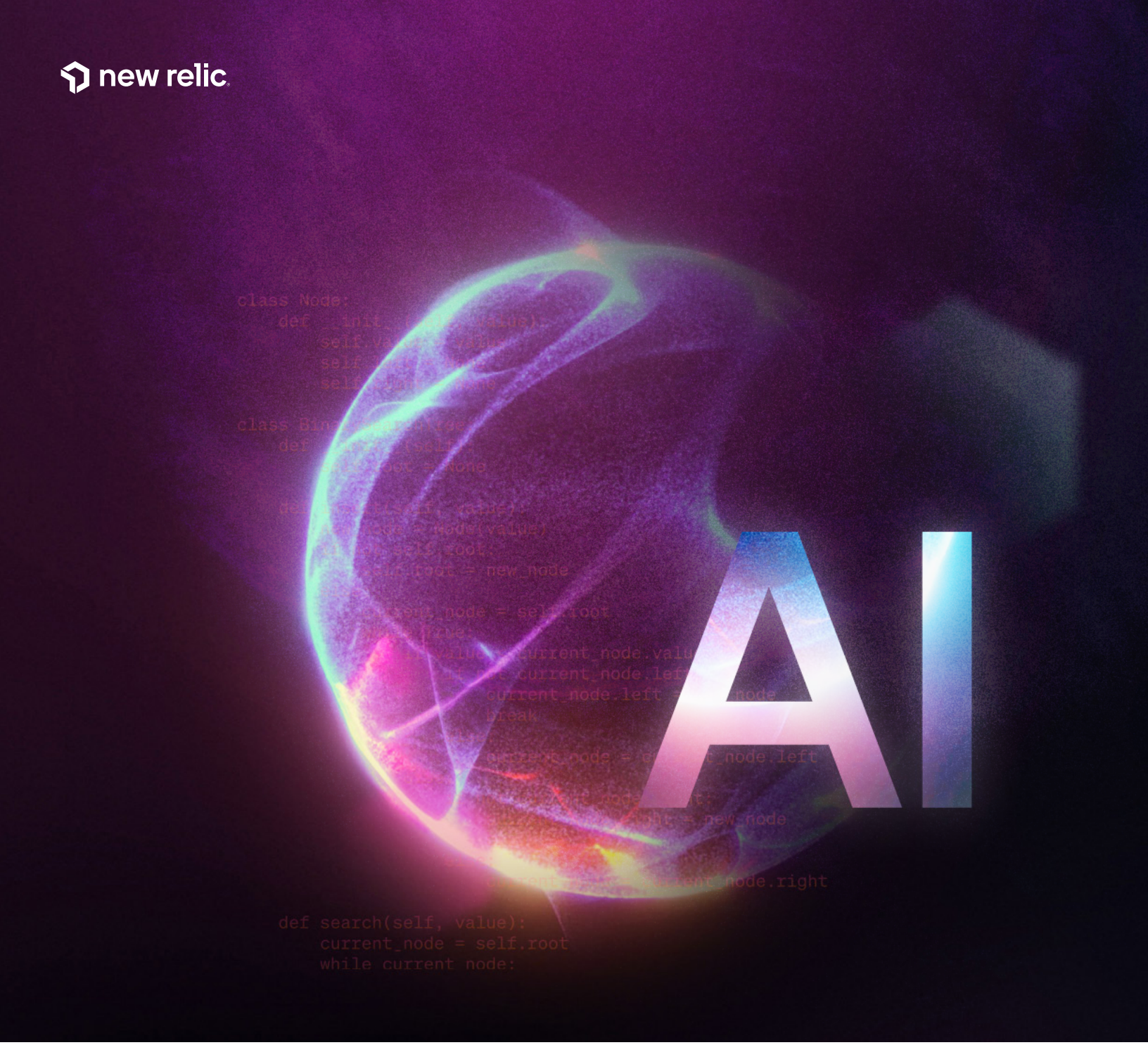
new relic



Komplexe KI-Strategien verlangen nach Full-Stack-KI-Observability

Massive Investitionen in KI machen ein zuverlässiges Monitoring von Performance, Risiko und Resultaten notwendig

Briefing / Dezember 2024

Die Herausforderung: KI-Transparenz

Führende Unternehmen aller Branchen profitieren bereits unmittelbar von generativer KI, ob durch Verbesserungen bei Kundenservice, Content-Erstellung, personalisierten Empfehlungen, Assistenten, intelligenten Chatbots, Risiko-/Betrugsvorhersage und -erkennung oder Prozessautomatisierung.

Aber neue Toolsets bergen auch neue Gefahren und neue Herausforderungen – und gleichzeitig neue Chancen.

Für Unternehmen, die auf KI setzen, ist es wichtig, alle Komponenten zu verstehen, die in ihre KI-Lösungen einfließen. Es steht viel auf dem Spiel, denn KI kann enorme Auswirkungen auf all die Bereiche haben, die Führungskräften und Vorständen besonders am Herzen liegen: Kundenerlebnis, Umsatz, Kosten, Cybersicherheit und sogar Markenwahrnehmung, um nur einige zu nennen. Nicht nur IT-Manager:innen und Entwickler:innen, auch Führungskräfte müssen wissen, wie ihre KI-Modelle funktionieren, welche Kosten damit verbunden sind, welche Schwachstellen es gibt und wie sie miteinander verzahnt sind.

Das mag selbstverständlich klingen, ist aber angesichts der immer komplexer werdenden KI-Technologien von heute eine große Herausforderung.

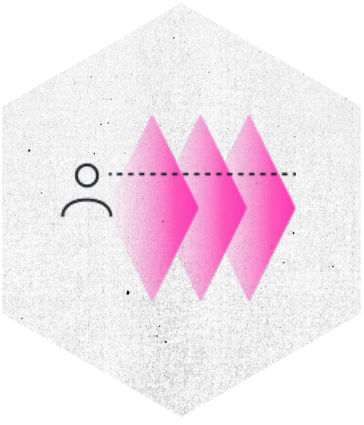
Mit KI-Lösungen erhalten auch neue Frameworks und Komponenten Einzug. Zum einen natürlich die Large Language Models (LLMs), aber auch umfangreiche Datenspeicher, Datenpipelines, Orchestrierungs-Frameworks und ML-Codebibliotheken. Für viele Unternehmen ist das Troubleshooting solcher Systeme eine knifflige Angelegenheit.

Selbst Routine-Monitoring kann schwierig sein, insbesondere wenn Tools verwendet werden, die vor dem Aufkommen von KI-Anwendungsfällen entwickelt wurden. Jedes System verfügt zudem über eine eigene API, eine eigene Methode zum Daten-Reporting und eine eigene Benutzeroberfläche. Deshalb müssen Monitoring-Tools genau auf jedes Modell zugeschnitten werden.

> „So gut wie alle Unternehmen sind momentan dabei, KI-Anwendungen in ihre Tech-Stacks zu integrieren, um ihren Kund:innen eine bessere UX zu bieten, effizienter zu agieren und letztendlich den Gewinn zu steigern. Allerdings hält mit der KI auch eine gewisse Komplexität Einzug in den Tech-Stack, und wir müssen verantwortungsbewusst mit ihr umgehen – besonders im Hinblick auf Qualität, Compliance und Kosteneffizienz."
>
> **Stephen Elliot**
> IDC Group Vice President

Ohne zuverlässige Daten über die Performance und den geschäftlichen Nutzen dieser KI-Systeme werden sich Unternehmen schwertun, fundierte Entscheidungen über neue KI-Investitionen zu treffen oder diese Investitionen an Geschäftsergebnisse und Mehrwert zu knüpfen.

Erst wenn Unternehmen Zugang zu verlässlichen Daten über Performance und Nutzen von KI haben, können sie auch fundierte Entscheidungen treffen.



Risikomanagement, Steigerung des ROI

Der Return on Investment (ROI) ist ein weiteres dringendes Anliegen von Führungskräften in der KI. KI-Investitionen auf Enterprise-Niveau verändern nicht nur Unternehmen, sondern ganze Branchen. Laut einer aktuellen Schätzung von Sequoia Capital gaben Unternehmen weltweit allein im Jahr 2024 so viel für Nvidia-KI-Infrastruktur aus, dass sie lebenslange KI-Einnahmen in Höhe von 600 Milliarden US-Dollar generieren müssten, um die Ausgaben zu rechtfertigen. Und das sind nur die Investitionen eines einzigen Jahres.

Zwar investieren die meisten Unternehmen (noch) nicht in dieser Höhe in KI, allerdings wird die Frage, ob sich die Investitionen rentieren, für alle Unternehmen von entscheidender Bedeutung sein. Vor allem wenn der KI-Hype nachlässt und Führungskräfte sowie Firmenvorstände sehen wollen, dass sich diese neue Technologie in einem höheren ROI widerspiegelt. Viele Unternehmen unterschätzen die Anzahl der Komponenten – und die Höhe der Kosten, die in ihre KI-Stacks fließen –, was die Kostenrechnung zu einer echten Herausforderung machen kann.

Damit sich die Ausgaben rechtfertigen lassen, müssen KI-Modelle einen messbaren Wert liefern und den Umsatz proportional mit der Investition steigern.

Die Lösung? Full-Stack-KI-Monitoring

Die ideale KI-Monitoring-Lösung liefert CTOs und CIOs die notwendigen Daten, um rasche, fundierte Entscheidungen zu treffen, die Kosten zu begrenzen und den ROI zu messen und zu maximieren. Und mit der richtigen Lösung ist es einfacher, die Zuverlässigkeit, Qualität und Effizienz über alle Komponenten des KI-Tech-Stacks hinweg ebenso wie in Services und Infrastruktur zu gewährleisten.

Dazu ist eine Observability-Lösung notwendig, die speziell für KI-Stacks entwickelt wurde.

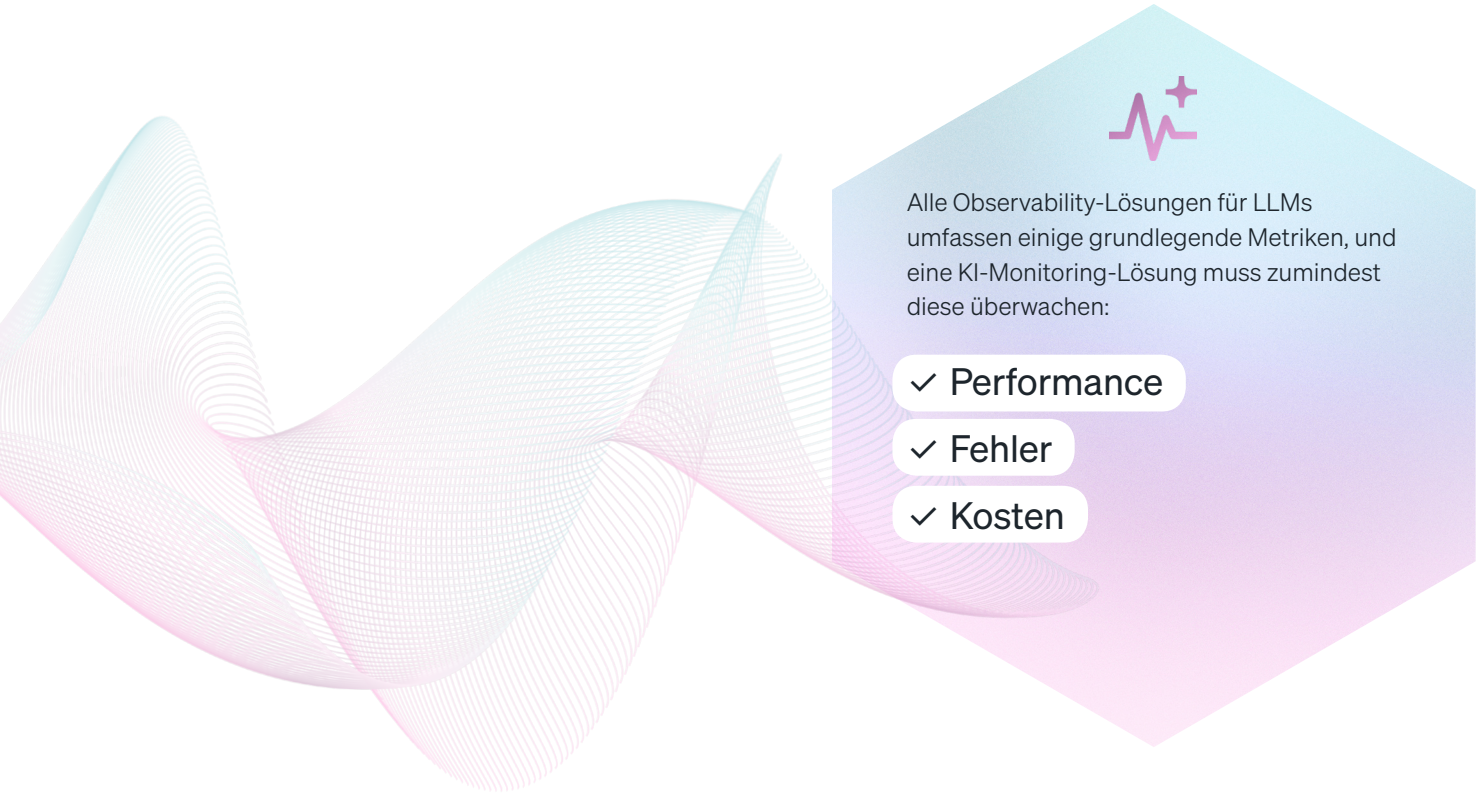
KI-Monitoring konzentriert sich jedoch meist auf die LLM-Ebene, dabei sind im KI-Stack noch zahlreiche andere Komponenten am Werk.

Wenn Sie sich nur auf die LLM-Ebene konzentrieren, entgehen Ihnen eventuell potenzielle Probleme, die sich auf die Performance und Kosten eines KI-Systems auswirken könnten.

Ein Beispiel: Server, Anwendungscode und Vektordatenbanken können alle zu Performance-Problemen beitragen und sind daher ganzheitlich zu betrachte.

End-to-End-KI-Monitoring

Full-Stack-Observability ist also in führenden Unternehmen ein Muss, damit Performance, Fehler und Kosten umfassend überwacht werden – ob in der Anwendungsebene, der LLM- und KI-Ebene oder der UI. Unternehmen müssen sehen können, wie viele KI-Apps in ihrem gesamten digitalen Bestand laufen, wie hoch die Gesamtkosten sind und wie viele Fehler auftreten.



Dies sind nur einige der Fehler, die in KI-Anwendungen häufig auftreten:

- Der Speicher Ihrer Cloud-Infrastruktur wird überlastet, was zu einem Absturz Ihrer Anwendungen führen kann.
- Festplatten-I/O-Bottlenecks können die App verlangsamen und dem Nutzungserlebnis schaden.
- Erscheint ein neues Release des LLM oder anderer Stack-Komponenten, kann dies zu unvorhergesehenen Performance-Problemen oder unerklärlichen Fehlern führen.

Diese Probleme hängen zwar mit der Infrastruktur zusammen, zum Vorschein kommen sie allerdings oft als Erstes in der Anwendung und in der UX. Eine Monitoring-Komplettlösung muss allumfassende Observability bieten, damit solche Probleme präzise nachverfolgt und diagnostiziert werden können.

Dies ist zunehmend von Bedeutung für Unternehmen, die an Service Level Commitments (SLCs) gebunden sind, und kann bei der Einhaltung gesetzlicher Vorschriften und Verpflichtungen gegenüber den Endbenutzer:innen helfen.

Marktveränderungen immer im Griff

KI entwickelt sich so rapide, dass man kaum Schritt halten kann. Selbst in Unternehmen mit milliardenschweren F&E-Budgets ist nicht garantiert, dass sie stets über die neuesten Entwicklungen auf dem Laufenden bleiben.

Was allerdings nahezu garantiert ist: Die aktuell eingesetzten Komponenten Ihres KI-Stacks sind höchstwahrscheinlich nach nur einem oder sogar einem halben Jahr nicht mehr genau das, was Sie brauchen. Anders gesagt: Das Einzige, worauf man sich verlassen kann, ist, dass man sich auf nichts verlassen kann.

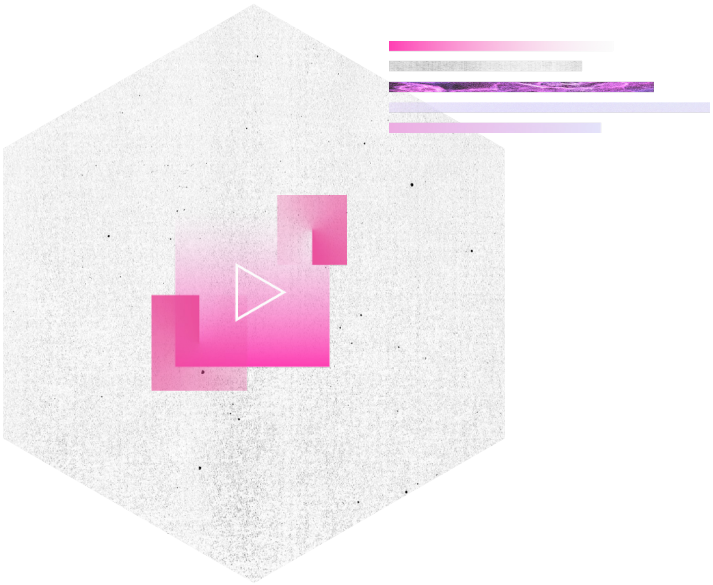
Zukunftsorientierte Unternehmen werden sich nach KI-Monitoring-Lösungen umsehen, die flexibel und erweiterbar sind und mit allen Veränderungen, Erweiterungen und Ergänzungen einer Infrastruktur Schritt halten.

Angesichts der Tatsache, dass viele KI-Stacks ein sehr heterogenes Gebilde sind, sollten KI-Führungskräfte zudem Lösungen auswählen, die aktuell die großen LLM-Anbieter und Frameworks abdecken: OpenAI, Azure OpenAI Service, Amazon Bedrock, Anthropic, LangChain, Nvidia NIM, Groq und Llama.

Bei dieser großen Anzahl an Optionen in Sachen KI ist es wichtig, den gesamten digitalen Bestand zu überwachen.

Was ist beispielsweise, wenn Ihr Unternehmen eine Lizenzvereinbarung zur Verwendung von OpenAI für die meisten Use Cases hat, ein Engineer für ein bestimmtes Projekt nun aber lieber Bedrock verwenden möchte?

Durchdachte Monitoring-Funktionen sollten Transparenz und Vulnerability Management bieten, auch für LLMs und KI-Codebibliotheken außerhalb der primären Anwendungsfälle des Unternehmens.



Qualitätskontrolle

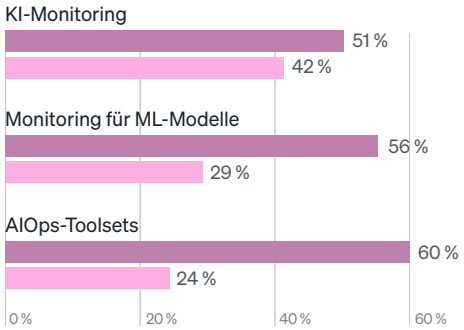
Auf lange Sicht werden Unternehmen nicht nur Performance, Kosten und ROI überwachen wollen, sondern auch die Qualität der KI-Antworten. Dies ist wichtig für das Monitoring und Alerting hinsichtlich Qualität und Fairness dieser Tools, denn dazu gehören auch KI-Probleme wie Toxizität und Bias, die durch KI-Modelle erzeugt werden können.

Eines ist klar: KI birgt eine Reihe einzigartiger Risiken. KI-Führungskräfte müssen ihre Teams in die Lage versetzen, mögliche LLM-Probleme wie Bias, Halluzinationen und Toxizität proaktiv anzugehen. Wie reagieren Sie schnell und angemessen auf Benutzerbeschwerden? Oder noch besser: Können Sie proaktiv herausfinden, ob die Antworten eines KI-Modells unzuverlässig sind, und das Problem so frühzeitig beheben, dass es sich gar nicht erst auf Benutzerseite bemerkbar macht?

Benutzer:innen erwarten zunehmend, dass die vom LLM ausgegebenen Texte inhaltlich richtig und authentisch sind. Wenn Ihre KI-App keine fachlich korrekten Antworten liefern kann, benötigen Sie eine Monitoring-Lösung, mit der Sie das Problem an der Wurzel packen können – ganz gleich, ob es am LLM, am Datenspeicher, an den ML-Bibliotheken oder am Anwendungscode selbst liegt.

Langfristig ist für KI-Führungskräfte vor allem wichtig zu wissen, wie sie ihre KI-Stacks kontinuierlich verbessern können, damit solche Probleme im Laufe der Zeit immer seltener auftreten.

Einsatz von Observability-Toolsets, 2024 bis 2027



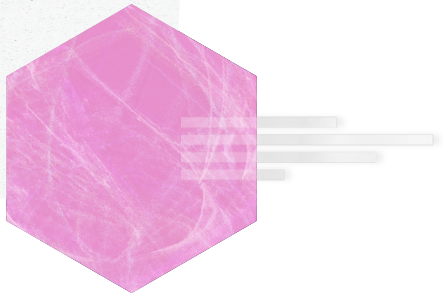
Einsatz innerhalb der nächsten 3 Jahre geplant
Schon jetzt im Einsatz

Observability Forecast 2024

Checkliste fürs KI-Monitoring

Eine Lösung fürs Full-Stack-KI-Monitoring sollte umfassende Observability für alle Ebenen des KI-Stacks bieten, nicht nur für das LLM:

- ✓ Orchestrierungs-Framework
- ✓ LLM
- ✓ ML-Bibliotheken
- ✓ Modellbereitstellung
- ✓ Vektordatenbanken
- ✓ KI-Infrastruktur



Wie werden Sie Ihre KI überwachen?

Mit maßgeschneiderter KI-Observability erhalten Unternehmen einen durchgängigen Einblick in ihre KI-Workflows. Auf diese Weise haben sie die notwendigen Infos, um Troubleshooting durchzuführen sowie verschiedene Strategien und Plattformen zu vergleichen und zu optimieren. Zusätzlich können sie damit ihre KI-gestützten Angebote verbessern und völlig neue Kundenerlebnisse schaffen, z. B. Chatbots, mit denen Kunden tatsächlich sprechen möchten, oder KI-gestützte Benutzeroberflächen, die sich wirklich intelligent und intuitiv anfühlen.

Durch Full-Stack-KI-Monitoring können Unternehmen zudem leichter ihre Kosten verwalten, die Performance verbessern, Störungen reduzieren, Risiken minimieren und den ROI erhöhen.

Da überrascht es nicht, dass 42 % der Unternehmen KI-Monitoring bereits implementiert haben und weitere 51 % das in den nächsten drei Jahren vorhaben, denn dieses Toolset ist inzwischen nicht nur ein Muss, sondern ein echter Wettbewerbsvorteil.

CTOs und CIOs müssen es ihren Teams ermöglichen, den gesamten KI-Stack umfassend zu überwachen. Mit KI-Monitoring erhalten Teams die Möglichkeit, tiefer in einzelne LLMs einzutauchen, die Performance verschiedener LLMs zu vergleichen, die Kosten unter Kontrolle zu behalten, Prompts und Antworten nachzuverfolgen, die Performance des gesamten Stacks zu optimieren und letztendlich Qualitätsprobleme wie Bias, Halluzinationen und Toxizität anzugehen.

Mehr erfahren

WEITERE RESSOURCEN

Report zu KI und Observability
Observability Forecast 2024
New Relic kontaktieren

Mit diesen durchgängigen Einblicken in den gesamten KI-Anwendungspool sind Führungskräfte in der Lage, die KI-Performance zu optimieren, die Qualität zu steigern, Kosten im Rahmen zu halten und einen höheren ROI zu erzielen.

„Mit Observability für KI-Anwendungen lassen sich diese Dinge auf eine intelligente und effiziente Weise angehen und Unternehmen können ihre Innovationsbestrebungen ausweiten und vorantreiben. Anbieter solcher Lösungen sorgen letztendlich dafür, dass Unternehmen bessere Produkte für zufriedenere Kund:innen auf den Markt bringen können."

Stephen Elliot
IDC Group Vice President