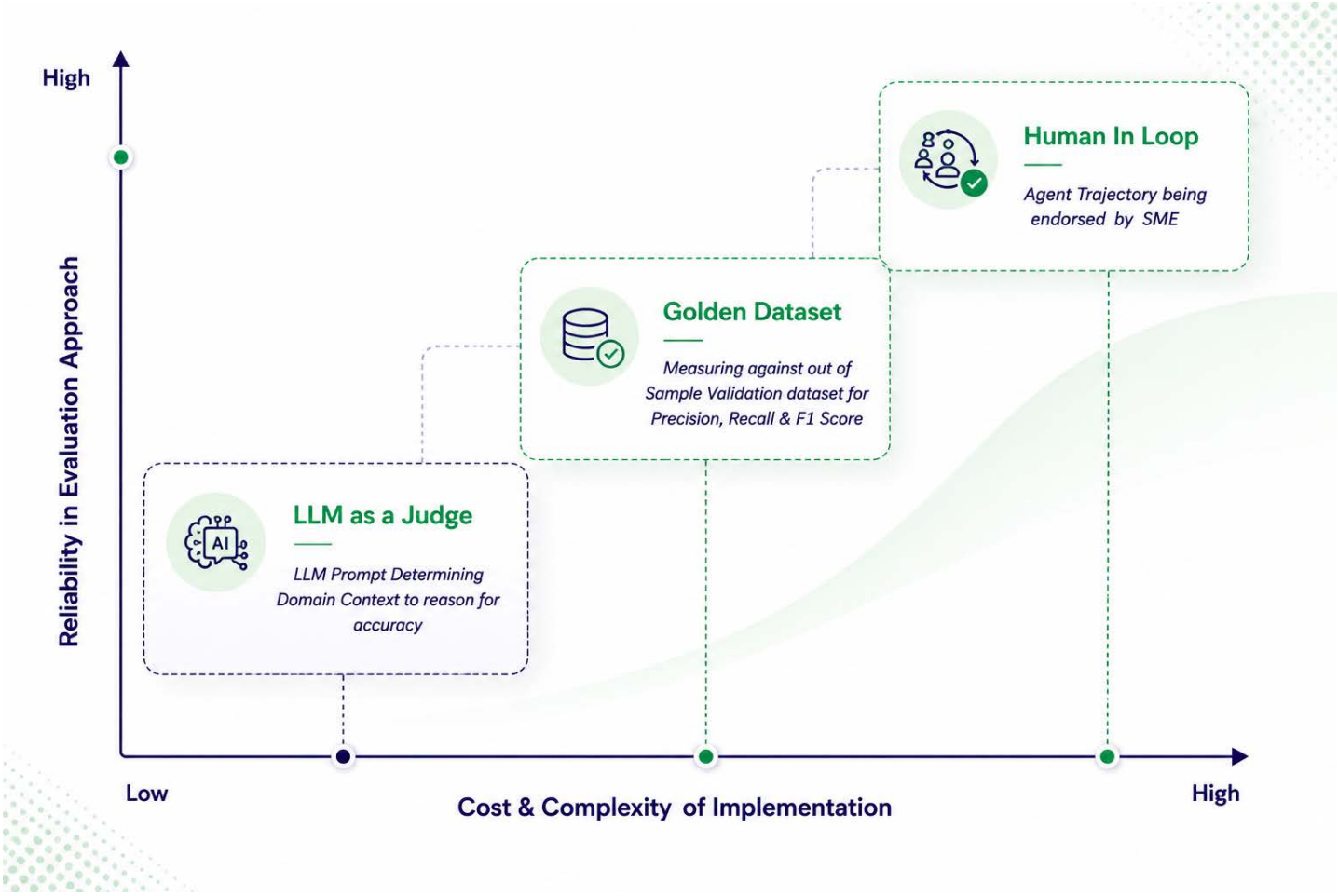
The approved action plane is the boundary defined by humans in advance—a set of permitted actions, thresholds, and escalation paths that the Control Tower agent can execute without further approval. This is governance through bounded autonomy: the Control Tower acts, but only within a space that humans have explicitly sanctioned.

Measuring agentic AI performance without gaming the metrics

When defining key performance indicators (KPIs) for an agent catering to business needs, two parameters must be kept in mind: the complexity of implementing the metric and the reliability of the approach used to derive it. The progression starts with LLM-as-a-judge, advances as clarity grows in mapping agentic system performance to business value and moves toward benchmarking with a golden dataset annotated by SMEs, enabling the business to project ROI estimations with confidence. The highest level of measurement maturity involves a real-time, dynamic human-in-the-loop approach where each agent trajectory is assessed by an SME.

“When a measure becomes a target, it ceases to be a good measure. Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.” — Goodhart's Law

Goodhart’s Law must be upheld in every measurement framework. The benchmarking golden dataset needs to be periodically aligned with evolving demand to prevent metric gaming and ensure evaluation fidelity.



What every enterprise must answer before granting agent autonomy

Accountability matrix

Governance without a clear accountability model collapses into bureaucracy. The accountability matrix gives the methodology its operational structure by resolving four fundamental questions before any agent is granted autonomy.

Phase of monitoring

Monitoring is not a post-deployment activity. It must be considered at discovery and design. The learnings from the evaluation methodology must be incorporated into the agent development lifecycle so that observability is a native property of the agent, not a layer applied after a failure. Teams that build observability from the start enter production with baselines already established.

Objective of control

The control objective of an agent determines the right granularity of observability and the right trade-off between evaluation cost and control granularity. A low-stakes informational agent with a narrow tool scope needs lighter observability than an autonomous agent that writes to systems of record, triggers financial transactions, or communicates externally on behalf of the organization. Defining the control objective before deployment allows teams to make a principled decision about which evaluation tier is appropriate, how frequently to evaluate, and what thresholds should trigger intervention.

Responsibility

Who is accountable for each agent's actions—and who is accountable for each monitoring action—must be defined before the agent is handed autonomy. This isn't a question to be resolved after an incident. The responsibility map covers the agent builder, the agent owner, the monitoring function, the escalation path when an intervention trigger fires, and the business owner it serves. Clear ownership at each point in the chain converts an observability framework into an operational reality.

Frequency of evaluations and monitoring

Continuous evaluation at the highest fidelity for every agent in the estate is financially unsustainable. The frequency decision is a deliberate trade-off between the total cost of ownership of the monitoring infrastructure and the ROI the agent generates. High-ROI agents operating in high-risk domains justify high-frequency, high-fidelity evaluations. Lower-value, lower-risk agents can operate on lighter cadences. What matters is that the trade-off is made explicitly, documented, and revisited as the agent matures and its risk profile changes.

From signal to action: How the situation-reaction assessment drives decisions

The situation-reaction assessment answers the question that every monitoring framework eventually faces: now that there is a signal, what should be done with it? It maps specific metric patterns—individual thresholds and multi-metric combinations—to the mitigations most likely to resolve the underlying risk, giving teams a structured path from detection to action rather than leaving response to improvisation.

The table below is illustrative and not an exhaustive list.

Dimension	Situation	Impact	Control Tower observability metrics	Business decisions enabled
Cost efficiency	**Finance team's multi-agent payment and dispute system running large LLMs with consistently low context window utilization (avg 13%) week-over-week (WoW)**	**LLMs billed for full context window regardless of usage, resulting in $18,400 per month wasted. Zero accuracy benefit from oversized models**	Context window utilization % (WoW trend), token consumption per agent, cost per transaction, LLM vs. small language model (SLM) cost differential	**Downgrade LLM to domain SLM for low-context tasks; right-size model per agent workload; reduce LLM spend 70% with no accuracy loss**
Security and trust	**AI platform reputation falling. Data clean, knowledge base (KB) audited, infra healthy — but 147 prompt injections and 89 jailbreak attempts per week silently corrupting all three agents**	**Tool selection accuracy collapsed to 34%. Wrong dispute routing, wrong payment decisions. Reputation dropped to 38% with zero alerts fired**	Prompt injection attempt frequency, jailbreak detection rate, tool selection accuracy %, agent trust score, silent failure event count	Deploy prompt guardrails and input validation; activate tool-call audit layer; isolate injection vector per agent; restcre reputation via targeted hardening
Answer quality	Agent responses not meeting user expectations. Team debating model vs. KB vs. retrieval-augmented generation (RAG) vs. prompts with no measurement — two agents, two different root causes	Vague, incomplete answers degrading user trust. Resources wasted on wrong fixes. Two distinct failure patterns invisible without metric tracking	Relevance score (WoW), similarity score (WoW), coverage score (WoW). Pattern A: all three declining. Pattern B: similarity stable, relevance and coverage falling	Pattern A: KB freshness refresh and fallback layer. Pattern B: Prompt completeness enforcement and task validation. Prescribe different fix per agent based on metric pattern
Embedding intelligence	General embedding model unable to distinguish 'chargeback,' 'reversal,' and 'refund' in cross-border contexts. Dispute analyst and compliance auditor agents building inconsistent world models	**Routing accuracy at 34–61%. Two agents disagreeing on same case. Embedding gap score rising 0.12 to 0.71 over eight weeks. Query and response vectors diverging geometrically**	Semantic gap score (query-to-response drift WoW), embedding cluster separation viz, routing accuracy per intent class, embedding engine latency and throughput, token efficiency %	Upgrade to domain-specific embedding model; shadow to canary to blue/green rollout (zero downtime); gate promotion on gap score < 0.15 and latency < 120 ms; align both agents to one shared semantic world
Portfolio resource allocation	**Q3 planning: $100K unallocated compute credits. Retail forecasting agent (VER 8.2x) and payment dispute agent (VER 1.4x) both lobbying for budget. C-suite choosing on instinct**	**Same $100K returns $820K via retail vs. $140K via payment. Without data, budget risks going to the louder agent, not the higher-yield one. Declining payment ROI (-18% WoW) invisible to leadership**	Value efficiency ratio (VER) per agent, 30-day run rate, ROI trajectory (WoW trend), cost per success (CPS), high-yield vs. high-maintenance classification	Deploy $100K immediately to retail action plane; zero budget to payment until sprint gates met; assign model downgrading sprint: CPS $18.50 to $3, VER 1.4x to 5x; unlock payment budget only when VER is 4.0x or higher

Where the ADAM Control Tower fits in the agentic AI observability landscape

The tools in this landscape provide instrumentation and evaluation infrastructure. The Control Tower is the governance operating model that sits above them, defining what gets instrumented, what thresholds mean, who owns the response, and under what conditions an agent is permitted to operate. The ADAM Control Tower provides the observability, accountability, and intervention layer that none of these tools supply natively.

Agent Evaluation Platform Comparison

Dimension	Arize Phoenix	Datadog LLM Obs	Galileo	AgentOps	ADAM Control Tower
Observe					
Agent behavior (step tracking & workflow)	**Strong:** OpenTelemetry-native distributed tracing, full tool call tracking, LLM span coverage and agent step visibility across multi-hop flows.	**Moderate:** LLM traces integrated into application monitoring; less purpose-built for agent trajectories.	**Moderate:** Behavior captured via logs and traces but limited real-time state transitions.	**Strong:** Purpose-built session replay, records tool invocations, state transitions and agent calls.	**Strong:** OTEL-native distributed tracing and full span/tool call visibility across multi-agent systems.
Data & Context Observability	**Strong:** Prompt/response logging, embedding drift detection, retrieved document relevance scoring and context window analytics.	**Moderate:** Payload capture with privacy masking but limited retrieval-specific analytics.	**Strong:** Deepest RAG observability including adherence, chunk relevance and completeness metrics.	**Moderate:** Session-level capture focused on logging rather than deep RAG evaluation.	**Strong:** Full observability of engineering pipeline, context retrievals, prompts and embedding vectors.
Measure					
Cost & Responsible AI	**Moderate:** Token usage tracking and custom hallucination evaluators.	**Strong:** Pre-built cost dashboards, budget alerts and governance controls.	**Strong:** Comprehensive AI suite with hallucination scoring, PII detection and toxicity metrics.	**Moderate:** Per-session cost tracking with limited policy enforcement.	**Strong:** Agent-level cost management, RBAC access and responsible AI metrics.
Eval & Quality Metrics	**Strong:** Custom evaluators, LLM-as-a-judge benchmarking and A/B comparison.	**Moderate:** Quality monitoring via SLOs and developer-defined thresholds.	**Strong:** Automated evaluations, retrieval quality scoring and benchmark comparisons.	**Moderate:** Task completion metrics and limited evaluation depth.	**Strong:** Three-layer evaluation framework with benchmarking and embedding drift analysis.
Control					
Next Best Action	**Limited:** Manual recommendations from dashboards.	**Moderate:** Alert-driven guidance and anomaly detection.	**Moderate:** Evaluation insights drive retraining and prompt improvements.	**Moderate:** Session replay assists root-cause analysis.	**Strong:** Contextual next-best-action recommendations benchmarked against control policies.
Guardrails & Governance	**Moderate:** Monitoring with custom hooks and alerting.	**Strong:** Incident lifecycle management, SLO enforcement, audit trail.	**Moderate:** Input/output guardrails and evaluation pipelines.	**Limited:** Visibility-focused with minimal runtime controls.	**Strong:** Agent control, runtime mitigation and governance processes.
Persona Fit					
Forward-Deployed Engineer	**Primary**	**Primary**	Supported	**Primary**	**Primary**
Business Analyst	Limited	Supported	**Primary**	Limited	**Primary**
CXOs	Limited	**Primary**	Supported	Limited	Supported

The case against over-instrumentation—and why observability still wins

Critics argue that observability layers add overhead and risk slowing autonomous systems. The concern is valid: over-instrumentation can become bureaucratic. Yet governance designed as a brief, deliberate pause enables teams to move faster with confidence, not slower with permission slips.

What leaders should do differently about agentic AI governance

- **Treat observability as a design input, not a diagnostic.** Agents built with instrumentation from the start enter production with baselines already established.
- **Match evaluation investment to risk and ROI.** High-value agents in high-risk domains justify costlier, higher-frequency evaluation; lower-stakes agents need lighter cadences.
- **Define accountability before granting autonomy.** The responsibility map covering builder, owner, monitoring function, and escalation path must be resolved before any incident.
- **Govern through bounded autonomy, not blanket restriction.** The goal is a deliberate pause before expanded scope, not a barrier that blocks agent velocity.

About Brillio

Brillio is The Enterprise AI Accelerator helping Fortune 1000 companies move from AI ambition to scaled impact, faster. Powered by our AI accelerator platform – Agentic Data and Application Management (ADAM), Brillio is one of the fastest-growing digital technology service providers, delivering transformation across five core workstreams: business-led transformation, customer experience transformation, AI and data engineering, digital engineering, and infrastructure engineering.

With 14 delivery locations across North America, Europe, and Asia and a team of over 6,000 customer-obsessed professionals, Brillio combines deep industry expertise, modern engineering, and accelerators to deliver measurable outcomes. Headquartered in Dallas, Texas, Brillio serves clients globally with a commitment to speed, scale, and measurable impact.

https://www.brillio.com/
Contact Us: info@brillio.com

brillio