SIGNAL 1

The New Economics of AI: Cheaper to Build, Costlier to Run

Why agentic AI changes how health systems need to measure and manage value

For years, the economics of enterprise AI programs looked a lot like the economics of traditional software development. Large fixed costs. Low marginal costs.

Consider a predictive sepsis model. Building it requires specialized technical talent, data preparation, model training and validation, integration, and deployment. But once it is running, generating another prediction costs relatively little, and expanding usage improves the economics by spreading the initial investment across more patients and facilities.

These economics shaped how health systems thought about internal AI programs. The critical decision typically happened before deployment: Is this problem large enough and valuable enough to justify the upfront investment required to solve it? After that, there was little reason not to keep a useful model running and even expand its use.

Agentic AI changes this equation. Applications built on foundation models can be significantly less expensive to develop, but they carry a meaningful marginal cost every time they run. Every action an agent takes consumes resources, whether it's summarizing a chart, drafting a message, calling another system or executing a workflow. A single task can trigger dozens of model calls.

Purchased AI is following suit, as vendors move from per-seat licenses to usage-based pricing that passes these costs through to health systems.

This change introduces a fundamentally different set of economic considerations:

> The defining AI decision is moving from “Should we build this?”, made once before deployment, to “Should we keep running it, and where?”, made continuously after deployment.

Agentic AI shifts the cost from building to running

	Traditional AI	Agentic AI
Cost to build	**High and fixed** talent, data, training, deployment	**Significantly lower** built on foundation models
Cost per use	**Relatively little** more use spreads the fixed cost	**Meaningful, on every run** model calls + clinician review
Scaling up	**Improves the economics** more patients, more facilities	**Depends on the setting** each new use adds cost
The key decision	**"Should we build this?"** made once, before deployment	**"Keep running it, and where?"** made continuously, after deployment

Three Implications for Health Systems

1. Cheaper to build means more to shut down

Historically, high development costs forced organizations to be selective about which AI projects they pursued. Once deployed, however, those costs were sunk and the marginal cost of continued use was generally low. Even solutions delivering modest value could remain economically viable, reducing the need to continuously measure their return.

Lower development costs make it easier to build applications, but every use now generates a cost. An agent that performs well technically may still fail to generate enough downstream value to justify its continued operation.

> Once an agent is built, the question that matters is whether each use creates more value than it costs.

This creates a new challenge: the realized benefits and costs of an agent may not be fully understood until it is operating in a real-world environment. Actual usage, costs and outcomes can differ substantially from initial projections.

Health systems will therefore need more than modeled business cases. They need to be able to continuously measure the costs and benefits of deployed agents. Acting on those measurements will require organizational discipline and a willingness to shut down agents that may be functioning as intended but aren't generating enough value to justify their ongoing costs.

2. Scaling AI is no longer always the obvious decision

With earlier generations of AI, scaling generally improved the return on investment, so moving successful pilots enterprise-wide became a central goal.

With agents, however, every expansion to additional users, departments or hospitals adds operating costs. Yet there is no guarantee that the agent delivers sufficient value in the new setting. When each use cost almost nothing, any setting where a model created value was worth adding. Now each use has a cost, and the value an agent creates varies from one setting to the next.

Consider an AI agent that manages routine patient portal requests, reviewing the message, retrieving relevant chart information, determining the appropriate next step and preparing a response for clinician approval. In a high-volume primary care practice, it may resolve routine requests efficiently and save substantial staff time. In a specialty clinic, however, more complex cases may require additional model calls, chart review and clinician intervention. The same agent could cost more to operate while delivering less value per request.

Under the changing economics of AI, a successful pilot does not necessarily justify enterprise-wide expansion. Each new setting introduces additional per-use costs that must be weighed against the value created by the agent in that setting. Health systems need to measure agent performance and economics in each environment to see where expansion justifies the cost and where it does not.

3. Lowering the cost of running AI becomes a strategic investment

Health systems have long invested in infrastructure that makes AI easier and less expensive to develop and deploy, including data lakes and MLOps environments. Now, agentic AI creates a new priority: applying FinOps principles to measure, manage and reduce the cost of running AI.

Organizations can lower these costs by routing tasks to less expensive models when appropriate, eliminating unnecessary model calls, and optimizing prompts and workflows. While savings from individual interactions may be small, they add up across millions of agent actions.

Lower operating costs can also make new use cases economically viable. An agent that costs more to run than the value it generates today may become worth deploying as its cost per task falls.

> At enterprise scale, even small reductions in operating costs can significantly improve the return on generative and agentic AI.

Managing AI for Economic Value

Together, these changes necessitate a new approach to enterprise AI management, combining technology that continuously measures cost and value with clear accountability for decisions about continued use, expansion and cost optimization.

Signal 1's AI Management System (AIMS) provides the technology foundation, with visibility into usage, performance, cost and value across the AI portfolio. Through its Agent Control Plane, Signal 1 is extending these capabilities to help health systems evaluate agent economics, optimize model selection and identify opportunities to reduce token costs.

As health systems deploy more agents, the financial implications of these decisions will grow. Continuous measurement, combined with the mandate and willingness to act, will be essential to ensuring AI investments deliver value as they scale.