

Observability: AI Is Eating the Budget That Keeps AI Safe

Monitoring Budgets Are Partly Funding AI Cost Overruns, so Buyers Must Ensure AI Observability Reveals the Full Cost of Approved Work

Analyst(s): Vikram Rathnam, Mitch Ashley

Publication Date: October 27, 2026

Document #: AIOVRMA202610

What You Need to Know

- **Judge AI Observability by Cost per Accepted Result:** By the end of 2027, we expect buyers running AI agents in production to judge AI observability on cost per accepted result, including the human time spent reviewing and fixing agent work, rather than cost per token. That cost covers every attempt, including failed ones.
- **More Teams Cap Agent Spending Than Check Agent Work:** Spending limits per agent or session are in place in 32% of organizations. Only 12% evaluate agent work before it is promoted to the next stage.
- **AI Overruns Are Reaching Monitoring Budgets:** In the past six months, AI spending exceeded plans for 52% of organizations. Among those covering it from other software budgets, 44% took money from IT operations and monitoring tools.
- **Observability Gets Many Small Raises, Few Big Ones:** In the next 12 months, 64% of buyers plan to spend more on observability, but only 19% plan a large increase, the lowest share of the five core Software Lifecycle Engineering (SLE) areas.

The Futurum View

By the end of 2027, we expect buyers running agents in production to judge AI observability on cost per accepted result: the full cost of every attempt, including failures and human fixes, divided by the work that passes review. Counting tokens shows what AI agents consume. It does not show whether their work was worth doing or whether it worked. AI observability means seeing how AI applications and agents actually run: the models and tools they call, the retries they make, and the cost of each step. Right now, many organizations are paying for AI overruns partly out of monitoring budgets, just as they let agents do more. That trade is hard to sustain. Researchers have begun pricing AI by the cost per correct result rather than per token, as in the [Cost-of-Pass framework](#). Futurum's view adds the human time spent reviewing and fixing agent work. The evidence comes from Futurum Research and Futurum Intelligence survey data, FinOps Foundation guidance, and academic research on agent cost and quality.

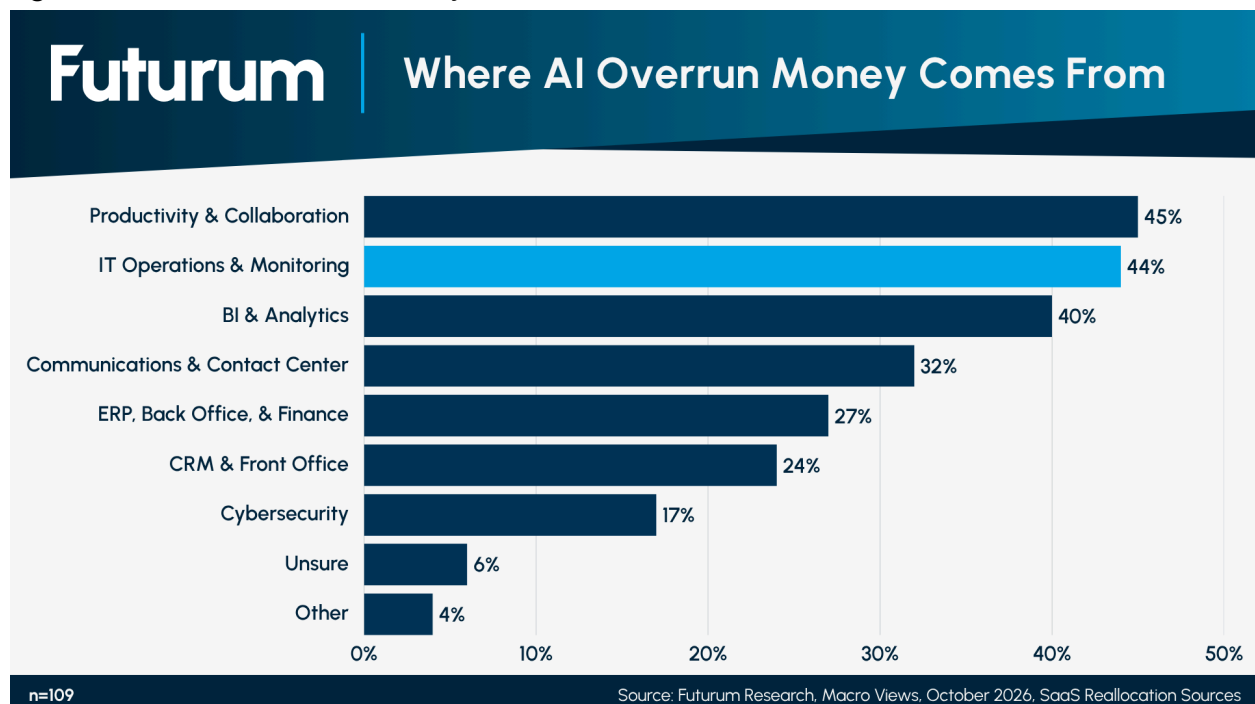
The scope is AI applications and agents through 2027, without vendor rankings or market sizing.

AI Overruns Are Drawing on Existing Budgets

AI spending is outpacing plans. In Futurum Intelligence's October 2026 Macro Views survey, 52% of 1,684 respondents said AI spending was above plan over the past six months. Of the 875 who said how they covered the gap, 45% asked for more money, 40% moved money from another part of IT, and 38% pushed the cost into next year's planning cycle. Respondents could pick more than one.

Monitoring tools are one source of money. Among the 109 organizations that took money from other software budgets, 44% took it from IT operations and monitoring tools, close to the 45% that took it from productivity and collaboration software. At the same time, AI raises the cost of observability itself. The [FinOps Foundation](#) says observability and telemetry spend is a common driver of AI budget overruns. Observability is being squeezed from both sides. Observability budgets are still growing overall, but they now have to cover AI workloads on top of everything else.

Figure 1: Where AI Overrun Money Comes From



Source: Futurum Research, Macro Views, October 2026, SaaS Reallocation Sources. Base: respondents moving money from SaaS or software budgets to cover AI overruns, N=109. Respondents could select more than one source. The survey did not measure how much money moved or whether teams lost visibility. SaaS is software as a service, BI is business intelligence, ERP is enterprise resource planning, and CRM is customer relationship management.

This is a bad moment to give up visibility. In Futurum’s 2H 2026 SLE Decision Maker Survey, 75% of 839 decision makers reported at least one production incident in the past year involving AI-generated code, agent actions, or AI tools. Yet buyers are adding to observability only in small steps. Of the same 839, 64% plan to raise observability spending in the next 12 months, but only 19% plan an increase of more than 15%, the lowest share among the five core SLE areas.

Table 1: AI Funding Choices, Observability Investment, and Agent Controls

Finding	Result	What Was Measured	Source and Response Base
AI spending is above plan	52%	Moderately or substantially above plan over the preceding six months	Futurum Research Macro Views, October 2026; N=1,684
Monitoring software as an overrun funding source	44%	Respondents reallocating SaaS/software budgets to cover AI overruns	Futurum Research Macro Views, October 2026; subgroup N=109
AI-contributed production incidents	75%	At least one incident in the prior 12 months involving AI-generated code, agent actions, or AI tooling	Futurum Research 2H 2026 SLE survey, Q38; N=839
Agent evaluation before promotion	12%	Evaluation/quality scoring before promotion is selected as an agent governance control	Futurum Research 2H 2026 SLE survey, Q37; N=839
Planned observability increases	64%/19%	Significant or slight increase/significant increase (>15%) over the next 12 months	Futurum Research 2H 2026 SLE survey, Q21; N=839

Source: Futurum Research Macro Views, October 2026, AI Spend vs. Budget, Actions for AI Budget Overruns, Budget Reallocation Sources, and SaaS Reallocation Sources dashboards, all filters set to All. Futurum Research, 2H 2026 SLE Decision Maker Survey (2H26DMSLEMA202607), July 2026, pp. 6, 8–9; fielded June 2026. Macro Views response bases: 1,684 (AI spend vs. budget), 875 (actions for overruns), 350 (reallocation within IT), and 109 (SaaS reallocation). These figures come from separate surveys and are not combined.

The overruns will also raise next year’s budget. Futurum’s 2H 2026 CIO & Technology Buyers survey found the same pattern: 43% of over-budget enterprises absorbed the overrun into the next cycle. Some of the cuts to monitoring may only remove duplicate tools. Buyers should check what visibility they still have after each cut.

Connect Agent Costs With Accepted Results

AI cost is becoming part of observability. The same traces that show what an agent did can show what each step cost. A single agent request can trigger multiple model calls, tool calls, retries, and human corrections. Judging that work takes four functions, which one product may combine: observability records what the agent did and what it cost; evaluation checks the result against what the task required; financial analysis adds up the full cost and value; and execution controls set spending limits and approval rules. Observability matters most here because it holds the trace: the record of each attempt and what it cost. Once evaluation results and review outcomes are added to that trace, teams can calculate the cost per accepted result task by task, which a finance report of total spend cannot show.

Human review is the cost that token-based measures leave out. In [a study of 567 pull requests](#) created by a single AI coding agent, 84% were merged, but 45% of the merged changes still required further human revision. Token prices can mislead in the other direction, too. The [Cost-of-Pass researchers](#) found reasoning models were the most cost-effective option for complex quantitative problems, despite higher per-token costs.

We recommend measuring cost per accepted result. Take one type of task over a set period. Add up the cost of every attempt, including model, tool, and infrastructure costs, as well as the human time spent on reviewing and fixing. Divide by the number of results that met the agreed standard. For example, if an agent makes 100 bug-fix attempts and 60 pass review, the cost per accepted fix is the total cost of all 100 attempts, plus the engineering time spent on review and fixing, divided by 60. A cheaper model that fails more often can worsen this number. The measure also gives teams a practical way to pay down what Futurum tracks as verification debt: AI-generated work piling up faster than people can check it.

The [FinOps Foundation](#) draws the same line between resource measures, such as cost per token, and business measures, such as cost per case resolved. Observability can supply the evidence. Product owners and finance still have to agree on what counts as an accepted result.

AI Budgets and Outcomes Need a Single Owner

Business units are paying part of the AI bill, too. In Futurum Intelligence's survey, 24% of the 875 organizations covering overruns moved money from a non-IT business-unit budget. Futurum's 2H 2026 CIO & Technology Buyers survey found a similar share: 23% of over-budget enterprises shifted spending to a non-IT business-unit budget. Our view is that once a business unit pays part of the AI bill, CIOs will find it hard to take that money back.

We expect 2027 planning to bring this to light. The team paying for model use, the team running the agent, and the team responsible for the outcome often sit on different budgets. Someone has to own the total cost and agree on what good work looks like. A bigger observability budget alone will not settle that.

Three groups need to act:

- **CIOs and Finance:** Agree with business units on who pays for model use, observability, evaluation, and human review. List the visibility you must keep before moving money out of software budgets.
- **Product and Engineering Teams:** Define what an accepted result means for each type of task, and track how much time people spend reviewing and fixing agent work.
- **Observability and Evaluation Vendors:** Show one task's steps, cost, and quality result together. Then show how that evidence changes model choice, workflow design, spending limits, and when to bring in a person.

What to Watch

- **Cost and Quality in Buying Criteria:** Watch for requests for proposal (RFPs) and renewals that ask vendors to show cost per accepted result. If buyers continue to buy mainly based on token and usage dashboards, our view is wrong.
- **Where the Measure Lives:** Watch whether the cost per accepted result shows up in observability platforms or in FinOps and AI cost-management tools. If it settles in finance tools instead, our forecast is wrong.
- **Budget Ownership Agreements:** Watch for CIOs, finance, and business units to agree in writing on who pays for agent use, observability, and evaluation. Clear ownership and a few budget fights would weaken our expectation of friction in 2027.
- **Evaluation Connected to Spending Controls:** Compare teams that use evaluation results to set model choice, retry limits, and human approvals with teams that rely on spending caps alone.
- **Measured Results From Model Changes:** Look for reports that compare task success, total cost, speed, and human fixing time after a model change. A cheaper model that needs more retries or rework can raise the cost of an accepted result.

Other Insights from Futurum

[46.9% of Enterprises Report AI Spend Over Budget in 2H 2026](#)

[Software Lifecycle Engineering Market to Reach \\$226 Billion by 2030 as Enterprise Hands AI the Keyboard Before the Guardrails](#)

[Harness Shows Why LLM Security Scanning Needs a Repeatable Verdict](#)

About Us

About the Authors

Vikram Rathnam is Research Director, Software Lifecycle Engineering at Futurum. His research examines how software is built, tested, secured, and operated as AI agents move from assisting developers to doing the work. His coverage includes observability, platform engineering, and the control planes that govern agents. With 24 years of experience on the vendor side, including roles at Cisco, Secureworks, Dell, and CMIT Solutions. He can be reached at vrathnam@futurumgroup.com.

Mitch Ashley is VP & Practice Lead of CIO & Technology Buyers and Software Lifecycle Engineering for The Futurum Group. A multi-time CIO and CTO with 30+ years leading technical organizations, Mitch built and operated production systems spanning cybersecurity for the U.S. Department of Defense, PKI services for the broadband and 5G industries, SaaS platforms, large-scale telecom and banking systems, and a national broadband network. He can be reached at mashley@futurumgroup.com.

About The Futurum Group

Every day, The Futurum Group's analysts, researchers, and advisors help business leaders worldwide anticipate tectonic shifts in their industries and leverage disruptive innovation. Unlike traditional analysts, The Futurum Group works not only in analysis and research but also takes that insight and knowledge even further, engaging all the way through the go-to-market process.

Futurum Research provides in-depth research and insights on global technology markets using advisory services, custom research reports, strategic consulting engagements, digital events, go-to-market planning, and message testing. It also creates, distributes, and amplifies rich media content that all stakeholders read, watch, and listen to.

See more details on The Futurum Group at futurumgroup.com.

Copyright & Use License

Copyright Notice

Copyright ©2026 by The Futurum Group, LLC. All rights, including that of translation into other languages, are specifically reserved. No part of this publication may be reproduced in any form, stored in a retrieval system, or transmitted by any method or means, electrical, mechanical, photographic, or otherwise, without the express written permission of The Futurum Group futurumgroup.com. United States copyright laws and international treaties protect this publication. Unauthorized distribution or reproduction of this publication, or any portion of it, may result in severe civil and criminal penalties and will be prosecuted to the maximum extent necessary to protect the publisher's rights.

License Notice

This document may be distributed within the licensed organization only.

The following acts are prohibited:

Transmittal to others outside your immediate organization, including partners, resellers, external consultants, etc., in any media format

Posting on a website that is accessible to others outside your immediate organization

The possession or use within an unlicensed organization

Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. The Futurum Group shall have no liability for errors, omissions, or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve their intended results. The opinions expressed herein are subject to change without notice.