HITACHI

Hitachi Application Reliability Centers' eBook

Operationalizing GenAI and AI at Scale for **Enterprises**

Hitachi Digital Services

Generative AI (GenAI) systems have revolutionized various industries by enabling the creation of sophisticated models that can generate human-like text, images and other content. However, the deployment and management of these systems in production environments pose significant challenges. This guide outlines practical strategies to overcome these challenges by ensuring the reliability, observability, and overall performance of GenAI systems.

The Rise of Generative AI: Transforming Enterprises

AI and Generative AI are becoming increasingly accessible to non-technical users, as major companies invest in user-friendly interfaces to enhance productivity and creativity. By 2025, worldwide spending on generative AI is projected to reach $644 billion, a 76.4% increase from 2024, according to Gartner [1].

Enterprise adoption is accelerating rapidly. According to Gartner, 30% of enterprises will have implemented generative AI in at least one business function by 2025, up from 5% in 2021 [2]. This reflects a broader trend of integrating AI into core business workflows, with adoption rates expected to continue rising.

[1] Gartner Forecasts Worldwide GenAI Spending to Reach $644 billion in 2025
[2] Gartner on Generative AI Applications and Adoption

Challenges in AI & GenAI Production Operations

01

Reliability & Observability

Challenge: Ensuring that AI & GenAI systems perform reliably at scale with the required precision is a significant challenge. GenAI systems are complex and can exhibit unpredictable behavior, making it difficult to maintain reliability.

Solution: Build robust architectures and implement comprehensive monitoring and management strategies. Ensure that AI systems operate at the desired levels of scale, precision and consistency by defining key Service Level Indicators (SLIs) and Service Level Objectives (SLOs) and designing observability instrumentation. (See **Approach to AI Observability** for further technical details).

02

Security & Responsible AI

Challenge: Maintaining accuracy, fairness and data integrity is crucial for building trust in GenAI systems. Bias and hallucinations in AI outputs can undermine the reliability and acceptance of these systems.

Solution: To build trust in Generative AI systems, it's essential to ensure accuracy, fairness, and data integrity through diverse training data, grounded outputs, and robust data governance. Embedding Responsible AI metrics—such as fairness, robustness, and transparency—enables continuous monitoring of model behavior. By implementing custom-curated guardrails aligned with organizational values and risk thresholds, teams can proactively detect and mitigate bias, hallucinations, and drift. These guardrails support automated or human-in-the-loop interventions, ensuring timely corrective actions and keeping AI systems ethically aligned and trustworthy over time.

03

Cost & Sustainability for AI

Challenge: Managing the compute, storage and network capacity required for AI systems as demand grows is challenging. Additionally, controlling costs and carbon emissions across on-premises, cloud and third-party APIs is essential for sustainable operations.

Solution: Determine the compute, storage and network capacity needed as AI demand expands. Manage costs and carbon emissions by optimizing model selection, tuning, training, inference and hosting costs.

04

Managing AI Workloads

Challenge: AI systems in production environments face unique operational challenges—ranging from unpredictable scaling demands and high infrastructure costs to model drift, security vulnerabilities, and compliance pressures. Traditional IT operations often lack the AI-specific tooling and processes needed to manage these complexities effectively.

Solution: Managing AI workloads requires a purpose-built operational model that integrates intelligent provisioning, cost-aware scaling, proactive maintenance, and robust security and governance. By implementing dynamic environment management, AI-specific capacity planning, real-time model monitoring, and automated incident response, organizations can ensure their AI systems remain performant, secure, and compliant. This integrated approach directly supports the operational needs outlined above—enabling enterprises to scale AI confidently while minimizing risk and maximizing value.

Hitachi Digital Services' Approach to AI Observability

01

Comprehensive Observability

Set up comprehensive observability tools and dashboards to monitor various aspects of GenAI systems. Cover key areas such as usage, prompt, data, response, model, infrastructure and cost. This holistic approach ensures that all critical components of the GenAI system are monitored and managed effectively.

02

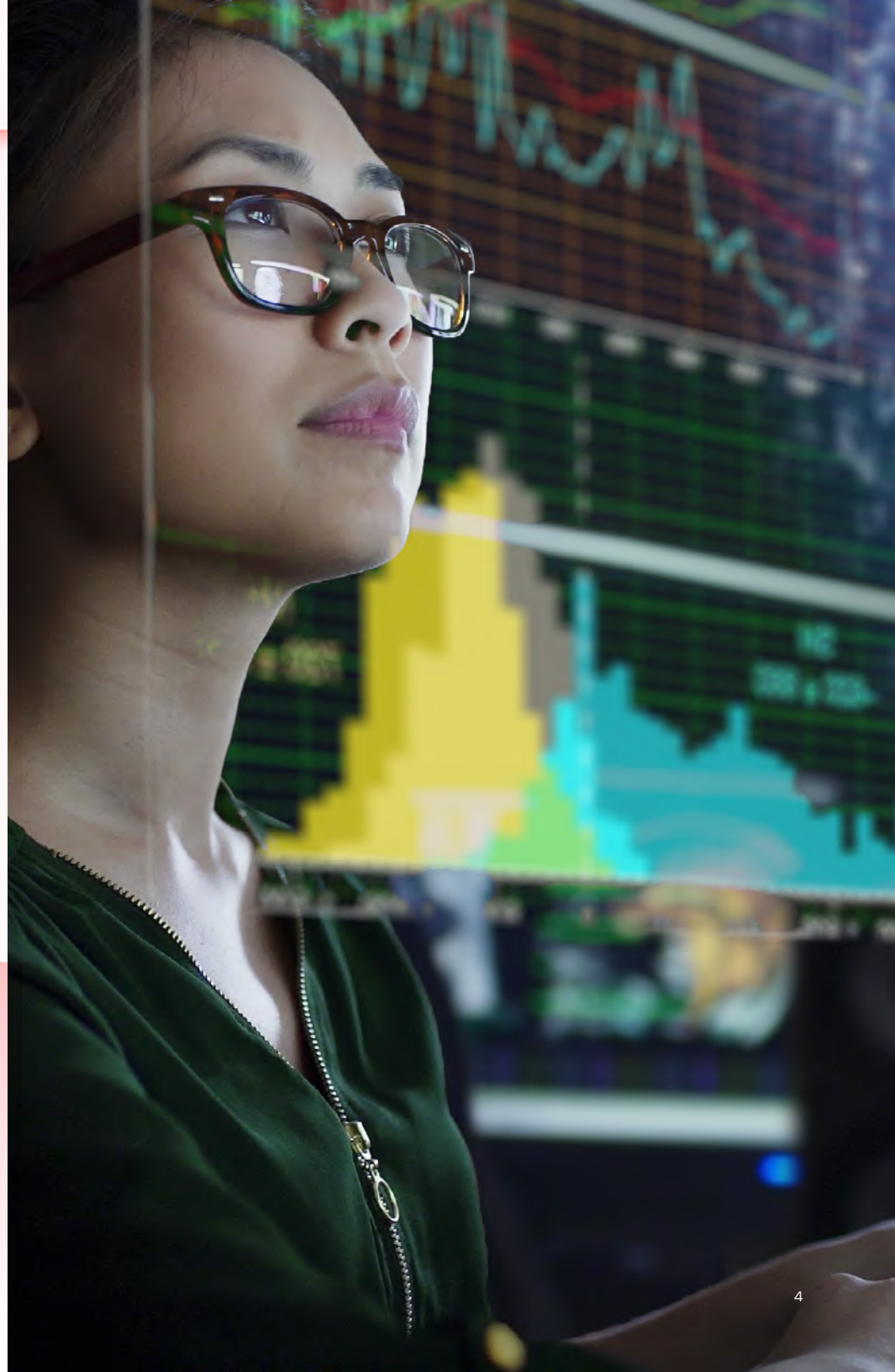
Integration with SRE Principles

Integrate Site Reliability Engineering (SRE) principles into GenAI systems management. Monitor GenAI observability metrics in a manner similar to traditional Non-Functional Requirements (NFRs), SLIs, SLOs and error budgets. This integration ensures that GenAI systems are managed with the same rigor and discipline as traditional software systems (see figure 1).

03

Focus on Key Metrics

To ensure the stability, performance and trustworthiness of GenAI applications in production, it is essential to define and monitor a comprehensive set of operational and performance metrics. These metrics provide visibility into system behavior, support proactive issue resolution, and guide continuous optimization.



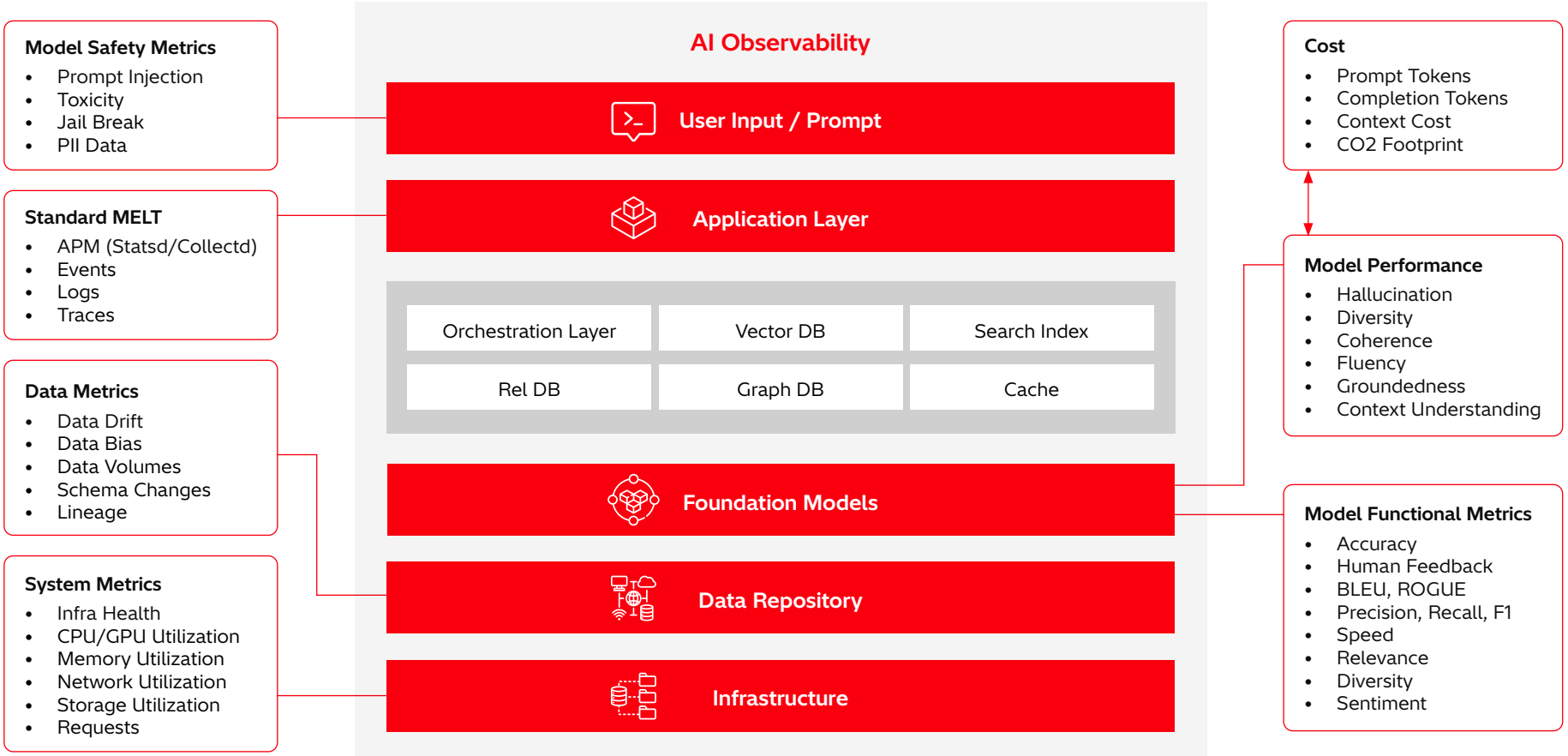


Figure 1

Key Aspects to Consider in AI Observability

Holistic Coverage: The observability approach covers the entire AI lifecycle, from conception, prototyping to production. It ensures that observability is maintained throughout all stages of application development.

Advanced Tooling: Leverage both open-source and commercial tools for observability. This includes tools like Langkit, LangSmith, MLOps, Arize, Phoenix, TensorBoard, AgentOps, WhyLabs, Domino, Azure Monitor, AWS CloudWatch, SageMaker Clarify, Seldon and Wandb.

Actionable Insights: Provide actionable alerts and dashboards that help in real-time monitoring and decision-making. Focus on providing insights that can be used along with prompt engineering and fine-tuning techniques to improve overall system performance.

Leveraging Learning Language Model (LLM) Tracing

Purpose of LLM Tracing

LLM tracing is a method used to monitor, analyze and debug the execution of large language models. It provides a detailed snapshot of each operation or invocation within an LLM application. This includes tracking API calls, prompt formatting and model responses. By organizing these operations into a trace—a collection of runs (or spans) organized in a tree or graph structure—developers can gain a comprehensive view of how their models operate.

Benefits of LLM Tracing

Comprehensive Context Capture:
Tracing captures the full execution context, including API calls, context, prompts and parallel operations. This detailed view allows developers to understand the entire lifecycle of a request, making it easier to diagnose issues and comprehend the model's behavior.

Performance Monitoring:
Tracing provides insights into latency times, token usage and the sequence of operations. This data helps identify bottlenecks and optimize model performance, ensuring efficient resource utilization.

Cost Tracking:
Managing the costs associated with model usage is another important aspect of LLM tracing. By monitoring token usage and API calls, developers can track expenses and make informed decisions about resource allocation and budget management.

Error Identification & Debugging:
Tracing helps in pinpointing low-quality outputs and understanding their origins. This is crucial for debugging issues, such as why certain responses might be unsatisfactory or incorrect. It enables developers to troubleshoot and refine model responses effectively.

User Feedback Collection:
Tracing allows for the collection of user feedback on model outputs. This feedback is essential for iterative improvement and fine-tuning, helping to enhance the model's accuracy and relevance over time.

Dataset Building:
Detailed logs from tracing can be used to build fine-tuning and testing datasets. This helps in creating datasets that reflect real-world usage, improving the model's performance and applicability in diverse scenarios.

Managing AI Workloads

Overview

Managing AI workloads requires a specialized operational framework that ensures AI systems run reliably, securely, and efficiently in production. This involves orchestrating cloud infrastructure, maintaining application health, enforcing security, and responding to incidents—all while optimizing for performance and cost.

Key Considerations

01

AI Platform Engineering

Design and operate scalable, production-grade AI platforms that support the full lifecycle of AI workloads—from experimentation to deployment. Maintain the health of AI applications through continuous monitoring, patching, and performance tuning.

Key Considerations: AI platforms must support heterogeneous compute, distributed training, and orchestration via Containerized platforms with ML-specific extensions (e.g., Kubeflow, Ray). For Agentic AI systems, platforms must also support long-running, stateful agents, memory stores, and event-driven workflows. CI/CD pipelines should include model reproducibility, prompt versioning, and automated predictable evaluation harnesses.

02

Data Quality & Reliability

To support reliable AI, GenAI, and Agentic AI systems, data quality and drift detection must be embedded into the data lifecycle. The process begins with enforcing strict data validation rules at ingestion—checking for schema consistency, completeness, and accuracy. Continuous profiling is used to establish baseline statistics, enabling automated detection of anomalies and inconsistencies. As data flows through the pipeline, transformations should be tracked with metadata and versioning to ensure transparency and reproducibility. For drift monitoring, statistical methods are applied to detect shifts in feature distributions and model outputs over time. These signals must be integrated into alerting and retraining workflows to ensure that models remain aligned with real-world data. Closed-loop feedback from model performance in production further informs data quality improvements and adaptive retraining strategies.

Key Considerations: Monitor model performance, data drift, and Responsible AI metrics. For Agentic AI, track agent task success rates, memory coherence, and reasoning loops. Automate retraining pipelines triggered by drift thresholds and maintain reproducibility with MLflow or Weights & Biases.

03

Capacity Management
Ensure the system can elastically scale to meet AI workload demands without overprovisioning.

Key Considerations: AI workloads—especially those involving LLMs or agentic orchestration—can be bursty and unpredictable. Use predictive autoscaling based on historical usage, token throughput, and latency SLAs. Implement queue-based job scheduling (e.g., KubeFlow Pipelines, Ray Serve) and resource-aware schedulers to optimize resource allocation.

04

Incident Management
Detect and resolve AI-specific incidents with minimal downtime. Leverage SRE principles like Direct Responsible Individual (DRI) and Shared-Responsibility model for proactive review and recalibrating thresholds.

Key Considerations: AI incidents may involve silent failures (e.g., hallucinations, degraded reasoning). Use anomaly detection on model outputs, integrate tracing (e.g., LangSmith, OpenTelemetry), and implement rollback mechanisms for model and prompt versions.

05

Security
Apply security updates across infrastructure, models, and data pipelines.

Key Considerations: AI systems are vulnerable to adversarial attacks, model inversion, and prompt injection (especially in LLM agents). Secure model artifacts with signed containers, encrypt memory stores, and enforce RBAC/IAM for agent actions. Regularly scan for CVEs in ML libraries and LLM frameworks.

06

Cost Optimization
Implement FinOps-aligned strategies to manage the high cost of AI workloads across training, inference, and agent orchestration.

Key Considerations: Use cost-per-inference and cost-per-agent-task as KPIs. Optimize for idle resource detection, autoscaling, and workload bin-packing. For Agentic AI, monitor agent orchestration overhead and memory usage. Leverage infrastructure optimization like low cost regions, spot/preemptible instances for non-critical tasks, apply prompt engineering and model compression to reduce inference costs.

To operationalize GenAI at scale, enterprises must adopt a resilient architecture grounded in observability, SRE-aligned metrics, and AI-specific workload orchestration. Hitachi Digital Service's HARC delivers this foundation—enabling secure, cost-optimized, and continuously improvable AI systems that meet the demands of modern production environments.

Hitachi Application Reliability Center (HARC): Powering Resilience at Scale

Since its inception, the Hitachi Application Reliability Center (HARC) has empowered numerous Fortune 500 enterprises to operate mission-critical workflows with enhanced resilience, security, and cost efficiency. Spanning high-stakes industries such as finance, healthcare, defense, and automotive, HARC has consistently delivered 99.8% uptime across enterprise environments.

Our Impact is Measurable:

$150+ million in total client savings

delivered by eliminating avoidable outages, streamlining operations, and optimizing cloud expenditures.

$14+ million in revenue leakage reduced

for a leading U.S. travel technology firm.

150+ high-risk vulnerabilities

remediated, reinforcing enterprise security postures.

HARC continues to be a trusted partner in building secure, compliant, and high-performing digital ecosystems for data-intensive enterprises worldwide.

Authors

Shrinath Venkatsubramaniam,
Chief Architect,
Center for Architecture & AI (CAAI) and HARC

Vitor Domingos,
Lead Solution Architect,
CAAI

Marimuthu Muthusamy,
Vice President,
HARC

Sai Subramanian,
Chief Architect,
CAAI and HARC

Learn More

Explore how Hitachi Digital Services can support your organization with next-generation application reliability and strategic IT solutions.
Website: Hitachi Application Reliability Centers | **Press Release:** Five-Year Strategic IT Partnership with Envista
Visit our website or read the press release to dive deeper into our approach and success stories.