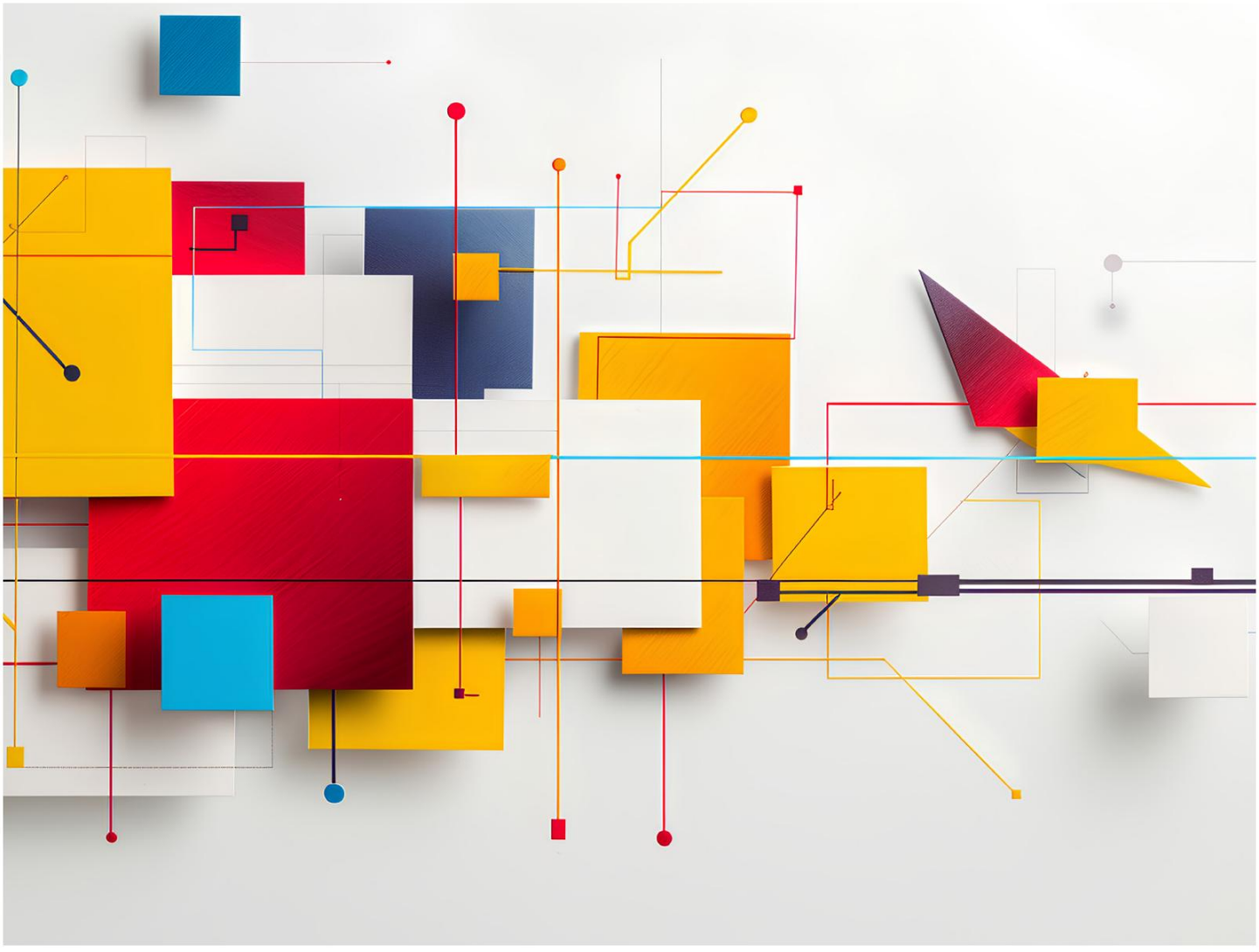
MIT Technology Review Insights

Réalisé en partenariat avec

Redis

Une nouvelle pile logicielle pour l'IA générative préparera les organisations aux défis du passage des pilotes d'IA à l'IA en production.

Passer de l'IA générative à la product



Principaux points à retenir

1 Les entreprises sont en train d'adopter le potentiel transformateur de l'IA générative et sont impatientes de mettre cette technologie en production. Cependant, les difficultés de mise en œuvre ralentissent le déploiement pour beaucoup d'entre elles.

2 Les dirigeants citent la qualité de sortie de l'IA, la complexité de l'intégration, les coûts élevés de l'inférence et de la formation des modèles et la latence des applications comme les principaux défis. et ils cherchent à intégrer des systèmes d'IA dans la solution.

3 Construire une pile d'IA solide et adaptable prenant en charge une variété de modèles de base, y compris des solutions d'intégration, et offrant des outils de nouvelle génération, seront essentiels au succès des entreprises avec IA générative.

Depuis l'introduction de ChatGPT en novembre 2022, les entreprises se sont tournées vers les grands modèles de langage (LLM) et les modèles d'IA générative à la recherche de solutions à leurs problèmes les plus complexes et les plus exigeants en main-d'œuvre. La promesse que le service client pourrait être confié à des plateformes de chat hautement qualifiées capables de reconnaître le problème d'un client et de présenter des commentaires techniques conviviaux, par exemple, ou que les entreprises pourraient décomposer et analyser leurs trésors de données non structurées, des vidéos aux PDF, a suscité un intérêt massif des entreprises pour cette technologie.

Ce battage médiatique se transforme en production. La part des entreprises qui utilisent l'IA générative dans au moins une fonction commerciale a presque doublé cette année pour atteindre 65 %, selon McKinsey. La grande majorité des organisations (91 %) Les applications d'IA générative devraient augmenter leur productivité, l'informatique, la cybersécurité, le marketing, le service client et le développement de produits étant parmi les domaines les plus touchés, selon Deloitte.

Pourtant, la difficulté de déployer avec succès l'IA générative continue d'entraver les progrès. Les entreprises savent que l'IA générative pourrait transformer leurs activités. que s'ils ne parviennent pas à adopter cette nouvelle stratégie, ils seront laissés pour compte, mais ils sont confrontés à des obstacles lors de sa mise en œuvre. Cela laisse les deux tiers des chefs d'entreprise ambivalent ou insatisfaits des progrès réalisés dans le cadre de leurs déploiements d'IA. Et tandis qu'au troisième trimestre 2023, 79 % des entreprises ont déclaré qu'elles prévoyaient pour déployer des projets d'IA générative au cours de la prochaine année, seulement 5 % ont déclaré avoir des cas d'utilisation en production en mai 2024.

« Nous commençons tout juste à comprendre comment rendre le déploiement de l'IA productif et rentable », explique Rowan Trollope, PDG de Redis, un créateur de plateformes de données en temps réel et d'accélérateurs d'IA. « Le coût et la complexité de la mise en œuvre de ces systèmes ne sont pas simples. »

Méthodologie

En août 2024, MIT Technology Review Insights a mené un sondage sur les défis et les choix auxquels les organisations sont confrontées lors du déploiement de cas d'utilisation d'IA générative. Les 250 dirigeants interrogés représentent un large éventail de secteurs d'activité et travaillent dans des organisations du monde entier.

Estimations de l' impact éventuel de l'IA générative sur le PIB

Les investissements dans l'IA vont d'un peu moins de 1 000 milliards de dollars à 4 400 milliards de dollars par an, avec des impacts prévus sur la productivité comparables à ceux d'Internet, de l'automatisation robotique et de la machine à vapeur. Pourtant, si la promesse d'une croissance accélérée des revenus et de réductions des coûts demeure, le chemin à parcourir pour atteindre ces objectifs est complexe et souvent coûteux. Les entreprises doivent trouver des moyens de créer et de déployer efficacement des projets d'IA avec des composants bien compris à grande échelle, explique Trollope.

« Nous commençons tout juste à comprendre comment rendre le déploiement de l'IA productif et rentable. »

Rowan Trollope, PDG de Redis

Défis et complexité du déploiement

Il n'est pas facile de faire passer l'IA générative en production. Interrogés sur les défis liés à la production, près des trois quarts des chefs d'entreprise (72 %) ont déclaré s'inquiéter de la qualité des résultats de leurs systèmes d'IA. Dans le même sondage MIT Technology Review Insights, environ 6 sur 10 ont déclaré s'inquiéter également de l'intégration, des coûts élevés liés à l'inférence des modèles et des coûts élevés liés à la formation des modèles.

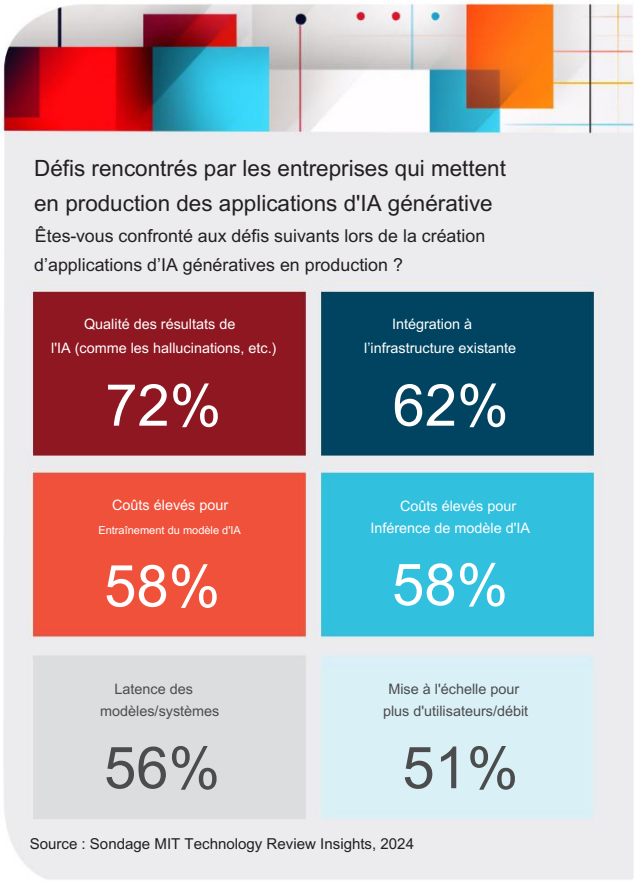
Étant donné que la plupart des entreprises ne forment pas leurs propres modèles d'IA génératifs, une intégration appropriée est nécessaire pour fournir les bonnes données et le bon contexte au modèle, explique Harrison Chase, cofondateur et PDG de LangChain, qui se concentre sur les solutions d'intégration et d'orchestration. « À un niveau élevé, l'un des principaux problèmes est de donner le bon contexte au modèle », explique-t-il. « Vous pouvez vouloir prendre un résultat d'un LLM précédent, et vous devez essentiellement récupérer toutes ces données et les transmettre. » Les développeurs d'IA ont besoin « d'une couche d'orchestration qui aide à cette orchestration du contexte », ajoute-t-il.

En outre, les entreprises doivent être en mesure de déterminer la quantité de contexte à fournir à un modèle. Les systèmes d'IA générative fonctionnent généralement mieux avec plus de contexte, mais plus de contexte équivaut à plus de coûts, il est donc important de trouver le bon compromis, explique Trollope.

Le nombre de jetons utilisés par un modèle est une bonne estimation de la quantité de contexte fournie. « Pour beaucoup, tout est une question de coût par jeton. Si vous pouvez réduire ce coût par jeton, vous pouvez rendre les inférences plus efficaces », explique Trollope.

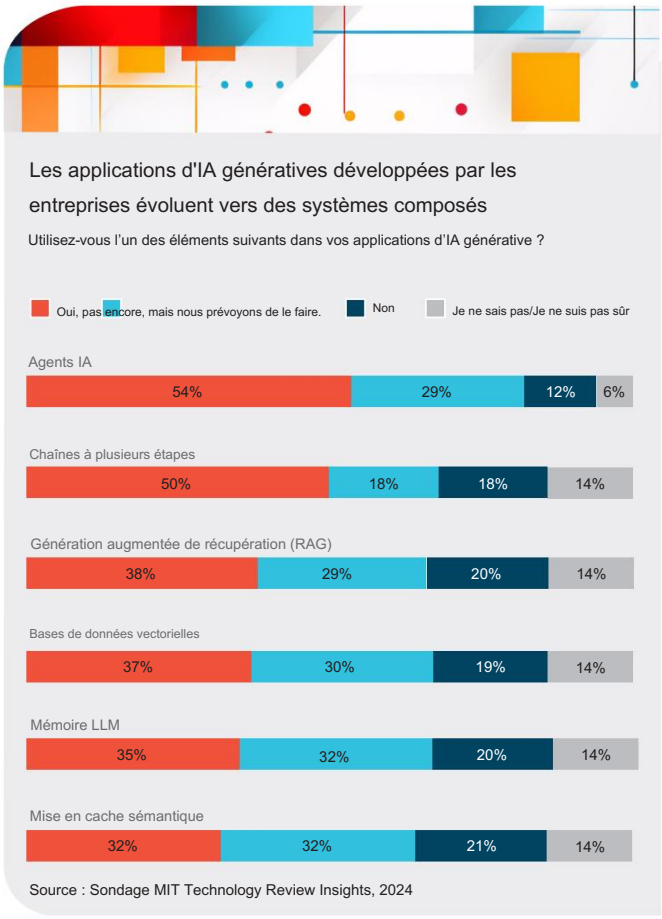
Les coûts ont été une préoccupation majeure dans l'enquête, peut-être en partie parce qu'il reste difficile de déterminer les avantages de l'IA générative. Le calcul du retour sur investissement des systèmes et produits d'IA est un processus complexe comportant des incertitudes importantes, et les coûts de déploiement et d'exploitation peuvent être difficiles à quantifier. Deloitte la recherche a trouvé que les entreprises « ont du mal à définir et à mesurer les impacts exacts de leurs initiatives d'IA générative », 48 % d'entre elles utilisant des indicateurs de performance clés et 38 % créant des cadres spécifiques à l'entreprise pour évaluer les investissements en IA générative.

Les entreprises continueront d'attendre de pouvoir déterminer de manière fiable le retour sur investissement, affirme Chase. « Les entreprises veulent estimer le retour sur investissement à l'avance, et c'est encore assez difficile à faire », dit-il. « Quantifier le retour sur investissement à l'avance ou même après coup est encore très difficile, et je pense que c'est un obstacle majeur à l'étape de la construction. »



Source : Sondage MIT Technology Review Insights, 2024

L'avantage de l'IA composite Dans de nombreux cas, les systèmes d'IA composite, qui rassemblent plusieurs modèles, technologies ou capacités d'IA, apparaissent comme des solutions efficaces aux défis de déploiement. Les systèmes d'IA composite, parfois appelés agents d'IA ou systèmes agentiques, peuvent relier différents modèles d'IA spécialisés dans différentes tâches, combiner plusieurs technologies liées à l'IA ou intégrer des modules d'IA individuels dont les différentes capacités se combinent pour s'attaquer à des tâches plus complexes. Ces types de systèmes ont suscité un large intérêt parmi les répondants à l'enquête, une majorité (54 %) déclarant qu'ils utilisent déjà des agents d'IA et 29 % prévoyant de le faire à l'avenir.



Les systèmes d'IA composites qui enchaînent plusieurs modèles spécifiques à une tâche peuvent être utilisés pour réduire les coûts et améliorer les performances, explique Chase. « Une fois que l'application commence à décoller, les entreprises se demandent si elles peuvent utiliser un modèle moins cher ici ? », explique-t-il. « Peut-être que j'utilise un modèle coûteux à un endroit, mais des modèles moins chers à un autre. Ou peut-être que je prends ce modèle coûteux et que je le divise en deux ou trois appels de modèles moins chers, car c'est ce qui est nécessaire pour le coût et la latence. »

Le routage sémantique est une autre solution courante dans ce domaine. « Il oriente l'utilisateur vers le bon outil, qui peut être un autre modèle, ou même une IA. Il peut orienter la demande de l'utilisateur vers un outil de planification ou même vers un humain pour répondre à son besoin », explique Trollope. Près de la moitié des personnes interrogées déclarent utiliser actuellement des chaînes à plusieurs étapes dans leurs applications d'IA générative, et 18 % d'entre elles déclarent qu'elles prévoient de le faire à l'avenir.

La génération augmentée par récupération (RAG) est une technique d'IA composée qui ajoute un composant de récupération à celui de génération. La capacité de récupération permet au système de rechercher dans les fichiers, les documents et les données, en trouvant des informations très pertinentes et fondées sur lesquelles baser ses résultats. Cela permet aux organisations d'adopter des modèles d'IA à usage général, puis de les adapter et de les affiner

Les personnes interrogées s'intéressent beaucoup au RAG, 38 % d'entre elles déclarant utiliser cette technique actuellement et 29 % supplémentaires déclarant qu'elles prévoient de le faire prochainement.

Les caches sémantiques et les bases de données vectorielles sont deux composants supplémentaires qui peuvent ajouter de l'efficacité aux systèmes d'IA. Les caches sémantiques regroupent les requêtes de modèles similaires et dupliquées en fonction de leur signification et de leur contexte (pas seulement des mots littéraux utilisés), puis stockent et réutilisent les réponses du modèle selon les besoins. Les bases de données vectorielles sont les bases de données spécialisées utilisées pour stocker les vecteurs de grande dimension, ou « incorporations », qui représentent ces requêtes et réponses et qui permettent

« Les entreprises souhaitent estimer le retour sur investissement à l'avance, ce qui reste encore assez difficile. Quantifier le retour sur investissement à l'avance ou même après coup reste un véritable défi, et je pense que c'est un obstacle majeur à l'étape de la construction. »

Harrison Chase, cofondateur et PDG de LangChain

Mise en correspondance par similarité. Les personnes interrogées adoptent ces deux outils, 37 % d'entre elles déclarant utiliser actuellement des bases de données vectorielles et 32 % la mise en cache sémantique. L'intérêt pour ces technologies augmente également rapidement, environ un tiers des personnes interrogées (30 % pour les bases de données vectorielles et 32 % pour la mise en cache sémantique) déclarant qu'elles prévoient de les adopter à l'avenir.

Redis estime qu'entre 30 et 80 % des requêtes adressées à un modèle de langage étendu (LLM) sont des doublons, de sorte que les gains d'efficacité liés à la mise en cache peuvent être substantiels. Un groupe de recherche a découvert répétition fréquente même au niveau d'un seul utilisateur, calculant qu'en moyenne 31 % des requêtes ChatGPT des participants étaient répétitives de leurs propres requêtes précédentes.

Les caches sémantiques peuvent donc accélérer considérablement ces requêtes et réduire considérablement les coûts d'inférence des modèles, explique Trollope. « Nous avons des clients dont 90 % des appels LLM sont traités grâce à notre mise en cache sémantique, qui repose sur une base de données vectorielle », explique-t-il. « Ces utilisateurs n'ont pas besoin d'attendre le LLM et le client contrôle ses coûts. »

Construire la pile

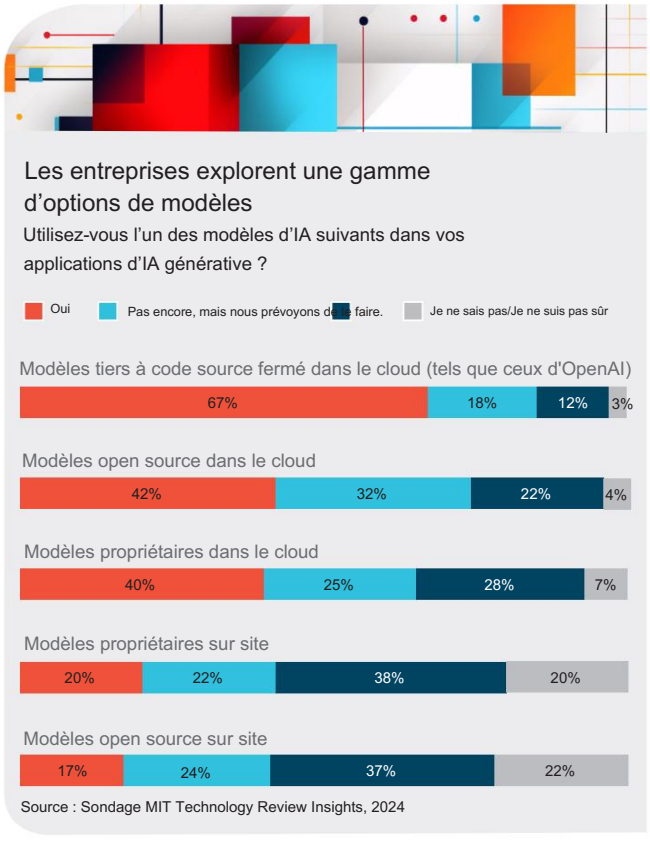
À mesure que les entreprises concrétisent leurs ambitions en matière d'IA et mettent en production leurs applications d'IA générative, elles devront se concentrer sur la création de la pile technologique nécessaire pour les soutenir. La première étape consistera à déterminer sur quel(s) modèle(s) de base elles souhaitent s'appuyer.

La majorité des organisations interrogées (67 %) ont commencé par créer des applications d'IA génératives avec des modèles tiers à source fermée, tels que ceux d'OpenAI. Mais nombreux sont ceux qui aspirent à intégrer également d'autres types de modèles.

Les modèles open source sont en vogue, probablement motivés par la volonté des entreprises de contrôler la sécurité et les coûts. Près des trois quarts des répondants déclarent utiliser actuellement des modèles open source basés sur le cloud (42 %) ou prévoient de le faire. le feront à l'avenir (32 %). 41 % supplémentaires déclarent être intéressés par les modèles open source sur site : 17 % les utilisent actuellement et 24 % prévoient de le faire à l'avenir.

Le développement d'intégrations entre les technologies est également important pour la pile d'IA adaptable. Bien que les invites ne soient pas nécessairement facilement transférées entre les modèles, les interfaces de programmation d'applications (API) courantes peuvent aider les entreprises à échanger des composants plus efficaces ou plus performants.

Des entreprises comme Redis et LangChain soutiennent cette flexibilité en fournissant des interfaces ou des intégrations standard entre les principaux outils d'IA. « Les API peuvent ne pas avoir toutes les mêmes paramètres, et c'est donc là que l'utilisation de quelque chose



« LangChain peut aider », explique Chase, « car nous avons une interface standard pour tous les différents modèles, et vous pouvez donc basculer entre eux. »

Les personnes interrogées font preuve d'une grande connaissance et d'une grande sophistication des prochaines étapes de l'IA générative, et une grande partie d'entre elles affirment que leurs entreprises disposent de capacités d'IA matures en place ou en cours de développement. Pourtant, l'écart persiste dans ce que les entreprises ont pu mettre en production jusqu'à présent.

Trollope affirme qu'il y a un retard dans l'adoption des outils. Les entreprises doivent déployer les outils et solutions d'IA de nouvelle génération qui leur permettront de créer, de gérer et d'intégrer les applications d'IA génératrices de leur imagination. « L'industrie doit rattraper son retard sur les avancées majeures de la recherche, avec les outils de mise en œuvre nécessaires pour faciliter cette tâche », ajoute-t-il. « Une fois ces outils en place, je pense que nous assisterons à une accélération du développement de l'IA. »

Alors que les entreprises construisent leurs piles d'IA, en collaboration avec des partenaires comme Redis et LangChain, elles se préparent également aux innovations de l'IA du futur.

« Les entreprises ont besoin de modèles qui fonctionnent rapidement avec une plate-forme de données à haut débit qui peut garantir que lorsque vous récupérez des données, elles arrivent rapidement », explique Trollope. « Il ne peut y avoir beaucoup de latence dans ces systèmes. »

> « L'industrie doit rattraper son retard sur les avancées majeures de la recherche, en se dotant des outils de mise en œuvre nécessaires pour faciliter cette tâche. Une fois ces outils en place, je pense que nous assisterons à une accélération du développement de l'IA. »
>
> Rowan Trollope, PDG de Redis

Gestion de la latence dans les applications d'IA générative

Même si les capacités de l'IA générative se développent, ses utilisateurs attendent davantage de sa réactivité. L'interactivité vocale en temps réel, par exemple, est une technologie puissante et intuitive. À mesure qu'elle se généralise, les utilisateurs exigeront qu'elle fonctionne à la vitesse d'une conversation humaine. Les modèles de raisonnement comme o1 d'OpenAI peuvent fournir de bien meilleures réponses si on leur donne le temps de réfléchir. Tout cela devra être fait rapidement pour répondre aux attentes
attentes des utilisateurs.

« Il ne peut y avoir aucun décalage dans ces expériences », déclare Rowan Trollope, PDG de Redis. « La latence est le nouveau temps d'arrêt. » Les personnes interrogées sont d'accord : 56 % d'entre elles déclarent que la latence des modèles ou des systèmes est un défi qu'elles ont rencontré lors de la mise en production d'applications d'IA génératives.

Harrison Chase, PDG de LangChain, constate que les entreprises reconnaissent la nécessité d'accélérer le développement de leurs applications d'IA générative. « L'un des grands défis consiste à trouver la bonne expérience utilisateur pour ces applications », explique-t-il. L'un des thèmes récurrents que nous entendons est que les gens passent souvent autant de temps sur l'UX que sur l'ingénierie des messages, car une grande partie de l'UX est liée à la latence.

À mesure que les organisations évoluent vers des solutions d'IA composées plus sophistiquées, la latence deviendra une préoccupation croissante. « L'IA composée est égale à la latence composée et à la composition « Le coût est un facteur important », explique Trollope. « Dans un système multi-modèle, chaque modèle ajoute sa propre latence au fur et à mesure que vous les additionnez dans la pile. » Alors que les entreprises explorent ces systèmes d'IA plus complexes, leurs effets sur la vitesse resteront un défi majeur.

« Moving generative AI into production » est un document d'information de MIT Technology Review Insights. Nous tenons à remercier tous les participants ainsi que le sponsor, Redis. MIT Technology Review Insights a collecté et rapporté toutes les conclusions contenues dans ce document de manière indépendante, indépendamment de la participation ou du parrainage. Teresa Elsey a été la rédactrice de ce rapport et Nicola Crepaldi en a été l'éditeur.

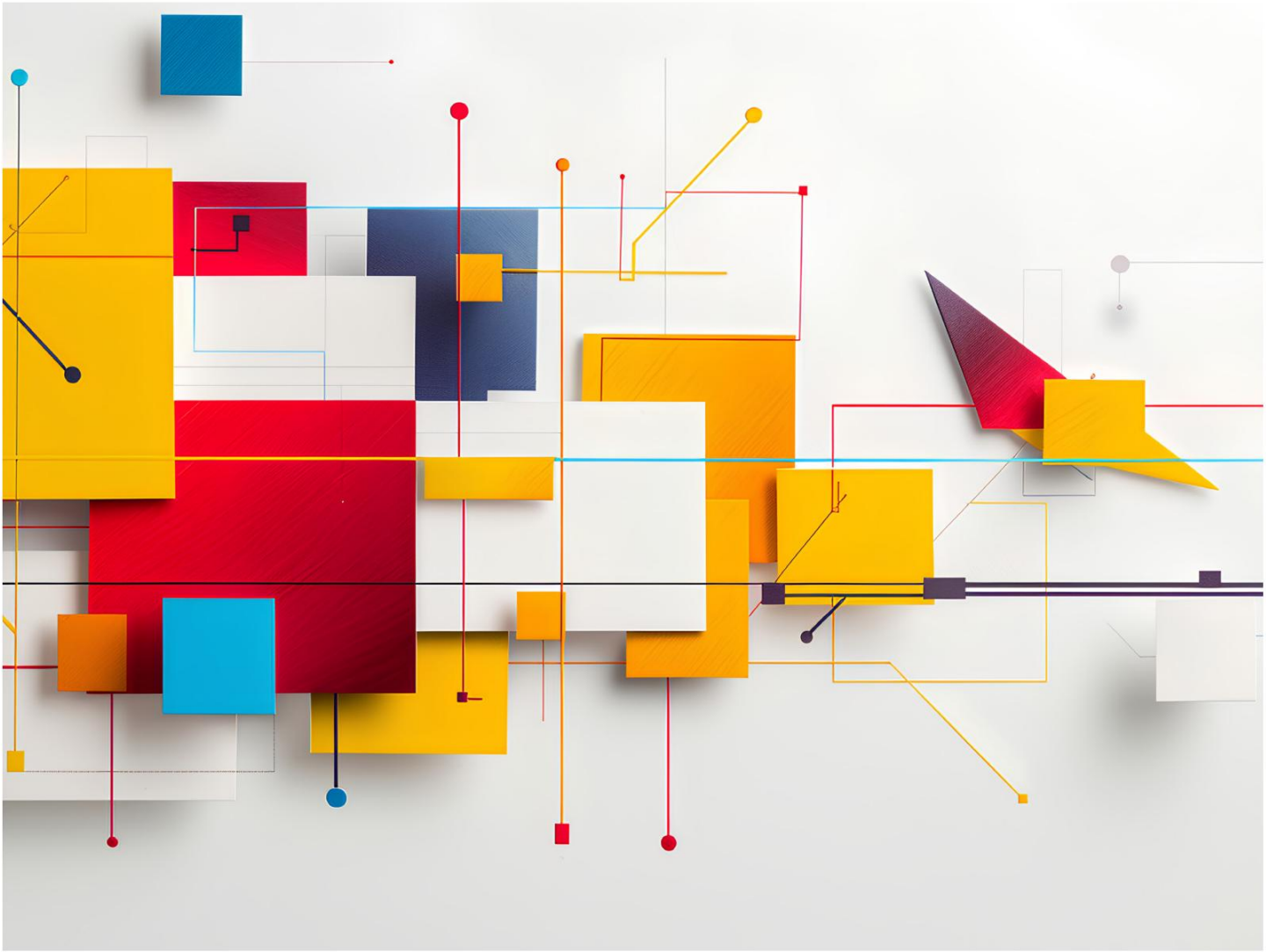
À propos de MIT Technology Review Insights

MIT Technology Review Insights est la division d'édition personnalisée de MIT Technology Review, le magazine technologique le plus ancien au monde, soutenu par la plus grande institution technologique au monde. Insights produit des événements en direct et des recherches sur les principaux défis technologiques et commerciaux du moment. Insights mène des recherches et des analyses qualitatives et quantitatives aux États-Unis et à l'étranger et publie une grande variété de contenus, notamment des articles, des rapports, des infographies, des vidéos et des podcasts. Et grâce à son panel MIT Technology Review Global Insights en pleine croissance, Insights a un accès inégalé aux cadres supérieurs, aux innovateurs et aux entrepreneurs du monde entier pour des enquêtes et des entretiens approfondis.

Du sponsor

Redis est la plateforme de données la plus rapide au monde. Depuis ses origines open source en 2011 jusqu'à devenir la marque la plus citée pour les solutions de mise en cache, Redis a aidé plus de 10 000 clients à créer, faire évoluer et déployer les applications sur lesquelles notre monde fonctionne. Avec des bases de données cloud et sur site pour la mise en cache, GenAI et plus encore, Redis aide les entreprises numériques à établir une nouvelle norme en matière de vitesse des applications. Implantée à San Francisco, Austin, Londres et Tel Aviv, Redis est reconnue internationalement comme le leader de la création rapide d'applications. Pour en savoir plus, rendez-vous sur redis.io.

Illustrations
Couverture et illustrations ponctuelles créées avec Adobe Stock.



Informations sur la revue technologique du MIT

www.technologyreview.com

insights@technologyreview.com