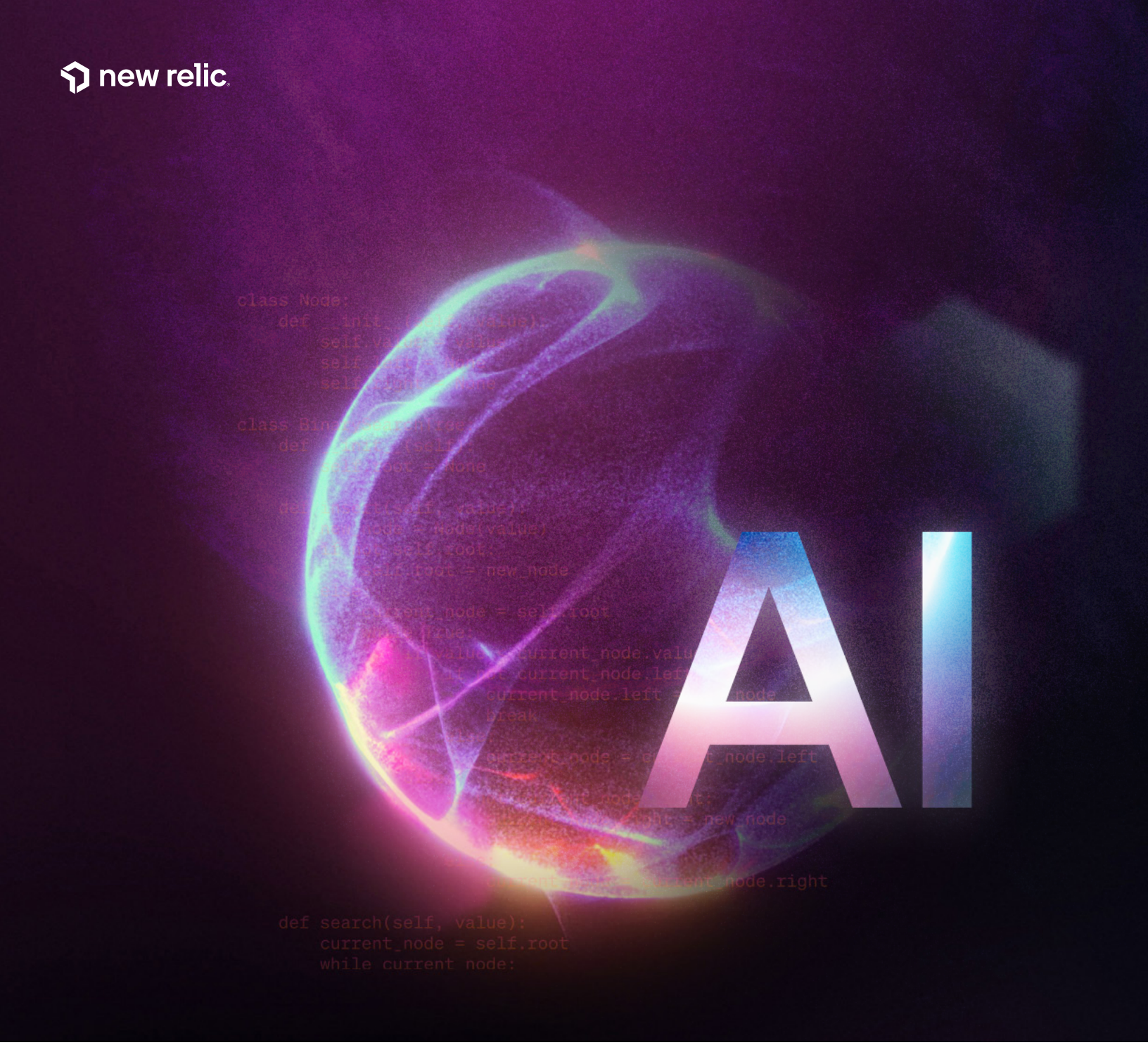
new relic



Advanced AI Strategies Demand Full-Stack AI Observability

Massive investments in AI require robust monitoring of performance, risk, and results

Brief / December 2024

The challenge of AI visibility

Leaders in every industry are seeing the immediate benefits of generative AI, starting with improved customer service, content generation, personalized recommendations, assistants, smart chatbots, risk/fraud prediction and detection, and process automation.

But with new capabilities come new risks, new challenges—and new opportunities.

For companies making big bets on AI, it's essential to understand all the components that go into their AI solutions. The stakes are high because AI has a potentially enormous impact on every area that executives and boards care about: customer experience, revenue, costs, cybersecurity, and even brand perception, to name a few. It's not just IT leaders and developers, but even executives need to know how their AI models are performing, their associated costs, potential vulnerabilities, and how they interact with one another.

That might sound like a straightforward request, but with today's increasingly complex AI tech stack, it's a tall order.

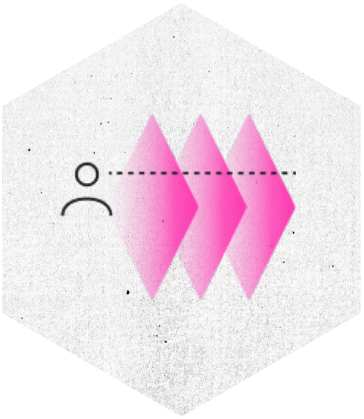
AI solutions introduce new frameworks and components. Yes, there are large language models (LLMs)—but also advanced data stores, data pipelines, orchestration frameworks, and machine learning (ML) code libraries. For many organizations, troubleshooting these systems is a whole new ball game.

Even monitoring their routine behavior can be tricky, particularly when using tools that were built before the advent of AI use cases. Each system also comes with its own unique API, its own way of reporting data, and its own interface—which means monitoring tools need to be tailored to each model.

> "Just about every organization is integrating AI applications into their tech stacks to provide better customer experiences and improve efficiency, with the hope of improving their bottom line. The trade-off is that AI creates complexity in their tech stack and must be adopted responsibly with security, quality, compliance, and cost top of mind."
>
> **Stephen Elliot**
> IDC Group Vice President

Without reliable data on the performance and business benefits of these AI systems, companies will struggle to make informed decisions about new AI investments or tie those investments to business outcomes and value.

As long as they lack access to reliable data on AI performance and benefits, companies will struggle to make informed decisions.



Managing risks, driving ROI

Return on investment (ROI) is another pressing concern for AI leaders. Enterprise AI investments are transforming not only companies but entire industries. For example, according to a recent estimate from Sequoia Capital, companies worldwide are spending so much on Nvidia AI infrastructure in 2024 alone that they'll need to generate $600 billion in lifetime AI-related revenue to justify the spend. And that's just one year's worth of investments!

While most companies aren't making such eye-popping investments in AI (yet), demonstrating ROI will be crucial for all organizations, particularly as the AI hype subsides and C-suite executives and their boards start to demand real results from this new technology. Yet many leaders are unaware of all the components and costs that go into their AI stacks, which can make accounting a challenge.

In order to justify the spend, AI models must deliver measurable value—and drive revenue that aligns with the investment.

The solution: Full-stack AI monitoring

The right AI monitoring solutions should give CTOs and CIOs the data they need to make timely decisions, limit expenses, and measure and maximize ROI. With the right solutions in place, leaders can better ensure reliability, quality, and efficiency throughout all components of the AI technology stack, alongside services and infrastructure.

That requires an observability solution designed specifically for AI stacks.

However, most AI monitoring focuses on the LLM layer. It's important to note that there are many other components involved in the AI stack.

Focusing only on the LLM layer can complicate understanding of all the potential issues that could affect an AI system's performance and cost. For example, consider performance: servers, application code, and vector databases could all contribute to issues, and all need to be considered holistically.

End-to-end AI monitoring

Leaders will need full-stack observability, providing comprehensive monitoring of performance, errors, and costs from the application layer, through the LLM and AI layer, and up to the user interface. They should be able to see how many AI apps are running across their entire digital estate, what the overall costs are, and overall error rates.

These are just a few of the errors that crop up frequently in AI applications:

- Your cloud infrastructure might run out of the memory allocated to it, causing your applications to crash.
- Disk input/output (I/O) bottlenecks might cause a slowdown in the performance of the app and degrade the user experience.
- A new release of the LLM or other stack components can lead to unanticipated performance issues or inexplicable errors.

These issues are related to infrastructure, but they might show up first in the application and user experience. A complete monitoring solution should provide observability from end to end in order to accurately track and diagnose these kinds of issues.

This will be increasingly important for companies that need to manage service level commitments (SLCs) and can help to ensure compliance with various regulations and other end-user commitments.

Keeping pace with market shifts

Advances in AI technology are coming so rapidly that it can be difficult to keep up. Even for companies with billion-dollar R&D budgets, staying on top of the latest developments is not guaranteed.

What is guaranteed is that the AI stack components you're using today may not be exactly what you want a year from now—or even six months. In other words, the one constant you can bet on is that change will be constant.

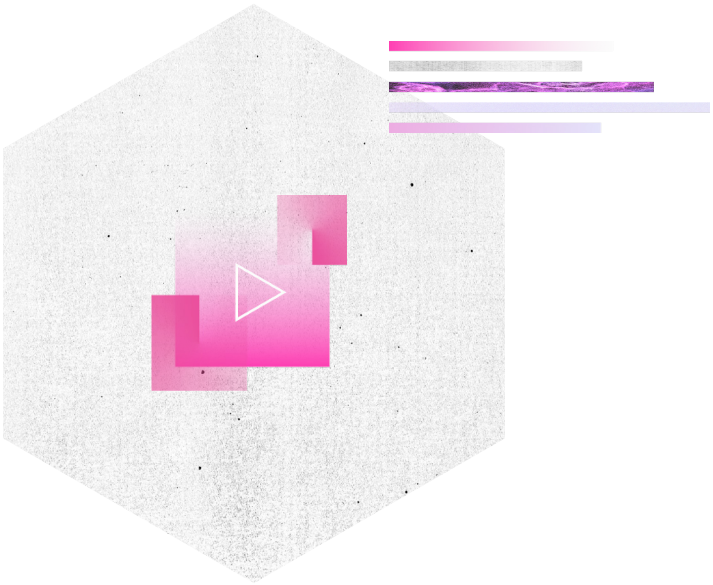
Smart companies will look for AI monitoring solutions that are flexible and extensible, enabling them to keep pace as their infrastructure grows, changes, and incorporates new components.

Additionally, because so many AI stacks are heterogenous, AI leaders should look for solutions that cover all of today's main LLM vendors and frameworks, including OpenAI, Azure OpenAI Service, Amazon Bedrock, Anthropic, LangChain, Nvidia NIM, Groq, and Llama.

Because there are so many options in the AI landscape, monitoring the entire digital estate will be important.

For example, what if your company has a licensing agreement in place to use OpenAI for most of its use cases, but an engineer determines there is a need for using Bedrock for a more discreet project?

Advanced monitoring features should be able to help provide visibility and vulnerability management, even for LLMs and AI code libraries that lie outside of the company's primary use cases.



Monitoring quality

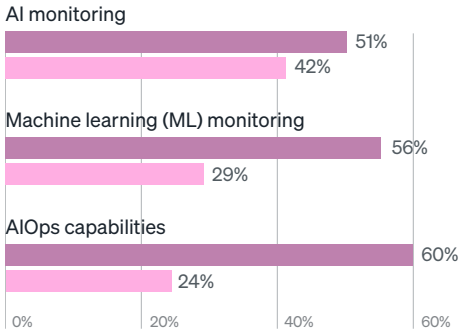
Eventually, companies will want to monitor more than performance, costs, and ROI—they'll want to measure the quality of AI responses. This is important for monitoring and alerting organizations to the quality and fairness of these tools, including possible toxicity and bias generated by their AI models.

The fact is, AI brings a unique set of risks. AI leaders must equip their teams to proactively manage quality concerns such as bias, hallucination, and toxicity that LLMs can sometimes generate. When user complaints surface, how do you respond quickly and appropriately? Better yet, can you detect poor reliability in an AI model's response and respond to it before it affects users?

User expectations also increasingly include assumptions about the accuracy of the facts and originality of the text returned by the LLM. If your AI app isn't able to deliver factual, original responses, you'll need a monitoring solution that can help you diagnose the problem at its source—whether it's the LLM, the datastore, the ML libraries being used, or the application code itself.

In the long term, AI leaders want to know how to continually improve their AI stacks so that the risk of these issues diminishes over time.

Observability capabilities deployment for 2024-2027

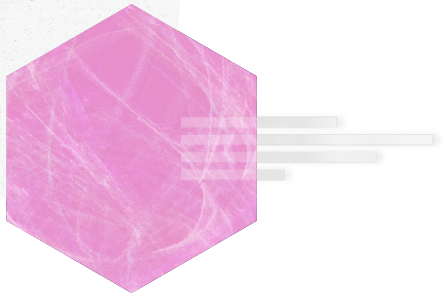


Observability Forecast 2024

AI Monitoring Checklist

A full-stack AI monitoring solution should provide end-to-end observability of all layers of the AI stack, not just the LLM:

- ✓ Orchestration framework
- ✓ LLM
- ✓ ML libraries
- ✓ Model serving
- ✓ Vector databases
- ✓ AI infrastructure



How will you monitor your AI?

With purpose-built AI observability, companies gain end-to-end visibility into their AI workflows. This gives them the necessary insights to troubleshoot, compare, and optimize different approaches and platforms. And it enables them to improve their AI-powered offerings to create completely new customer experiences—like chatbots that customers actually want to talk to or AI-powered user interfaces that feel truly intelligent and intuitive.

Full-stack AI monitoring also enables companies to manage costs, improve performance, reduce glitches, minimize risks, and increase ROI.

It's no wonder that 42% of companies have already deployed AI monitoring, and another 51% plan to do so within three years—this capability is essential. But even more than that, it's becoming a competitive advantage.

CTOs and CIOs need to empower their teams with the ability to monitor the entire AI stack comprehensively. With AI monitoring, teams gain the ability to dive deeper into individual LLMs, compare performance among different LLMs, manage costs, trace prompts and responses, optimize performance of the entire stack, and, eventually, manage quality concerns such as bias, hallucination, and toxicity.

Learn More

MORE RESOURCES

AI and Observability Report
2024 Observability Forecast
Connect with New Relic

With these end-to-end insights across the breadth and depth of the AI application ecosystem, leaders will be well-positioned to optimize AI performance, increase quality, control costs, and deliver ROI.

"Applying observability to AI applications is a smart and efficient way to address [their] complexities so that companies can scale and drive innovation. Any company that provides these solutions is ultimately enabling organizations to deliver better products and customer experiences."

Stephen Elliot
IDC Group Vice President