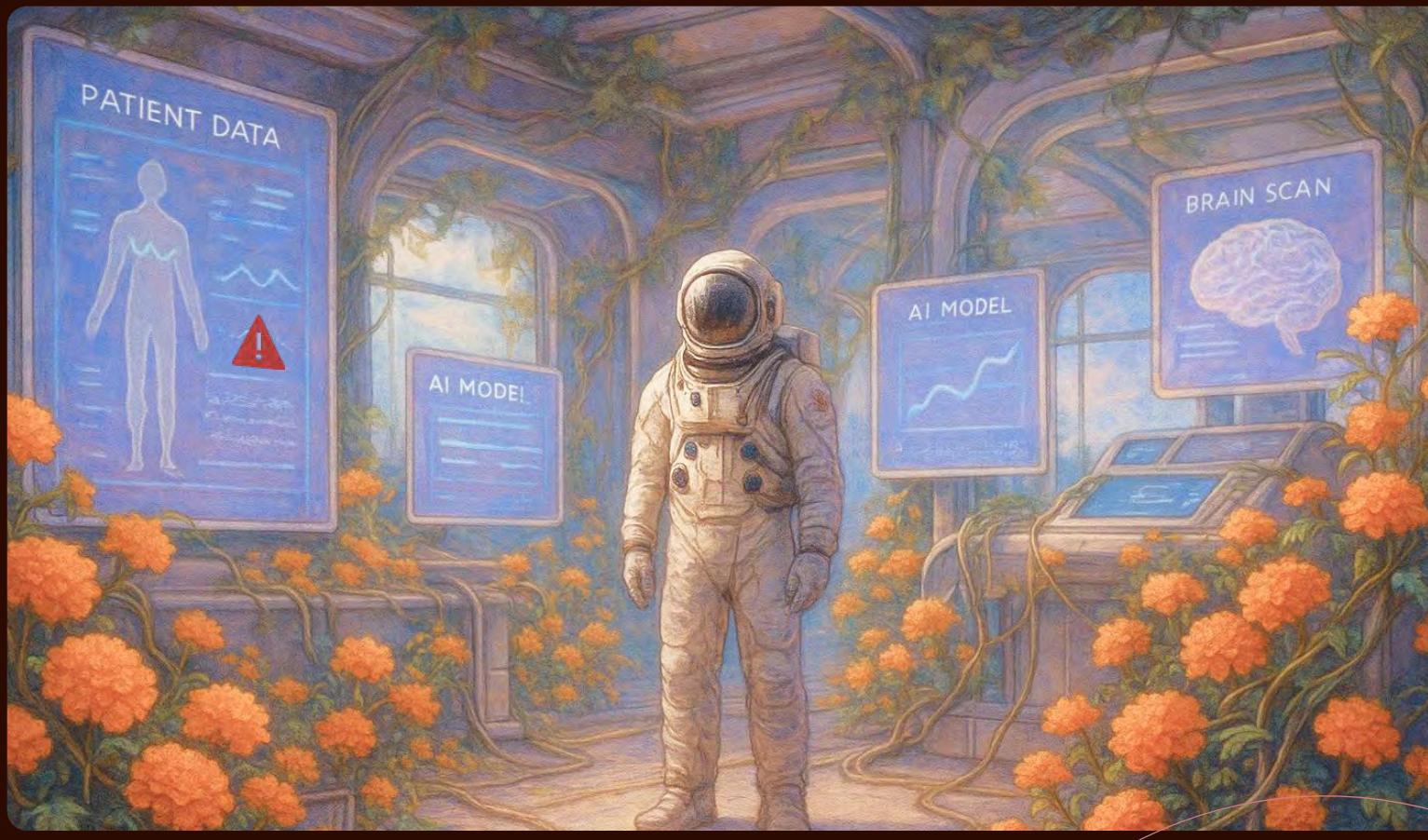
GUIDE

Monitoring AI Agents in Enterprise Healthcare



fiddler

Part 1: The New Reality: From Models to Agents in Healthcare 3

Introduction: The Next Wave of Healthcare Transformation 4

Why AI Observability is Now a Clinical Imperative 4

How External Evaluation Drives Up Your TCO 5

What the Evaluation Trust Tax Looks Like at Scale 5

Part 2: A Real-World Scenario: Monitoring the Multimodal AI Scribe 6

Use Case Deep Dive: The End-to-End AI Scribing Agent 7

Deconstructing the Agent: A Look at the Multimodal Workflow 7

The Unified Monitoring Challenge: Why a Fragmented View Creates Risk 8

Part 3: The Fiddler AI Observability Framework in Action 9

Monitoring the Speech-to-Text Model (Predictive AI) 10

Monitoring the Summarization & Extraction Model (Generative AI) 10

Monitoring the Orchestrator Agent (Agentic AI) 11

Part 4: Building a Culture of Trustworthy AI in Healthcare 12

Beyond the Technology: Governance and Best Practices for Clinical AI 13

Fiddler's Point of View: Observability as the Foundation for Responsible AI 13

Part 5: Powering Clinical Innovation with Trustworthy AI 14

The New Reality: From Models to Agents in Healthcare



Introduction: The Next Wave of Healthcare Transformation

Artificial intelligence is reshaping the healthcare landscape at an unprecedented pace. What was once theoretical is now a clinical and operational reality. Recent data shows that nearly two-thirds of U.S. physicians now use AI tools[1] — a staggering 78% increase from 2023.[2] This is not just an incremental change; it is a seismic shift in how care is delivered, managed, and optimized. And yet the industry continues to blaze ahead.

The next frontier is **Agentic AI**: autonomous systems that orchestrate multiple AI systems and workflows, including traditional machine learning (predictive AI) and large language models (generative AI), to execute complex, multi-step tasks. From automating prior authorizations to summarizing patient histories,[3] these agents promise to unlock immense value.

This guide provides a practical blueprint for navigating this new frontier. It is designed for healthcare leaders who understand that with great power comes great responsibility. Our goal is to demonstrate how a comprehensive AI Observability strategy is the essential foundation for deploying powerful AI agents that are safe, reliable, and fundamentally trustworthy.

Why AI Observability is Now a Clinical Imperative

In healthcare, the stakes are uniquely high. An AI failure can directly impact patient safety, violate foundational privacy regulations like HIPAA, and irrevocably damage the trust between patients and providers. As organizations move toward deploying sophisticated AI agents, the risks multiply across three distinct layers:

Predictive AI: Traditional ML Models

These systems are vulnerable to performance degradation, data drift as patient populations change, and algorithmic bias that can perpetuate and even amplify health inequities.

Generative AI: Large Language Models

LLMs introduce new challenges, including factual inaccuracies (hallucinations) that can lead to dangerous clinical advice, safety issues like leaking Protected Health Information (PHI), and vulnerabilities to prompt injection attacks.

Agentic AI

The agent itself introduces a new "black box" problem. Its internal logic for orchestrating other models can be opaque, creating the risk of cascading failures where a small, undetected error in one component leads to a critical failure in the final output.

For healthcare organizations operating across a complex and often fragmented technology stack (spanning multiple clouds and platforms) achieving a "single pane of glass" for governance is a common and critical goal. A piecemeal approach to monitoring is no longer sufficient.

This is where the Fiddler AI Observability Platform provides an essential solution. It is the infrastructure that manages this complexity, offering a single, unified view to monitor, analyze, and control the entire AI agentic system from end to end.

How External Evaluation Drives Up Your TCO

Every observability platform promises visibility into your agents. But most don't tell you what that visibility actually costs.

When an agent generates a trace, it needs to be scored for quality, safety, and performance. The most common approach sends each trace to an external LLM to act as a judge, and that API call shows up on your foundation model provider, not from your observability vendor. At enterprise scale, this hidden cost adds up fast and directly bloats your total cost of ownership (TCO).

This hidden cost affects you in three critical ways:

The Evaluation Trust Tax

You are charged every time a trace is evaluated via an external LLM call. This shows up on your LLM provider's bill. At enterprise scale, it compounds fast.

Incident Risk Exposure

To control costs, some teams may consider sampling, but the traces you skip could be the ones that matter most. These are low-frequency, high-impact events that sampling can miss.

Operational Overhead

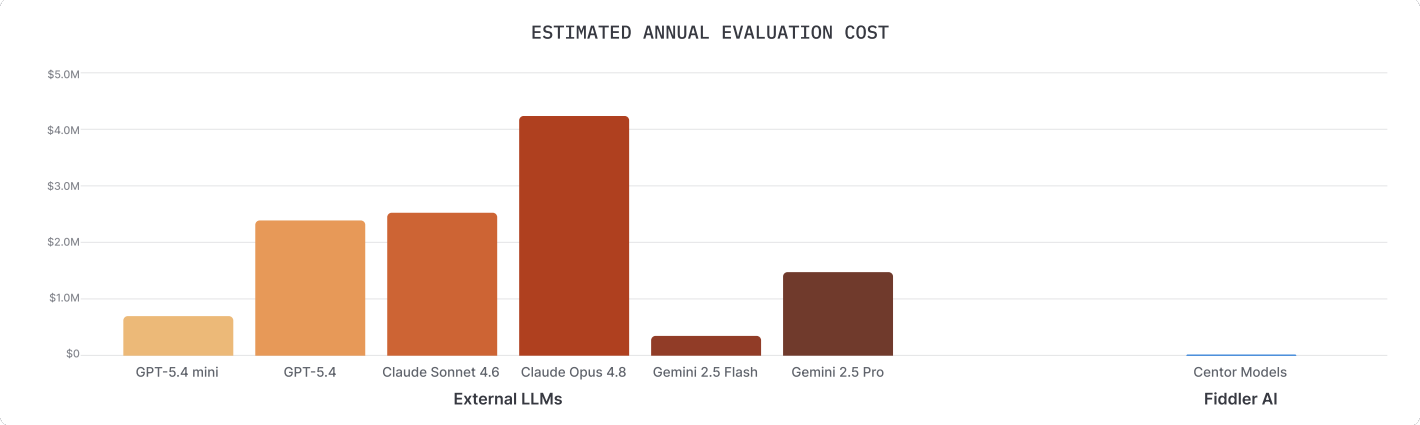
Engineering time and effort should go to building agents, not maintaining evaluation infrastructure: API orchestration, model hosting, prompt versioning, and calibration.

These three problems share a common root: platforms that rely on external LLMs to evaluate your agents.

Fiddler Centor Models are batteries-included. Secure and cost-effective models, for task-specific and complex reasoning with accuracy use cases, run inside your environment, with no external API calls, no data leaving your environment, and no hidden costs as you scale your agents.

What the Evaluation Trust Tax Looks Like at Scale

These are potential external LLM API costs you pay annually, in addition to your tooling fees.

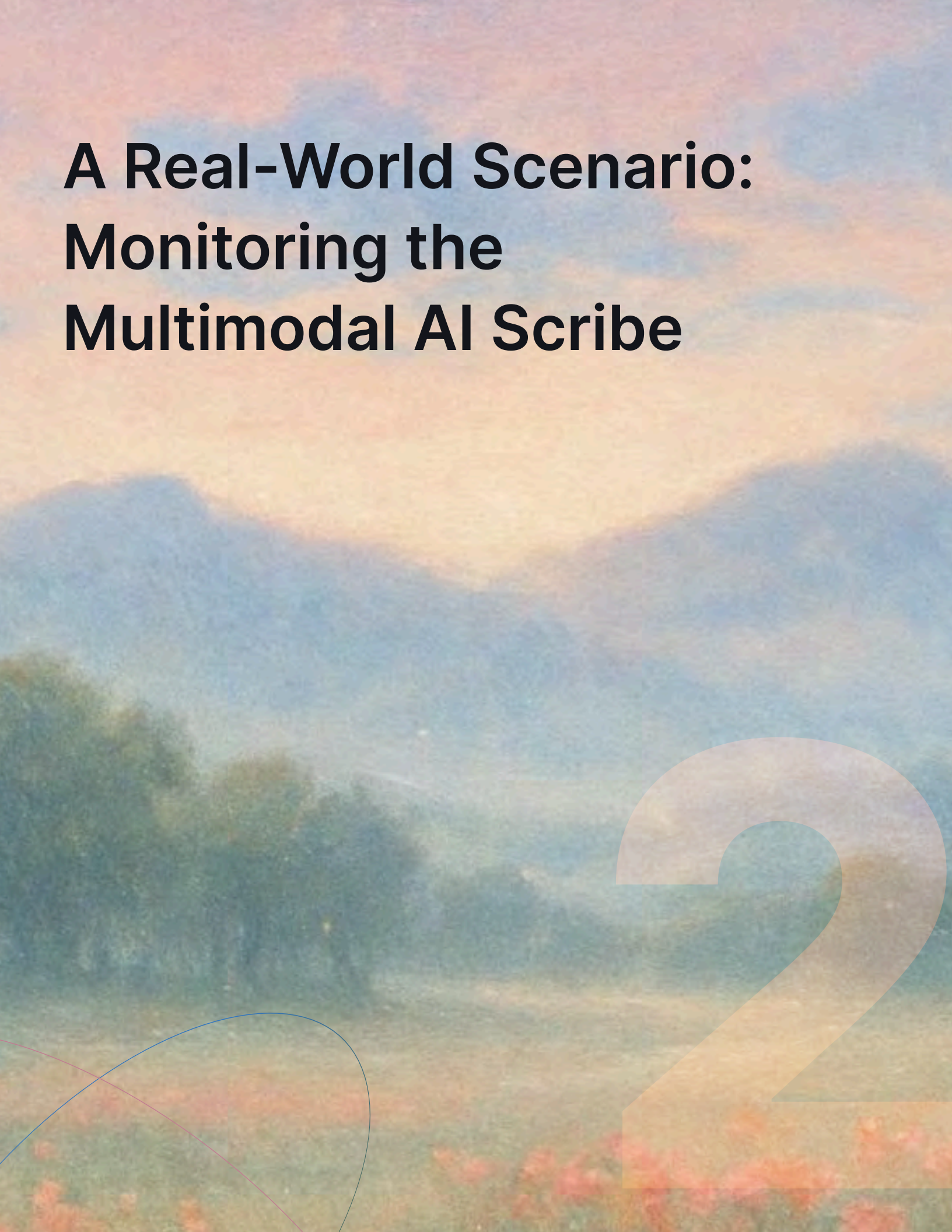


How many tokens are you burning?

Find out what the Evaluation Trust Tax is really costing you, the per-trace API cost charged on every evaluation.

fiddler.ai/evals-tco-calculator ↗

A Real-World Scenario: Monitoring the Multimodal AI Scribe



Use Case Deep Dive: The End-to-End AI Scribing Agent

To make these challenges concrete, let's explore a real-world application that is rapidly gaining traction: the AI-powered ambient scribe. This agent listens to a doctor-patient conversation and autonomously generates a structured clinical note for the Electronic Health Record (EHR).

Leading institutions like **Mass General Brigham** have already deployed AI copilots that reduced physician documentation time by 60%,[4] directly combating the crisis of clinician burnout and allowing doctors to focus more on their patients.

Deconstructing an Agent: A Look at the Multimodal Workflow

An AI scribe is a perfect example of a multimodal agent. Its workflow relies on a chain of specialized AI components working in concert:

PREDICTIVE AI

Audio Ingestion & Transcription

A speech-to-text model first converts the raw audio of the conversation into a written transcript. This is a classic machine learning task.

GENERATIVE AI

Clinical Information Extraction

A large language model then analyzes this transcript. It identifies and summarizes the critical clinical elements: patient symptoms, the physician's diagnosis, prescribed medications, and necessary follow-up actions.

AGENTIC AI

Workflow Orchestration

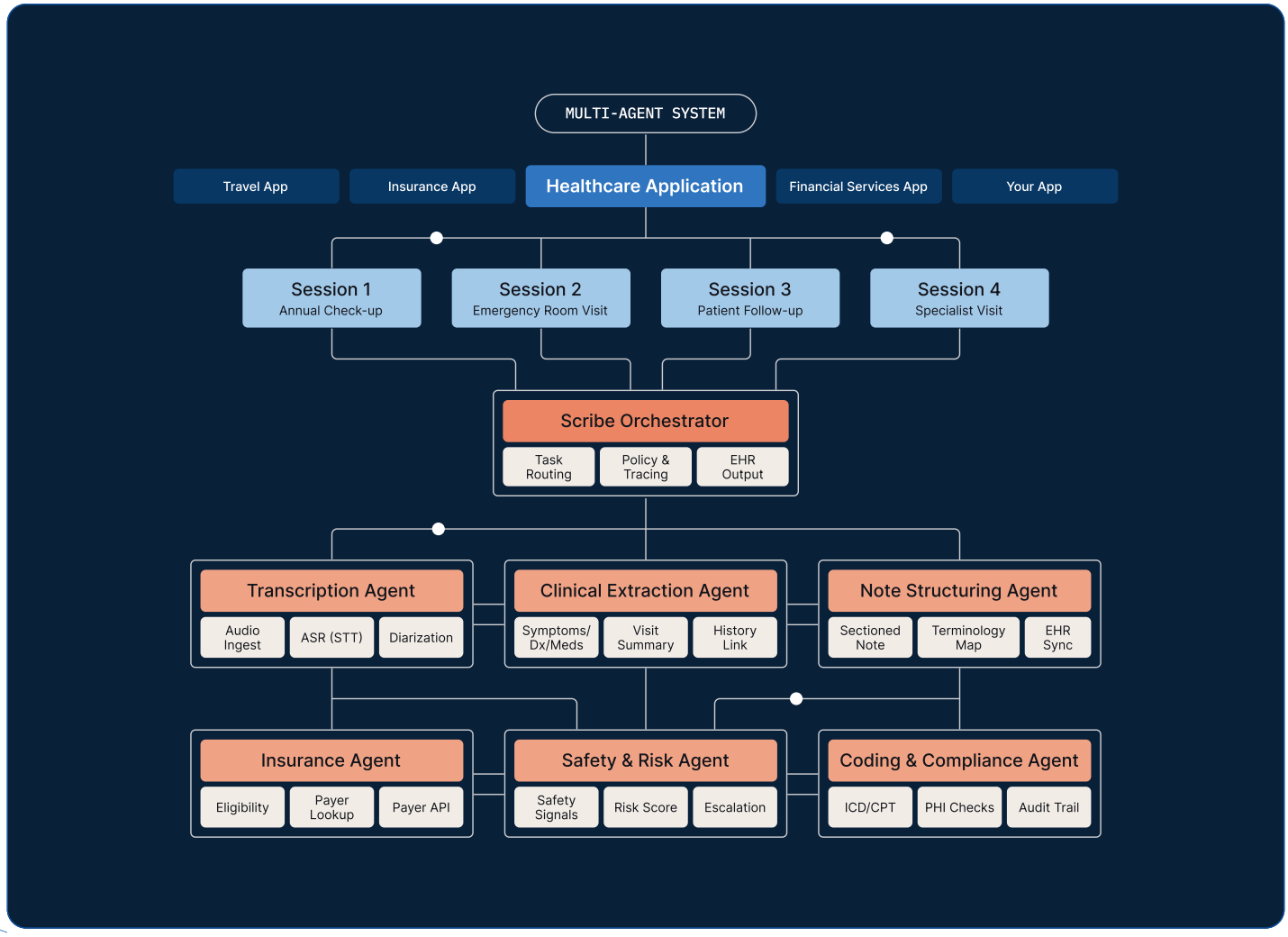
Finally, the orchestrating agent takes the extracted information and performs a series of tasks. It formats the summary into a compliant EHR note, may suggest relevant medical billing codes based on the diagnosis, and flags action items for the clinical staff, completing the end-to-end workflow.

The Unified Monitoring Challenge: Why a Fragmented View Creates Risk

The power of this agent lies in its interconnected workflow, but so does its greatest risk. A subtle error at any stage can create a domino effect. For instance:

1. A minor inaccuracy in the initial **ML transcription** (e.g., mistaking "hyperthyroidism" for "hypothyroidism")...
2. ...could be misinterpreted and amplified by the **LLM summarization**...
3. ...leading the **agent** to generate a dangerously incorrect patient report, potentially resulting in the wrong prescription or follow-up plan.

To ensure patient safety and clinical reliability, healthcare organizations cannot afford to monitor these systems in silos. They require an end-to-end observability strategy that provides a holistic view, capable of tracing an outcome back through the entire agentic workflow/hierarchy to its root cause.



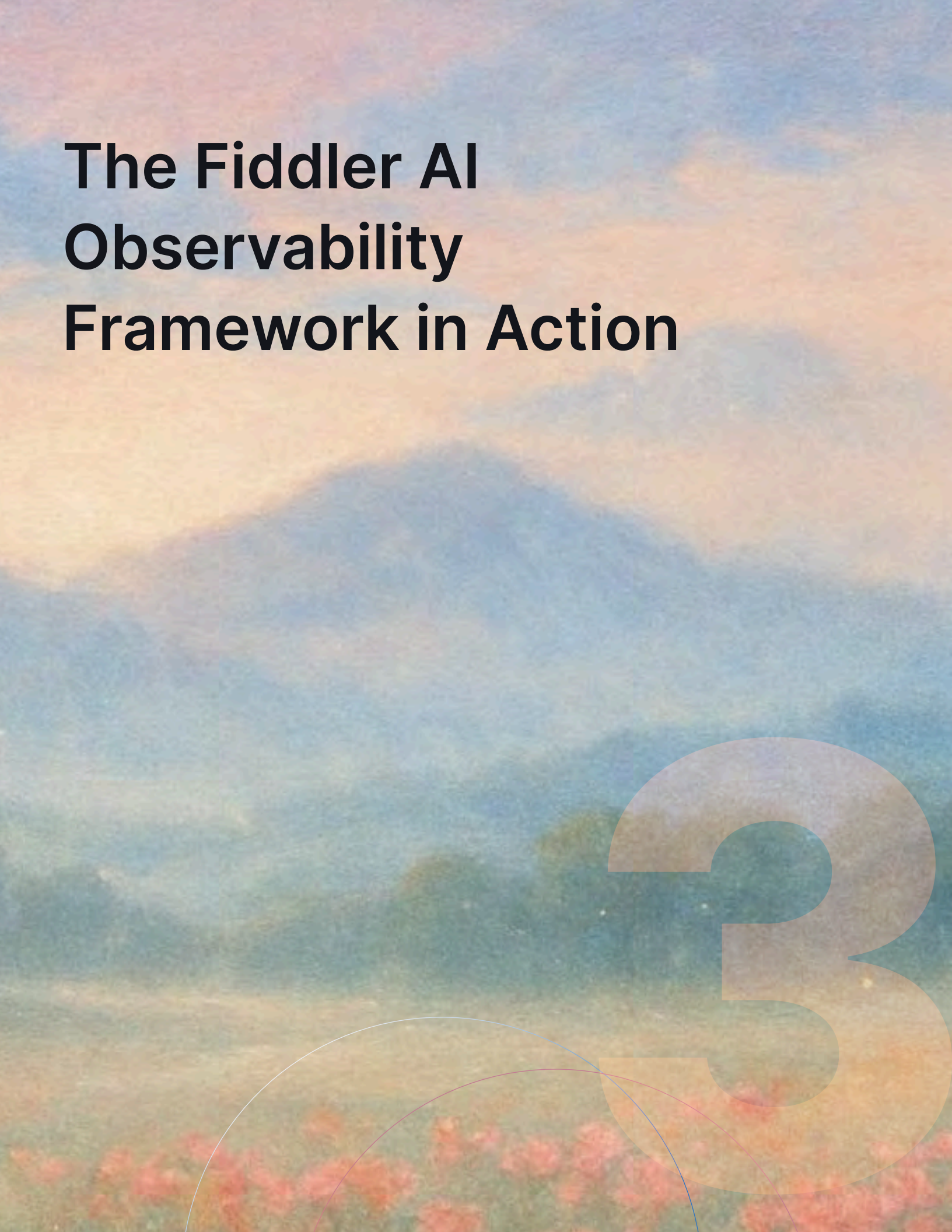
* **ASR/STT:** Automated Speech Recognition/Speech to Text

Diarization: Automatic Speaker Identification

ICD/CPT: Current Procedural Terminology

EHR: Electronic Health Record

The Fiddler AI Observability Framework in Action



Fiddler provides this unified view, offering tailored monitoring capabilities for each component of the AI scribe agent. This approach allows organizations to manage specific risks at each stage while maintaining a complete picture of the agent's overall health.

PREDICTIVE AI Monitoring the Speech-to-Text Model

Risks	Fiddler Predictive AI Observability
⚠ Performance degrades due to accents, audio quality, or changing terminology.	☑ **Performance & Drift Monitoring:** Track accuracy (e.g., Word Error Rate) and monitor data shifts to get alerts on performance decay.
⚠ Model performs less accurately for certain demographics, creating systemic bias.	☑ **Fairness Analysis:** Analyze performance across patient subgroups to detect and mitigate bias.
⚠ Transcription errors are difficult to diagnose, slowing down resolution.	☑ **Explainability & Root Cause Analysis:** Use explainability tools to find the root cause of errors for faster resolution.

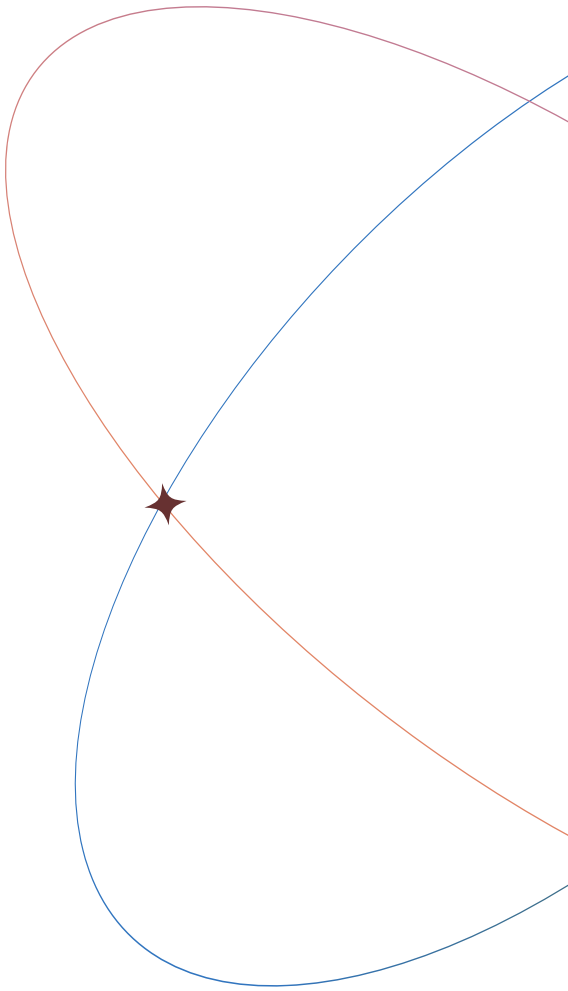
GENERATIVE AI Monitoring the Summarization & Extraction Model

Risks	Fiddler Generative AI Observability
⚠ Model may generate harmful, biased, or off-topic responses that violate clinical or brand safety policies.	☑ **Real-time Guardrails:** Implement real-time guardrails to prevent the model from generating unsafe or inappropriate content, ensuring responses adhere to predefined clinical and ethical standards.
⚠ Model may invent clinical details (hallucinations) or provide factually incorrect information.	☑ **Response Quality Monitoring:** Use custom metrics to evaluate factual accuracy and relevance, with alerts for high hallucination rates.
⚠ Model may leak Protected Health Information (PHI), creating compliance and privacy risks.	☑ **Safety Monitoring & Alerts:** Set up instant alerts for safety violations like PHI detection or toxic language.
⚠ Unmonitored models can lead to high costs and latency, hindering scale.	☑ **Cost & Latency Tracking:** Monitor token usage and response times to manage operational costs and performance.
⚠ Difficult to identify prompts that cause poor or unsafe model behavior.	☑ **Prompt & Response Analysis:** Log and analyze prompt-response pairs to find patterns that cause unsafe outputs.

AGENTIC AI

Monitoring the Orchestrator Agent

Risks	Fiddler Agentic AI Observability
⚠ The agent's opaque decision-making process ("black box") hinders trust, debugging, and clinician adoption.	☑ **Agent Transparency:** Visualize the agent's chain of thought and tool usage to enable debugging, auditing, and clinician trust.
⚠ Agents can amplify upstream errors from ML or LLM systems, resulting in a critically flawed output.	☑ **End-to-End Tracing & Root Cause Analysis:** Get a unified trace of the agent's entire workflow. Instantly find the source of any error, from any component.
⚠ Lack of a unified view makes it difficult for governance teams to monitor the complete system's health.	☑ **Unified Dashboard:** Monitor the ML model, LLM, and agent logic in a single dashboard, providing a "single pane of glass" for governance.



Building a Culture of Trustworthy AI in Healthcare

Beyond the Technology: Governance and Best Practices for Clinical AI

Technology alone is not enough. Building a trustworthy AI program requires a culture of responsibility and a robust governance framework.

The Human-in-the-Loop

Successful AI deployment depends on a multidisciplinary governance team. This team should include clinicians who understand the context of use, data scientists, IT leaders, and compliance officers who can oversee AI performance and make informed decisions about its application.

Regulatory Alignment

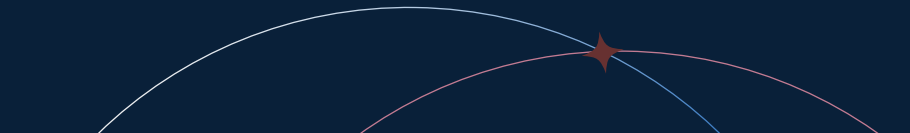
The practice of continuous monitoring is not just a best practice; it is increasingly becoming a regulatory expectation. The FDA's forward-thinking frameworks, such as the Total Product Life Cycle (TPLC)[5] approach and Predetermined Change Control Plans (PCCP),[6] explicitly acknowledge that AI models are not static, and recommend ongoing oversight for AI-enabled medical devices.[7]

A comprehensive observability platform provides the evidence and audit trails necessary to meet these evolving standards.

Fiddler's Point of View: Observability as the Foundation for Responsible AI

At Fiddler, we believe AI observability is more than a set of technical tools — it is a strategic framework for risk management, regulatory compliance, and building institutional trust. As healthcare organizations like yours stand on the cusp of the agentic AI revolution, the ability to see, understand, and govern these complex systems is paramount.

AI Observability is the key to achieving that transparency at scale.



Powering Clinical Innovation with Trustworthy AI



fiddler

Fiddler is the AI Control Plane, the system of trust, for first-party and third-party agents from the creation layer to production. Continuous evaluation, reliable monitoring, enforceable policy, and auditable governance give enterprises the centralized controls and actionable insights to scale AI with trust. Integral to the platform are the secure Fiddler Centor Models, which power the industry's fastest guardrails and evaluations that require low latency and cost-effectiveness or complex reasoning with accuracy. Compared to external LLMs, Fiddler Centor Models offer a low TCO with no hidden costs.

Fortune 100 organizations use Fiddler to deliver high performance agentic and predictive applications, protect from costly risks, and maximize ROI.

fiddler.ai sales@fiddler.ai

The transition to sophisticated, multimodal AI agents represents a pivotal moment for healthcare, offering the potential to dramatically reduce administrative burdens, accelerate diagnoses, and personalize patient care. To capitalize on this opportunity, organizations must ensure their AI initiatives can deliver **high performance, avoid costly risks, and maximize ROI.**

A comprehensive AI observability strategy is the essential framework for achieving these goals. It provides the continuous, real-time insights needed to turn the "black box" transparent, manage the inevitable risks of bias and drift, and build trust with clinicians, patients, and regulators.

By embedding observability into the AI lifecycle, healthcare organizations can ensure their models operate at peak effectiveness, prevent the clinical and financial damage of unforeseen failures, and ultimately, secure the highest possible return on their strategic AI investments. For any organization looking to lead in the new era of AI, establishing a unified observability practice is the most critical step toward ensuring long-term success, safety, and value.

Ready to deploy AI agents with confidence?

Explore how the Fiddler AI Observability Platform provides the unified monitoring, protection, and control needed to build the future of healthcare.

Request a demo: **fiddler.ai/demo** ↗

References:

1. DemandSage. (2025, June 8). AI in healthcare: statistics and insights. DemandSage. https://www.demandsage.com/ai-in-healthcare-stats/
2. Strategic Market Research. (2025, September). AI in healthcare statistics: latest data & facts. Strategic Market Research. https://www.strategicmarketresearch.com/blogs/ai-in-healthcare-statistics
3. World Wide Technology. (2025, July 1). Agentic AI: Strategic value and high-impact use cases for healthcare systems. WWT. https://www.wwt.com/blog/agentic-ai-strategic-value-and-high-impact-use-cases-for-healthcare-systems
4. Mass General Brigham. (2025, February 17). AI screening for clinical trials speeds up enrollment, study finds. Mass General Brigham. https://www.massgeneralbrigham.org/en/about/newsroom/press-releases/ai-screening-for-clinical-trial-speeds-up-enrollment
5. U.S. Food & Drug Administration. (2023, September 6). Total product life cycle for medical devices. FDA. https://www.fda.gov/about-fda/cdrh-transparency/total-product-life-cycle-medical-devices
6. U.S. Food & Drug Administration. (2024, August 4). Predetermined change control plans for medical devices. FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-medical-devices
7. U.S. Food & Drug Administration. (2025, January 7). Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations. FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing

