Red Hat

Meilleurs conseils pour

poser les bases de votre IA générative

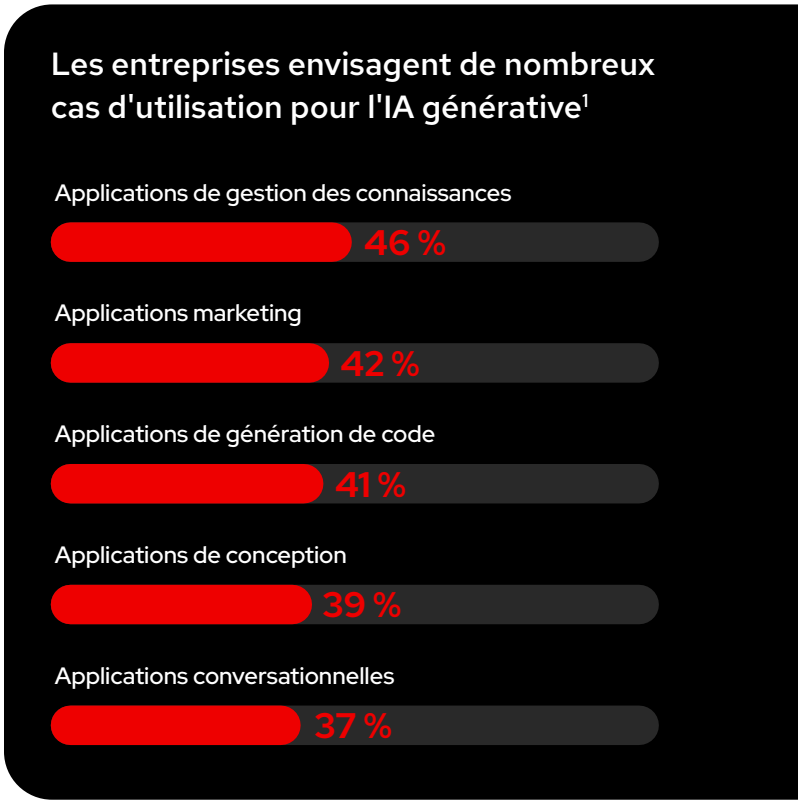
Sommaire

Explorer de nouvelles possibilités pour l'innovation

L'**intelligence artificielle (IA) générative** constitue un outil puissant pour les entreprises qui souhaitent créer des produits innovants, optimiser les processus et obtenir un avantage concurrentiel sur des marchés en rapide évolution. Basée sur les avancées en matière d'apprentissage profond et de réseaux neuronaux, cette technologie permet non seulement de traiter des données, mais aussi de générer du contenu nouveau et original, contrairement à l'IA prédictive. Elle transforme la collaboration entre l'homme et la machine, favorise l'émergence de nouvelles approches en matière de résolution des problèmes et offre des avantages métier significatifs dans divers secteurs.

Des entreprises du monde entier conçoivent de nouvelles applications innovantes à l'aide de technologies d'IA générative. En effet, 39 % d'entre elles investissent actuellement dans ces technologies, tandis que 37 % étudient des cas d'utilisation potentiels[1]. Voici quelques-uns des nombreux cas d'utilisation de l'IA générative à l'heure actuelle.

- **Générer des prévisions pour des scénarios complexes.** L'IA générative analyse des données historiques, identifie des schémas et établit des prévisions précises. Elle facilite ainsi la planification stratégique et la gestion des risques.
- **Personnaliser les campagnes marketing.** En analysant les données de la clientèle pour comprendre ses préférences et ses comportements, l'IA générative est en mesure de créer des supports marketing personnalisés, notamment des e-mails, publicités et promotions, qui contribuent à maximiser l'engagement et le taux de conversion.
- **Automatiser et personnaliser le service clientèle.** Intégrée aux chatbots et assistants virtuels intelligents, l'IA générative répond automatiquement aux demandes de la clientèle et interagit avec elle pour fournir un service efficace et personnalisé.



1 IDC Web Conference Proceeding, « Unlocking Business Success with Generative AI », document n° US50789223, juin 2023

L'IA générative source de nouvelles inquiétudes

Bien que les avantages et les inconvénients de l'IA générative ne soient pas encore complètement définis, de nombreuses entreprises souhaitent investir sans tarder dans cette nouvelle technologie. Il est toutefois important qu'elles comprennent les problèmes liés à l'IA générative afin d'établir des consignes éthiques et des frameworks de développement clairs, de se conformer aux réglementations gouvernementales et sectorielles, ainsi que de détecter et de corriger les problèmes potentiels.

- **Confidentialité des données.** L'usage de données personnelles ou sensibles pour l'entraînement des modèles d'IA générative suscite l'inquiétude ainsi que des interrogations sur la protection de la vie privée des individus.
- **Propriété des données.** L'utilisation de modèles propriétaires (ou de modèles pré-entraînés à l'aide de données propriétaires) cause des problèmes relatifs à la propriété des données susceptibles de provoquer des litiges.
- **Biais et objectivité.** On sait déjà que les réponses fournies par les outils d'IA générative reflètent les biais humains, notamment des stéréotypes dangereux et des discours haineux.
- **Utilisation éthique.** Les modèles d'IA générative peuvent créer du contenu synthétique et des deepfakes utilisés pour des activités malveillantes telles que les violations de confidentialité et les campagnes de désinformation.
- **Explicabilité et interprétabilité.** Le manque de transparence des outils d'IA générative complexifie l'interprétation, la compréhension et l'explication des résultats du modèle, ce qui entraîne un problème de responsabilité en cas d'informations incorrectes ou inventées.
- **Conséquences involontaires.** La nature autonome de l'IA générative risque d'induire des conséquences involontaires susceptibles de porter atteinte à certaines personnes et entreprises.
- **Défis en matière de réglementations.** La progression des technologies d'IA générative risque de dépasser celle du développement des cadres qui les régissent, ce qui complique la création et l'application de consignes garantes d'une utilisation éthique et responsable.
- **Consommation d'énergie.** L'entraînement des modèles d'IA demande de nombreuses ressources de calcul et énergétiques, ce qui soulève des interrogations quant à l'impact environnemental et à la durabilité de ce processus.

Ce livre numérique présente plusieurs conseils utiles pour poser les bases d'une infrastructure de confiance capable de prendre en charge les initiatives en matière d'IA générative.

Préparez-vous à l'IA générative

Dans son document « Unlocking Business Success with Generative AI », IDC recommande aux entreprises plusieurs actions pour se préparer à l'adoption de l'IA générative[2] :

- **Créez un environnement propice à l'expérimentation agile** des cas d'utilisation prioritaires qui répond à vos besoins.
- **Développez des politiques d'entreprise** sur l'utilisation responsable pour décourager les comportements malveillants.
- **Évaluez les effets de l'IA générative sur vos équipes** et procédez à une gestion proactive du changement.
- **Collaborez avec des fournisseurs de technologies** et de services fiables pour votre infrastructure d'IA.
- **Assurez-vous de disposer durablement des compétences techniques nécessaires** grâce au recrutement, à la formation ou à la prise en charge de services professionnels.

2 IDC Web Conference Proceeding, « Unlocking Business Success with Generative AI », document n° US50789223, juin 2023

Conseils pour poser les bases de l'IA générative

La base technologique que vous choisissez pour vos initiatives en matière d'IA générative influence considérablement votre capacité à adopter la technologie et réussir sa mise en œuvre. Dans ce chapitre, nous présentons plusieurs conseils pour vous aider à poser les bases de votre IA générative.

Conseil n° 1 : utilisez des outils éprouvés

Le développement d'applications basées sur des modèles d'IA générative peut devenir une tâche complexe. Avec les bons outils (composés de langages, de frameworks et d'environnements d'exécution basés sur des projets Open Source et des solutions commerciales), vous pouvez accélérer l'ajustement des modèles et simplifier le développement ainsi que le déploiement des applications.

Optez pour une base d'IA qui prend en charge les outils de votre choix pour développer rapidement et efficacement des solutions d'IA. La réalisation d'analyses exploratoires des données, de l'entraînement et de l'ajustement par le biais d'interfaces interactives contribue à simplifier la collaboration. À l'aide d'outils pré-intégrés et de fonctionnalités en libre-service, il est possible de rationaliser l'exploitation informatique et ainsi de maintenir la portabilité et la cohérence entre les environnements.

Conseil n° 2 : ajustez rapidement vos modèles

Puisque l'entraînement des modèles d'IA générative constitue un processus coûteux et chronophage, la plupart des entreprises conçoivent des solutions d'IA à partir de modèles de fondation pré-entraînés à l'aide de données générales. Les data scientists ajustent ensuite ces modèles avec des données diverses et propres à un domaine pour qu'ils puissent effectuer des tâches spécialisées. Cependant, ce processus d'ajustement nécessite souvent de nombreuses ressources, notamment des processeurs puissants ainsi qu'une infrastructure de cloud hybride distribuée.

Optez pour des plateformes d'IA avec des capacités de gestion et d'orchestration des charges de travail distribuées capables de déployer des sessions d'entraînement dans les environnements de cloud hybride, quels que soient la taille du modèle, le volume de données ou la durée. Les options d'ajustement des modèles de fondation dans des datacenters sur site simplifient la mise en conformité avec les exigences techniques et réglementaires pour les modèles restreints. Enfin, les fonctions d'entraînement par lots vous permettent d'anticiper l'ajustement des charges de travail et de faciliter le partage et la gestion des ressources.

Autres solutions pour ajuster des modèles

Des chercheurs tentent de trouver comment ajuster des modèles de fondation plus rapidement et efficacement. La **génération augmentée de récupération (RAG)** est un framework d'IA qui permet de récupérer des faits à partir de sources externes, telles que des bases de données internes, des intranets d'entreprise ou Internet, afin de fournir les informations les plus précises et à jour possible aux modèles d'IA générative.

Avec le **« prompt tuning »**, les modèles d'IA reçoivent des signaux ou des « prompts » (invites) front-end, par exemple des mots supplémentaires ou des numéros générés par IA, afin de guider les modèles vers la décision souhaitée et de permettre aux entreprises qui disposent d'un faible volume de données de personnaliser un modèle de fondation pour une tâche précise.

Conseil n° 3 : distribuez efficacement les modèles

Il peut s'avérer difficile pour les équipes d'exploitation de proposer des expériences utilisateur de haute qualité à partir de solutions d'IA génératives. D'un côté, la demande variable en applications requiert une infrastructure évolutive et une gestion automatisée. De l'autre, le déploiement efficace de modèles nécessite des capacités de surveillance des performances et de restauration rapide des versions précédentes. Enfin, puisque les solutions d'IA traitent de vastes quantités de données, il est également essentiel d'appliquer des normes de sécurités strictes au sein des différents environnements.

Optez pour des plateformes capables de déployer et mettre à l'échelle des modèles ainsi que des applications d'IA générative dans des clouds hybrides, notamment des infrastructures sur site, des ressources de cloud public et des appareils d'edge computing. Les options de distribution des modèles d'IA générative à partir d'environnements sur site ou isolés assurent la non-utilisation des données propriétaires pour le réentraînement des modèles disponibles publiquement. En outre, la prise en charge des déploiements selon le modèle Canary ainsi que des outils d'explicabilité améliore la cohérence et la fiabilité des réponses fournies par les modèles.

Conseil n° 4 : automatisez la gestion du cycle de vie

Les pipelines **d'intégration et de distribution continues (CI/CD)** peuvent déployer et gérer automatiquement les solutions d'IA générative. Grâce au réentraînement et à la mise à jour des modèles et des applications par le biais de modifications rapides et incrémentielles, vous pouvez accélérer le développement et augmenter les performances des modèles. Les pipelines d'IA sont toutefois plus complexes que les workflows CI/CD standard, car ils incluent souvent des étapes supplémentaires telles que l'extraction de données, l'entraînement, l'ajustement, la validation et le réentraînement.

Optez pour une base qui vous offre la possibilité de créer et d'intégrer des pipelines d'IA (à partir d'outils CD/CD tels que Tekton et Jenkins) dans des workflows DevOps existants, afin de développer, d'entraîner, de surveiller et de réentraîner rapidement et efficacement les modèles d'IA générative. Les outils **GitOps** de distribution continue comme ArgoCD permettent de définir et d'automatiser des déploiements de solutions d'IA complexes en tant que code pour distribuer des modèles et applications de manière cohérente.

Conteneurs pour l'IA générative

Les technologies de **conteneurs** et **Kubernetes** offrent des capacités de déploiement agile, de gestion et d'évolutivité pour accélérer le développement cloud-native de solutions d'IA générative. Provisionnez des environnements à la demande pour les datacenters sur site, les clouds publics et les appareils d'edge computing. Assurez la création, le déploiement, la mise à l'échelle et la gestion automatiques d'instances de conteneurs au sein d'infrastructures physiques et virtuelles. Intégrez également des composants et des magasins de données à partir d'un écosystème robuste de fournisseurs Open Source et commerciaux dans des solutions d'IA générative. Découvrez les **avantages des conteneurs pour l'IA**.

Conseil n° 5 : surveillez les modèles de manière cohérente

Les modèles d'IA générative peuvent avoir un impact majeur sur les personnes et les entreprises. Grâce au suivi du comportement des modèles, vous êtes en mesure d'analyser les décisions et les justifications, d'identifier les mauvaises performances et de signaler immédiatement tout comportement problématique. Ces données améliorent l'efficacité de la gouvernance qui garantit que les modèles produisent des informations justes, correctes et non biaisées au sein des environnements de production.

Examinez les bases de votre IA à l'aide de capacités de surveillance centralisées qui mesurent le biais et les écarts de données, détectent les anomalies et offrent une explicabilité point par point afin de vous aider à analyser, entretenir et corriger les modèles d'IA générative. La surveillance automatique et continue dans les environnements de production améliore la conformité avec les normes d'entreprise en matière de gouvernance des modèles. De plus, les interfaces d'outils intuitives ainsi que les rapports non techniques et lisibles par un humain favorisent l'utilisation et la maintenance responsables des modèles.

Définition des principaux concepts relatifs à l'IA générative

- **Un biais** désigne la présence dans un modèle de schémas comportementaux qui influent sur l'éthique et l'inclusivité des informations produites, notamment en favorisant certains groupes ou en générant des réponses qui alimentent les stéréotypes.
- **Un écart de données** survient lorsque les propriétés statistiques des données d'entraînement évoluent au fil du temps. Résultats : une réduction des performances du modèle et des réponses moins précises ou pertinentes.
- **La détection des anomalies** désigne le processus d'identification et de signalement des comportements inhabituels des modèles ou des comportements qui s'écartent des exemples fournis pendant l'entraînement.
- **L'explicabilité point par point** se réfère à la capacité à comprendre pourquoi les modèles produisent certains résultats, ce qui améliore la visibilité sur les applications pour lesquelles la transparence est essentielle.

Conseil n° 6 : tirez parti des écosystèmes de partenaires

Afin de proposer des expériences utilisateur innovantes, les solutions d'IA générative nécessitent de nombreux composants intégrés. Avec la bonne combinaison de technologies issues d'un écosystème collaboratif de fournisseurs de confiance, vous pouvez accélérer le développement des applications, corriger les biais et écarts de données, et ainsi garantir des performances cohérentes et fiables pour l'ensemble de votre solution.

Optez pour des fournisseurs de plateformes qui disposent d'un écosystème étendu de partenaires certifiés dont les solutions complètes permettent de développer et déployer des modèles et des applications d'IA générative. En vous appuyant sur une vaste sélection de composants (intégration et préparation des données, entraînement et distribution des modèles), vous serez en mesure de développer et déployer des solutions d'IA de manière rapide et efficace. Avec des solutions certifiées à l'interopérabilité éprouvée, vous pouvez également réduire le nombre de demandes d'assistance informatique et augmenter la productivité.

Conseil n° 7 : collaborez avec des spécialistes du domaine

Pour déployer et gérer efficacement des solutions d'IA générative, il faut de l'expérience et des connaissances spécialisées. Les exigences en matière d'évolutivité, les inquiétudes concernant la fiabilité et l'intégration aux systèmes existants peuvent compliquer les déploiements en production. L'utilisation inefficace des ressources de calcul risque d'entraîner des coûts inutiles, tandis que le non-respect des normes de sécurité, des politiques de confidentialité et des frameworks réglementaires sur l'IA est susceptible de provoquer des conséquences indésirables.

Optez pour des fournisseurs dont les équipes de spécialistes offrent une assistance complète ainsi que des conseils pour la création de vos solutions d'IA générative. Par exemple, une équipe d'ingénieurs dédiée peut fournir les outils, ressources et connaissances nécessaires pour accélérer vos projets d'IA. Des consultants spécialistes peuvent vous aider à surmonter les défis liés au déploiement, à optimiser l'efficacité de l'infrastructure et à garantir l'interopérabilité de votre solution d'IA. Enfin, vous pouvez acquérir les connaissances et l'expertise nécessaires pour lancer plus rapidement de nouveaux projets grâce à des services de formation professionnelle.

Pas d'IA générative sans collaboration

Pour mener à bien vos projets d'IA générative, il est nécessaire de créer une équipe pluridisciplinaires[3].

- **Les responsables métier** utilisent la solution ou en subissent les impacts.
- **Les spécialistes de l'IA** ajustent les modèles d'IA générative et en assurent la maintenance ainsi que la mise à jour.
- **Les data scientists** traitent les données d'entraînement des modèles en amont pour garantir leur précision et éviter les biais.
- **Les responsables de l'éthique et de la conformité** veillent à ce que tous les projets d'IA générative respectent les réglementations applicables.
- **Les spécialistes de l'exploitation** intègrent des solutions aux infrastructures existantes et mettent en œuvre des politiques de sécurité.

3 Kearney, « Standing up tiger teams to tackle generative AI complexity », novembre 2023

Innover rapidement à l'aide d'une base ouverte et flexible

Avec notre gamme complète de technologies, notre expérience éprouvée et nos partenariats stratégiques, nous pouvons vous aider à atteindre vos objectifs en matière d'IA générative. Nous proposons une base de développement et de déploiement pour les modèles et les applications d'IA générative, ainsi que des services et des formations pour accélérer leur adoption.

Red Hat® OpenShift® est une plateforme d'applications unifiée pour les entreprises, conçue pour l'innovation cloud-native. Cette solution offre à vos équipes la rapidité et la flexibilité dont elles ont besoin pour réussir, avec les ressources de calcul à la demande, la prise en charge de l'accélération du matériel et la cohérence dans les environnements sur site, de cloud public et d'edge computing. Vous pouvez ainsi créer une plateforme en libre-service pour que les équipes de science des données, d'ingénierie des données et de développement puissent rapidement concevoir des applications intelligentes. Les fonctions de collaboration permettent aux équipes de créer des résultats de modélisation dans des conteneurs, puis de les partager de manière cohérente avec leurs homologues ainsi qu'avec les développeurs.

Red Hat OpenShift AI s'appuie sur Red Hat OpenShift afin de fournir une plateforme complète pour concevoir, entraîner, ajuster, déployer et surveiller des modèles et des applications, tout en répondant aux exigences des solutions d'IA modernes en matière de charges de travail et de performances. Les équipes peuvent passer rapidement de l'expérimentation à la production dans un environnement cohérent et collaboratif qui intègre les offres des principaux partenaires certifiés, notamment NVIDIA, Intel, Starburst, Anaconda, IBM, Run:ai et Pachyderm. Associée à notre écosystème de technologies, la solution Red Hat OpenShift AI fournit des composants et capacités qui accélèrent le développement ainsi que le déploiement de solutions d'IA générative innovantes dans des clouds hybrides.

IBM watsonx.ai AI studio propose une sélection de modèles et d'options de déploiement avec les fonctionnalités d'IA générative dont vos applications intelligentes ont besoin. Déployez des modèles (Open Source, tiers, modèles de fondation IBM, etc.) dans tous les emplacements où résident vos charges de travail afin d'optimiser les performances et l'efficacité de vos solutions d'IA. Avec les **modèles de fondation développés par IBM** et entraînés à partir de données pertinentes pour votre entreprise, vos solutions d'IA générative sont également en mesure de comprendre les nuances liées à votre domaine d'activité afin de vous procurer un avantage concurrentiel.

Le service d'IA générative **Red Hat Ansible® Lightspeed with IBM watsonx Code Assistant** permet aux équipes de créer, d'adopter et de gérer des contenus d'automatisation de manière plus efficace. Connectée à IBM watsonx Code Assistant, la solution Red Hat Ansible Lightspeed vous aide à transformer vos idées d'automatisation en code Ansible avec des invites en langage naturel. Vous pouvez ainsi améliorer la productivité ainsi que l'accessibilité de l'automatisation au sein de votre entreprise.

Se lancer avec l'IA générative

L'IA générative est un outil puissant qui permet de créer des contenus originaux et de transformer la manière dont nous interagissons avec les applications et la technologie.

Notre gamme de produits s'appuie sur notre expertise, nos technologies et nos partenariats afin d'offrir à vos équipes une base commune pour créer et déployer des applications d'IA ainsi que des modèles d'AA, de manière transparente et contrôlée. Nous utilisons d'ailleurs nos propres outils et plateformes d'IA afin d'améliorer l'efficacité d'autres logiciels Open Source. De plus, la compatibilité avec les solutions de nos partenaires vous permet d'accéder à un écosystème d'outils d'IA fiables qui fonctionnent avec les plateformes Open Source telles que Red Hat OpenShift AI.

En savoir plus et essayer Red Hat OpenShift AI gratuitement

Lancez-vous plus rapidement avec les services de consulting Red Hat

Travaillez avec nos spécialistes pour démarrer sans attendre vos projets d'IA/AA. Nous proposons des services de consulting et de formation pour aider votre entreprise à adopter l'IA/AA plus rapidement.

- Découvrez nos services pour l'IA/AA : red.ht/aiml-consulting
- Organisez une session de découverte gratuite : redhat.com/consulting

1057715_0324_KVM

Red Hat