new relic

Les stratégies d'IA avancées exigent l'observabilité full-stack de l'IA

Les investissements massifs dans l'IA nécessitent le monitoring robuste des performances, risques et résultats

Synthèse / Décembre 2024

Les défis posés par la visibilité de l'IA

Les leaders de tous les secteurs voient les avantages immédiats de l'IA générative, notamment un service à la clientèle plus précis, la génération d'un meilleur contenu, des recommandations personnalisées plus efficaces, des assistants et chatbots plus intelligents, des prédictions et détections des risques/fraudes améliorées et une plus grande automatisation des processus.

Mais si ces capacités offrent de nouvelles opportunités, elles présentent également de nouveaux risques et difficultés.

Pour les entreprises qui parient le plus sur l'IA, il est essentiel de comprendre tous les composants qui sont inclus dans leurs solutions d'IA. Ces enjeux sont dus au potentiel d'impact énorme de l'IA dans tous les domaines qui sont importants pour les cadres et les comités de direction : l'expérience client, le chiffre d'affaires, les coûts, la cybersécurité et la perception de la marque, pour n'en citer que quelques-uns. Outre les leaders et développeurs des technologies de l'information, les cadres supérieurs doivent aussi savoir comment performent leurs modèles d'IA, et connaître les coûts associés, les vulnérabilités potentielles et leurs interactions les uns avec les autres.

Cette demande peut sembler évidente, mais dans un monde où les stacks technologiques sont de plus en plus complexes, la tâche peut s'avérer gigantesque.

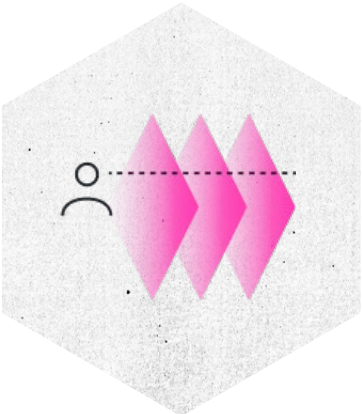
Les solutions d'IA proposent de nouveaux frameworks et composants. Elles comptent évidemment les grands modèles de langage (LLM), mais il ne faut pas oublier aussi les datastores, les pipelines de données, les frameworks d'orchestration et les bibliothèques de code pour l'apprentissage machine (ML). Dans de nombreuses organisations, le dépannage de ces systèmes est une toute autre paire de manches.

Le monitoring des comportements de routine peut lui-même s'avérer compliqué, notamment lors de l'utilisation d'outils développés avant l'arrivée de cas d'utilisation de l'IA. Chaque système est également accompagné de sa propre API, de sa propre façon de rapporter les données et de sa propre interface — ce qui signifie que les outils de monitoring doivent être personnalisés pour chaque modèle.

> « Quasiment toutes les organisations intègrent des applications d'IA à leurs stacks technologiques pour garantir une meilleure expérience client et pour améliorer l'efficacité dans l'espoir d'améliorer aussi les résultats. Mais l'IA crée une plus grande complexité au niveau du stack technologique et elle doit être adoptée de façon responsable en gardant à l'esprit la sécurité, la qualité, la conformité et les coûts. »
>
> **Stephen Elliot**
> Vice-président, IDC Group

Sans données fiables sur les avantages que ces systèmes d'IA apportent en matière de performances et d'activité, les entreprises ont du mal à prendre des décisions éclairées sur les nouveaux investissements d'IA ou à les associer aux résultats et à la valeur de l'entreprise.

Tant qu'elles n'ont pas accès à des données fiables sur les performances et les avantages de l'IA, les entreprises auront du mal à prendre des décisions éclairées.



Gestion des risques et augmentation du RSI

Pour les leaders de l'IA, le retour sur investissement (RSI) est une autre préoccupation importante. Les investissements en IA transforment non seulement les organisations, mais aussi des secteurs entiers. Par exemple, selon une estimation récente de Sequoia Capital, les entreprises du monde entier ont dépensé tant d'argent sur la seule infrastructure Nvidia AI en 2024 qu'elles devront générer 600 milliards d'USD en revenus associés à l'IA sur toute sa durée de vie pour justifier les dépenses. Et cela pour une seule année d'investissements !

Bien que la plupart des entreprises n'investissent pas de telles sommes dans l'IA (pour l'instant), toutes vont devoir prouver le RSI, notamment après que l'emballement pour l'IA se calme et que les cadres supérieurs et leurs comités de direction commencent à exiger des résultats tangibles de cette nouvelle technologie. Cependant, de nombreux leaders ne sont pas au courant des composants et coûts consacrés aux stacks d'IA, ce qui peut rendre la comptabilité très difficile.

Pour justifier les dépenses, les modèles d'IA doivent apporter une valeur mesurable — et stimuler un chiffre d'affaires en adéquation avec l'investissement.

La solution : le monitoring full-stack

Les bonnes solutions de monitoring de l'IA devraient apporter aux CTO et CIO les données dont ils ont besoin pour prendre des décisions en temps voulu, limiter les dépenses, et mesurer et optimiser le RSI. Une fois ces bonnes solutions en place, les leaders peuvent mieux assurer la fiabilité, la qualité et l'efficacité de tous les composants du stack technologique d'IA, mais aussi des services et de l'infrastructure.

Pour cela, une solution d'observabilité conçue spécifiquement pour les stacks d'IA est nécessaire.

Toutefois, la plupart des solutions de monitoring de l'IA se concentrent sur la couche LLM alors que de nombreux autres composants sont également impliqués dans le stack d'IA.

Si l'on se concentre uniquement sur la couche LLM, cela peut compliquer la compréhension de tous les problèmes potentiels qui pourraient avoir un impact sur les performances et le coût du système d'IA. Par exemple, considérez les performances : les serveurs, les codes des applications et les bases de données vectorielles pourraient tous contribuer aux problèmes et doivent tous être examinés de manière holistique.

Le monitoring de l'IA de bout en bout

Les leaders ont besoin d'une observabilité full-stack qui leur fournit un monitoring complet des performances, erreurs et coûts de la couche des applications jusqu'à l'interface utilisateur en passant par la couche des LLM et de l'IA. Ils doivent pouvoir voir le nombre d'applications d'IA en cours d'exécution sur tout leur domaine numérique, les coûts globaux et les taux d'erreur totaux.

Toutes les solutions d'observabilité pour les LLM comprennent quelques métriques de base et leur monitoring est le principal défi des solutions de monitoring de l'IA :

- ✓ Performances
- ✓ Erreurs
- ✓ Coûts

Voici seulement quelques-unes des erreurs qui surviennent fréquemment dans les applications d'IA :

- Il se peut que trop peu de mémoire ait été allouée à votre infrastructure cloud, ce qui peut entraîner le plantage de votre application.
- Les goulots d'étranglement au niveau des opérations d'entrée/sortie (I/O) d'un disque peuvent être causés par un ralentissement des performances de l'application et détériorer l'expérience utilisateur.
- Une nouvelle version des LLM ou d'autres composants du stack peut entraîner des problèmes de performances non anticipés ou des erreurs inexplicables.

Ces problèmes sont liés à l'infrastructure, mais il est possible qu'ils apparaissent d'abord dans l'application et au niveau de l'expérience utilisateur. Une solution de monitoring complète doit fournir l'observabilité de bout en bout afin de faire le suivi exact et de diagnostiquer précisément ce type de problèmes.

Cela va gagner en importance pour les entreprises qui doivent gérer les engagements au niveau des engagements de niveau de service (SLC) et peut assurer la conformité à différentes réglementations et divers engagements aux utilisateurs finaux.

Suivi de l'évolution du marché

Les avancées de la technologie de l'IA arrivent tellement vite qu'il peut être difficile de garder le rythme. Même les entreprises avec des budgets de R et D en milliards de dollars ont du mal à garder la tête hors de l'eau face à la déferlante des tout derniers développements.

Ce qui est sûr, c'est que les composants du stack d'IA que vous utilisez aujourd'hui ne seront peut-être pas exactement ce que vous voudrez dans un an — voire six mois. En d'autres termes, la seule constante sur laquelle vous pouvez compter, c'est le changement.

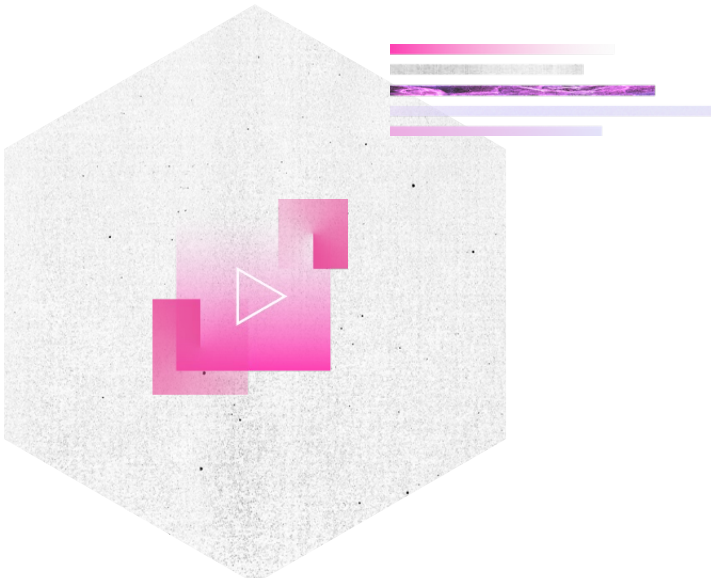
Les entreprises intelligentes vont se tourner vers des solutions de monitoring de l'IA qui sont flexibles et extensibles afin de garder la cadence avec la croissance et les changements de l'infrastructure, et d'incorporer de nouveaux composants.

En outre, un grand nombre de stacks d'IA sont tellement disparates qu'il faudrait que les leaders d'IA cherchent des solutions couvrant tous les fournisseurs et frameworks LLM d'aujourd'hui, y compris OpenAI, Azure OpenAI Service, Amazon Bedrock, Anthropic, LangChain, Nvidia NIM, Groq, et Llama.

Étant donné qu'il y a tant d'options dans le monde de l'IA, le monitoring de tout le domaine numérique va être très important.

Par exemple, que se passe-t-il si votre entreprise dispose d'un accord de licence pour utiliser OpenAI dans la plupart de ses cas d'utilisation, mais qu'un ingénieur se rend compte que Bedrock est plus adapté pour un projet unique ?

Les fonctionnalités avancées du monitoring devraient aider à fournir la visibilité et à gérer les vulnérabilités, même pour les LLM et les bibliothèques de code d'IA qui se trouvent en dehors des principaux cas d'utilisation de l'entreprise.



Qualité du monitoring

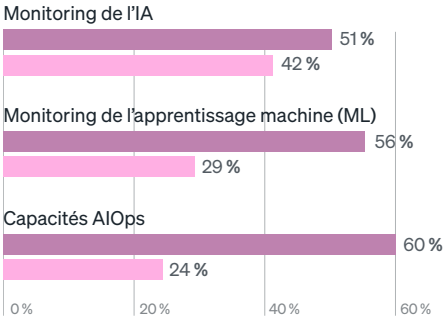
Tôt ou tard, les entreprises voudront monitorer plus que les performances, les coûts et le RSI, et connaître la qualité des réponses de l'IA. Il s'agit-là d'un élément important pour le monitoring et les alertes envoyées aux organisations en matière de qualité et d'impartialité de ces outils, y compris la toxicité et les biais possibles générés par leurs modèles d'IA.

Il faut bien admettre que l'IA apporte un ensemble unique de risques. Les leaders de l'IA doivent donner à leurs équipes les moyens de gérer de manière proactive les soucis de qualité tels que les biais, les hallucinations et la toxicité parfois générés par les LLM. Lorsque les plaintes des utilisateurs font surface, comment y répondez-vous rapidement et correctement ? Mieux encore, pouvez-vous détecter le manque de fiabilité dans la réponse d'un modèle d'IA et y répondre avant que les utilisateurs ne soient touchés ?

Les attentes des utilisateurs présupposent de plus en plus l'exactitude des faits et l'originalité du texte renvoyé par le LLM. Si votre application d'IA ne peut pas apporter de réponses factuelles et originales, vous aurez besoin d'une solution de monitoring qui vous aidera à diagnostiquer le problème à la source — qu'il s'agisse du LLM, du datastore, des bibliothèques ML utilisées ou du code de l'application lui-même.

À long terme, les leaders d'IA veulent savoir comment continuellement améliorer leurs stacks d'IA afin de réduire dans le temps les risques associés à ces problèmes.

Capacités de déploiements 2024-2027



Prévisions 2024 sur l'observabilité

Liste de contrôle du monitoring de l'IA

Une solution full-stack du monitoring de l'IA doit observer de bout en bout toutes les couches du stack d'IA au lieu des seuls LLM :

- ✓ Framework d'orchestration
- ✓ LLM
- ✓ Bibliothèques ML
- ✓ Mise en service du modèle
- ✓ Bases de données vectorielles
- ✓ Infrastructure d'IA

Comment allez-vous monitorer votre IA ?

Avec l'observabilité de l'IA sur mesure, les entreprises gagnent une visibilité de bout en bout dans leurs workflows d'IA. Cela leur donne les informations détaillées dont elles ont besoin pour le dépannage, mais aussi pour comparer et optimiser les différentes approches et plateformes. Et cela leur permet aussi d'améliorer leurs offres optimisées par l'IA pour créer une expérience client complètement nouvelle — comme des chatbots auxquels les clients veulent réellement parler ou des interfaces utilisateur optimisées par l'IA qui semblent vraiment intelligentes et intuitives.

Le monitoring de l'IA full-stack permet également aux entreprises de gérer les coûts, d'améliorer les performances, de réduire les bogues, de minimiser les risques et d'augmenter le RSI.

Il n'est donc pas étonnant que 42 % des entreprises aient déjà déployé le monitoring de l'IA et qu'un autre 51 % prévoie de le faire dans les trois ans. Cette capacité est non seulement essentielle, mais, ce qui est encore plus important c'est qu'elle devient un véritable avantage concurrentiel.

Les CTO et les CIO doivent donner à leurs équipes la capacité de monitorer le stack d'IA dans son entier. Avec le monitoring de l'IA, les équipes peuvent plonger plus en profondeur dans les LLM, comparer les performances entre différents LLM, gérer les coûts, faire le suivi des prompts et des réponses, optimiser les performances de tout le stack et enfin, gérer les soucis de qualité tels que les biais, les hallucinations et la toxicité.

En savoir plus

RESSOURCES SUPPLÉMENTAIRES

Rapport sur l'IA et l'observabilité
Prévisions 2024 sur l'observabilité
Contactez New Relic

Avec ces informations précieuses de bout en bout sur toute l'ampleur et la profondeur de l'écosystème des applications d'IA, les leaders seront bien placés pour optimiser les performances de l'IA, améliorer la qualité, contrôler les coûts et augmenter le RSI.

« L'application de l'observabilité aux applications d'IA est une façon intelligente et efficace de gérer [leurs] complexités afin que les entreprises puissent faire évoluer et stimuler l'innovation. Toute entreprise fournissant ces solutions permet en fait aux organisations de livrer de meilleurs produits et expériences client. »

-Stephen Elliot
Vice-président, IDC Group