

# Guide destiné aux cadres pour mettre en place l'IA générative

De l'expérimentation et la mise en œuvre : comment utiliser l'IA générative  
dans des scénarios concrets



# Table des matières

## En route vers l'IA générative 3

### Première partie

#### Comprendre l'environnement de l'IA générative 5

Qu'est-ce que l'IA générative ? 6

Qu'est-ce que le Machine Learning ? 6

Qu'est-ce qu'un grand modèle de langage (LLM) ? 6

Qu'est-ce que la génération augmentée de récupération (RAG) ? 8

Qu'est-ce qu'une base de données vectorielle ? 11

#### De quoi l'IA générative est-elle capable ? 12

Étape 0 : Définir la base et déterminer les possibilités 13

#### Adaptation en fonction de votre secteur 16

Télécommunications 19

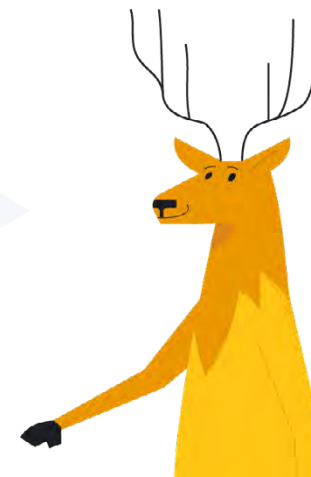
Services financiers 20

Commerce de détail 21

Automobile et fabrication 23

Secteur public 24

[Commencez par ici si vous savez déjà ce que vous attendez de l'IA générative](#)



### Deuxième partie

#### Mise en place de l'IA générative : 27

##### un petit pas pour la machine, un grand pas pour votre entreprise

Étape 1 : Définir les résultats que vous voudriez atteindre dans l'idéal 28

Étape 2 : Déterminer l'impact et mesurer le succès 29

Étape 3 : Sélectionner un modèle 30

Étape 4 : Faire des essais et apprendre de ses échecs 34

Étape 5 : Gouvernance et opérations 36

Ce qu'il faut savoir à propos de la sécurité des données 37

Étape 6 : Établir une chronologie et des jalons 38

#### Le début d'une nouvelle ère 40

# En route vers l'IA générative

Le moins qu'on puisse dire, c'est que l'IA générative est la technologie la plus révolutionnaire à avoir vu le jour en 2023. Dans tous les secteurs, on prédit que l'IA générative jouera un rôle quasi-incontournable dans les années à venir. Mais combien d'entreprises peuvent affirmer qu'elles ont compris comment l'IA générative fonctionne et comment en tirer le meilleur parti ?

Alors que les entreprises surfent sur la vague de l'IA générative, certaines d'entre elles commencent à en voir les résultats. Par exemple, chez Cisco, les ingénieurs peuvent instantanément trouver des réponses synthétiques et pertinentes dans des cas similaires traités par le support technique, des forums de discussion internes ou des articles de la base de connaissances concernant les problèmes rencontrés par les clients. Cisco récolte déjà les fruits de l'IA générative, en résolvant 90 % des demandes de support technique grâce à sa solution de recherche repensée et en économisant quelque 5 000 heures d'intervention des ingénieurs par mois.<sup>1</sup>

Dans le royaume du e-commerce, vous avez sans doute déjà vu l'IA générative à l'œuvre. L'IA générative est capable d'analyser les achats passés d'un client en parcourant son historique et ses préférences pour générer des recommandations de produit personnalisées à l'aide d'un chatbot. Au final, l'utilisation de l'IA générative promet de renforcer l'implication et la fidélisation des clients, d'améliorer la détection des fraudes, et bien plus encore.

Pour démystifier les capacités de l'IA générative et comprendre comment en tirer parti, vous avez besoin d'un guide détaillé qui vous explique comment exploiter vos données. Dans cet ebook, nous vous guiderons tout au long de votre parcours pour faire de vous **un véritable expert de l'IA**. Considérez-le comme une roadmap pour vous aider à révolutionner les résultats de votre entreprise grâce à l'IA générative.

99%

*des entreprises sont convaincues que l'IA générative a le potentiel pour révolutionner leur organisation, aussi bien en interne qu'en externe<sup>2</sup>*

Mais seulement

32%

*des leaders ont confiance dans leur capacité à mettre en œuvre l'IA dans leur entreprise<sup>3</sup>*

<sup>1</sup> Elastic, Cisco crée des expériences de recherche adossées à l'IA avec Elastic sur Google Cloud (2024).

<sup>2</sup> Elastic, Rapport sur l'IA générative d'Elastic (2024).

<sup>3</sup> Russell Reynolds, Embracing the Unknown: How Leaders Engage with Generative AI in the Face of Uncertainty (2024).

# Qu'attendre de ce guide ?

## Étape 0

Déterminez en quoi l'IA générative peut vous aider. Quels résultats visez-vous avec l'IA générative ?

## Étape 1

Définissez les résultats que vous voudriez atteindre dans l'idéal. À quoi une mise en œuvre réussie ressemble-t-elle pour votre cas d'utilisation ?

## Étape 2

Évaluez l'impact et mesurez le succès. Déterminez quels processus organisationnels ont été affectés et comment.

## Étape 3

Sélectionnez une stratégie de mise en œuvre. Étudiez les options qui s'offrent à vous.

## Étape 4

Effectuez un test et procédez par itération.

## Étape 5

Établissez des normes de gouvernance et palliez les préoccupations en matière de sécurité des données.

## Étape 6

Définissez un calendrier avec des jalons pour votre équipe.

Ignorez cette étape si vous savez déjà ce que vous attendez de l'IA générative

**Mais tout d'abord, étudions les bases.**



# Première partie : comprendre l'environnement de l'IA générative

Vous n'avez pas besoin d'être un expert en IA générative pour instrumenter un plan permettant de la mettre en place. Toutefois, en ayant une bonne compréhension des composants en jeu, vous pourrez prendre des décisions avisées et stratégiques tout au long du processus. Commençons par explorer les bases.



## Qu'est-ce que l'IA générative ?

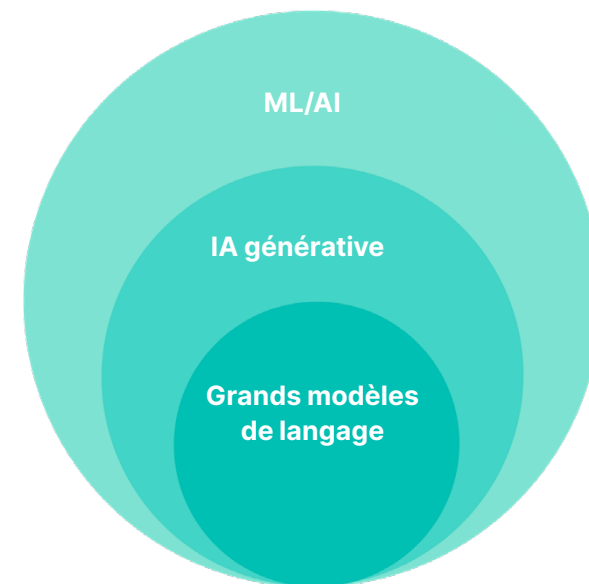
L'intelligence artificielle générative, ou IA générative, fait référence à des modèles de Deep Learning capables de générer des sorties sur invite. Il est important de comprendre que la capacité de génération de cette technologie dépend de son aptitude à prédire les sorties probables d'un point de vue statistique, chose qu'elle peut faire avec l'aide du Machine Learning. **Pour tirer le meilleur parti de l'IA générative, les données sont la clé, aussi bien pour sa mise en œuvre que pour les résultats qu'elle produit.** Nous y reviendrons.

## Qu'est-ce que le Machine Learning ?

Le Machine Learning, ou ML, est un sous-ensemble de l'intelligence artificielle qui exploite des algorithmes pour faire ressortir des connaissances à partir des données. Ces algorithmes analysent les données et "apprennent" les schémas et les similarités, de manière supervisée, semi-supervisée et non supervisée, ce qui leur permet de prendre des décisions. Le Machine Learning est la technologie sous-jacente qui donne à l'IA générative, et notamment aux grands modèles de langage, la capacité à "apprendre" en continu.

## Qu'est-ce qu'un grand modèle de langage (LLM) ?

Un grand modèle de langage, ou LLM, est un modèle informatique qui se trouve en dessous du Machine Learning. Il s'agit d'un type d'IA générative qui s'occupe spécifiquement du langage humain. Après avoir été entraîné sur de vastes ensembles de données de langage principalement publics, un LLM peut exécuter toutes sortes de tâches du traitement du langage naturel (NLP), telles que la reconnaissance, l'analyse, la synthèse, la prédiction, la traduction et la génération de texte. Lors de la mise en place de l'IA générative, les LLM sont l'élément qui permet à l'IA générative de communiquer en langage naturel (ou humain).





## Et si on parlait des hallucinations ?

Une hallucination est un résultat incorrect ou erroné généré par un LLM. Vous avez déjà probablement utilisé ChatGPT et obtenu parfois des réponses contestables. La sortie semble légitime... Mais est-ce réellement le cas ? ChatGPT, par exemple, s'appuie sur un LLM. Or, si le LLM ne parvient pas à trouver la réponse, il a tendance à en créer une. Cet angle mort est important à connaître lors des discussions concernant l'utilisation des LLM dans des applications d'entreprise. Comment vous assurer que les sorties générées soient pertinentes et précises ? C'est là que la génération augmentée de récupération (RAG) entre en jeu.

Vous : Combien de congés payés me reste-t-il ?



IA : Il reste 200 jours dans l'année.

Vous : Comment réparer ma sonnette vidéo qui ne veut pas se connecter au Wi-Fi ?

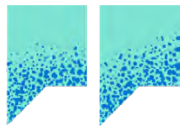


IA : Les meilleures sonnettes vidéo proposent un enregistrement 4K et...

## Qu'est-ce que la génération augmentée de récupération (RAG) ?

Considérez la génération augmentée de récupération, ou RAG, comme votre ligne de défense contre les hallucinations. La sortie générée par un LLM est renforcée, ou "vérifiée", en récupérant des informations à partir d'un ensemble de données spécifique, ou données contextuelles, que vous fournissez. Cette vérification a lieu lors d'une recherche optimisée par une base de données vectorielle. Par exemple, suite à la requête d'un utilisateur, une entreprise peut se servir de la RAG pour faire une recherche dans ses politiques et fournir des réponses pertinentes à un LLM, afin qu'il puisse à son tour répondre aux questions en s'appuyant sur les politiques de l'entreprise. En plus d'être un moyen de défense contre les hallucinations, **la RAG vous permet d'utiliser l'IA générative avec votre ensemble de données propriétaires**. C'est son plus grand avantage.

Lors de la mise en place de l'IA générative dans les applications d'entreprise, la RAG joue un rôle important pour de nombreuses raisons : elle peut fournir des résultats optimaux et vous permettre d'utiliser rapidement votre propre ensemble de données propriétaires. Elle coûte également moins cher que l'entraînement ou la création de votre propre LLM. C'est pourquoi la RAG est la clé pour une intégration de l'IA générative réussie. La RAG va au-delà des limites des LLM généraux pour créer le "moteur de recherche nouvelle génération".



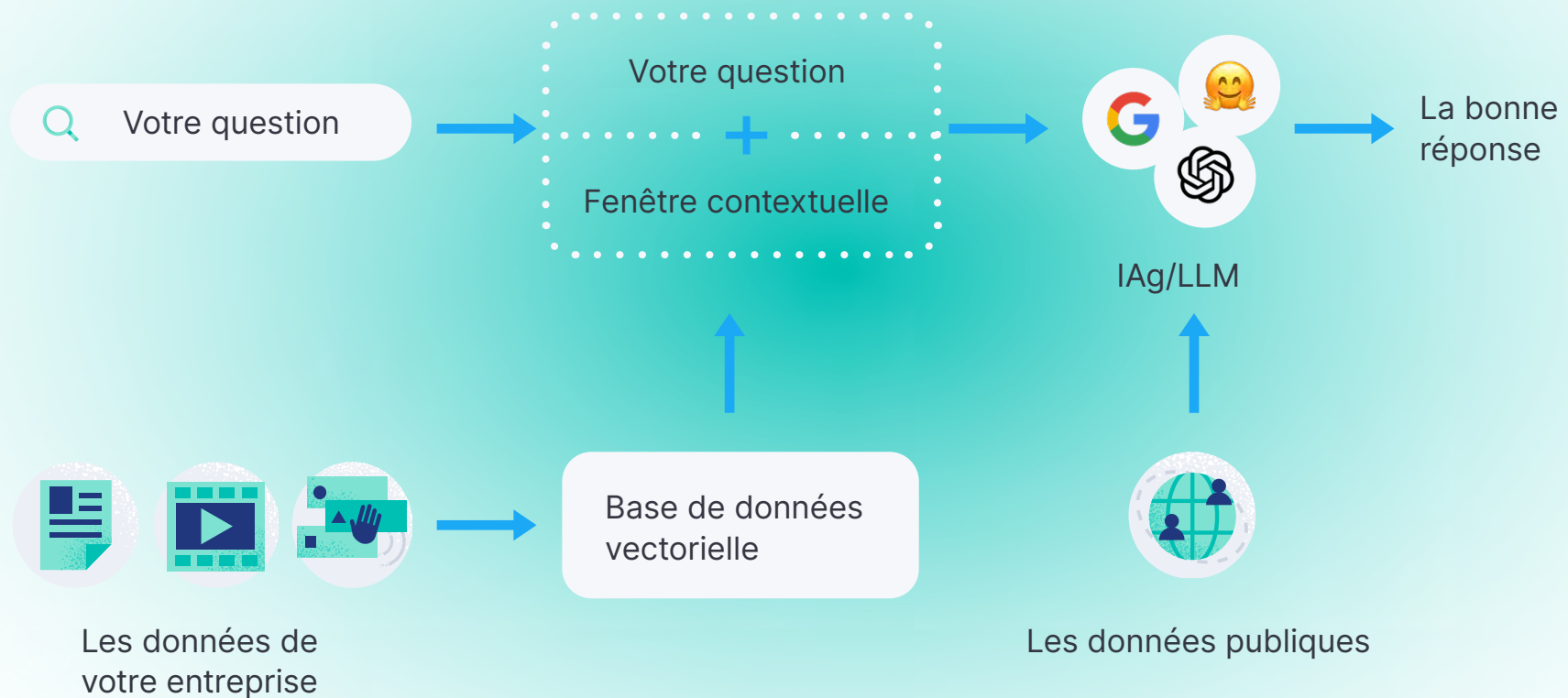
**"La RAG aide à créer le moteur de recherche nouvelle génération."**

— **Baha Azarmi**  
**VP de l'ingénierie client au niveau mondial chez Elastic**



# Génération augmentée de récupération (RAG)

La RAG vous permet d'utiliser l'IA générative avec votre ensemble de données propriétaires.



# La RAG est une nouvelle façon de répondre aux questions des utilisateurs :

## Recherche normale

Une personne recherche **un terme**.

 Politique sur le télétravail

Une requête est exécutée.

```
query = {
  "bool": {
    "should": [
      {
        "text_expansion": {
          "ml.inference.text_expanded_predicted_value": {
            "model_id": "model_id",
            "model_text": "question"
          }
        }
      }
    ]
  }
}
```

Les résultats sont présentés.

	Titre du document	Date d'ajout
1	Code de conduite des employés	01/01/2010
2	Politique sur l'utilisation de l'informatique	01/01/2015
3	Travaux en cours sur la page d'accueil	01/01/2022
	etc.	


L'utilisateur choisit un document et lit son contenu.

Code de conduite des employés

Lorem ipsum Lorem ipsum  
Lorem ipsum Lorem ipsum  
Lorem ipsum Lorem ipsum  
Lorem ipsum Lorem ipsum  
Lorem ipsum

## IA générative sans RAG

Une personne pose **une question**.

 Quelle est notre politique sur le télétravail ?

Une requête est exécutée.


```
query = {
  "bool": {
    "should": [
      {
        "text_expansion": {
          "ml.inference.text_expanded_predicted_value": {
            "model_id": "model_id",
            "model_text": "question"
          }
        }
      }
    ]
  }
}
```

Une réponse est fournie sans le contexte propre à votre domaine.

Une politique sur le télétravail est nécessaire lorsque vous avez des employés qui travaillent de manière hybride ou...

## Avec RAG

Une personne pose **une question**.

 Quelle est notre politique sur le télétravail ?

Une requête est exécutée.

```
query = {
  "bool": {
    "should": [
      {
        "text_expansion": {
          "ml.inference.text_expanded_predicted_value": {
            "model_id": "model_id",
            "model_text": "question"
          }
        }
      }
    ]
  },
  {
    "match": {
      "text": "question"
    }
  }
}
```

Les résultats obtenus servent de contexte.

	Titre du document	Date d'ajout
1	Code de conduite des employés	01/01/2010
2	Politique sur l'utilisation de l'informatique	01/01/2015
3	Travaux en cours sur la page d'accueil	01/01/2022
	etc.	

**Une réponse est fournie à partir des résultats de recherche par un LLM.**

✨ Les employés sont encouragés à faire du télétravail, dans la mesure où ils sont capables... etc.

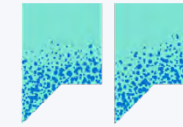
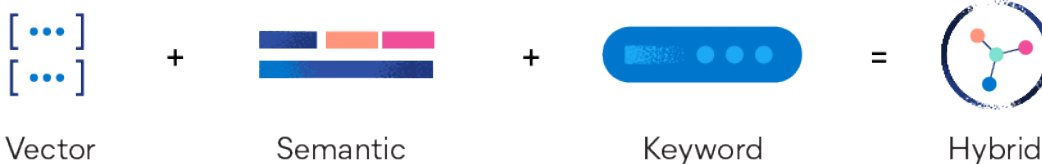
## Qu'est-ce qu'une base de données vectorielle ?

Une base de données vectorielle stocke des plongements vectoriels, c'est-à-dire des représentations numériques de mots, d'images ou de vidéos. Ces plongements ont plusieurs dimensions et permettent d'effectuer une recherche sémantique, qui est un type de recherche déterminant l'intention et le sens contextuel d'une requête. A contrario, une recherche textuelle s'intéresse uniquement aux résultats qui correspondent aux mots-clés indiqués dans la requête.

Dans le cadre de la RAG, une base de données vectorielle permet d'effectuer rapidement une recherche sémantique en fonction de l'invite fournie à l'IA générative. C'est ce qui rend la RAG possible.

L'IA générative gère très bien le NLP, toutefois la recherche traditionnelle par mots-clés est incapable d'assimiler le langage naturel et de proposer les résultats les plus appropriés pour alimenter l'IA générative. De ce fait, lorsqu'on utilise une base de données vectorielle alimentant l'IA générative avec des résultats de recherche similaires d'un point de vue sémantique à l'invite d'origine, l'IA générative peut générer des réponses qui sont plus pertinentes. Considérez la base de données vectorielle comme une banque de connaissances permettant à l'IA générative de répondre aux questions avec des informations précises.

Notons néanmoins que l'IA générative ne se limite pas aux bases de données vectorielles. Grâce à la RAG, l'IA générative peut puiser dans les bases de données relationnelles, dans les bases de données orientées graphe, dans les bases de données orientées document, ou même dans les moteurs de recherche par mots-clés. Le type de base de données qui vous convient le mieux dépend souvent de la nature des données, des algorithmes spécifiques utilisés et des exigences de performances du système. Par exemple, si vous avez des données structurées, privilégiez les bases de données relationnelles. Si vous avez des données avec des relations complexes, optez pour les bases de données orientées graphe. Et si vous souhaitez effectuer une recherche full-text, un moteur de recherche traditionnel répondra parfaitement à vos besoins.



## "Toutes les routes mènent à la recherche hybride."

**Serena Chou, directrice  
de la gestion des produits  
chez Elastic**

Même si la **recherche sémantique** donne des résultats qui correspondent au sens d'une requête, la **recherche par mots-clés** joue toujours un rôle important en faisant coïncider les résultats aux mots-clés exacts des requêtes. La recherche hybride est la pratique qui consiste à utiliser à la fois la recherche vectorielle (souvent dans le cadre de la recherche sémantique) et la recherche par mots-clés afin de fournir les résultats vraisemblablement les plus appropriés à l'IA générative.

**En bref : une solution de recherche hybride fournira probablement les résultats les plus pertinents pour les expériences d'IA générative de votre entreprise.**

# De quoi l'IA générative est-elle capable ?

Nous avons beaucoup parlé des technologies sous-jacentes et des concepts fondamentaux. Mais concrètement, de quoi l'IA générative est-elle capable ?



## Création

En identifiant des schémas dans ses données d'entraînement, l'IA générative peut "créer" ou *générer* des sorties. Elle se sert des données existantes et procède par itération pour produire de nouvelles idées, images, informations, et bien plus encore.



## Synthèse

Grâce à ses capacités de traitement du langage naturel, l'IA générative peut analyser un texte et le synthétiser. Vous avez besoin de passer en revue des documents très longs dans un laps de temps court ? L'IA générative est là pour vous aider.



## Découverte

**La clé de l'IA générative repose sur sa technologie de recherche sous-jacente.** C'est elle qui permet à l'outil d'IA générative de recevoir une requête, de faire des recherches dans un vaste ensemble de données privées ou publiques, puis de produire une réponse.



## Automatisation

Supposons que votre organisation utilise deux plateformes cloud différentes pour des services différents. Ces plateformes généreront chacune des logs aux formats différents. En convertissant automatiquement ces données au même format et en les associant à l'IA, votre équipe peut synthétiser les données et poser des questions à leur sujet en utilisant l'IA générative. Votre équipe informatique peut consacrer ses efforts au monitoring et à la gestion de vos systèmes, plutôt qu'à des tâches laborieuses.

## Étape 0 : Définir la base et déterminer les possibilités

Face aux multiples possibilités qui s'offrent à vous pour générer de la valeur, il est important de définir un point de départ. Comment maîtriser la puissance de l'IA générative pour donner plus de moyens à votre équipe, répondre aux attentes en constante évolution des clients et propulser votre entreprise vers de nouveaux sommets ? En définissant le domaine dans lequel l'IA générative peut générer le plus de valeur pour votre entreprise.

### Posez-vous les questions suivantes :

# 1 Quel problème essayez-vous de résoudre ?

Y a-t-il un domaine dans votre entreprise dans lequel les performances sont particulièrement mauvaises ? Où vos employés passent-ils le plus de temps sur des tâches répétitives ? Cherchent-ils constamment des informations dans des bases de données internes ou des moteurs externes ?



### Par exemple :

Vos employés éprouvent-ils des difficultés à trouver les informations qu'ils recherchent, qu'il s'agisse de mises à jour de projets ou d'informations dans le domaine des RH ? Y a-t-il des tâches que votre équipe de sécurité pourrait automatiser afin de pouvoir se consacrer à une approche plus proactive ? Existe-t-il une certaine désorganisation dans vos processus de service client, du fait que les ingénieurs ne puissent pas toujours communiquer en temps réel avec les agents concernant les problèmes courants et les correctifs récents ?

### Rappel :

Vous devez travailler avec l'équipe dont les workflows seront impactés. Par exemple, si vous décidez de mettre à jour un processus RH inefficace, il est important d'impliquer l'équipe RH dès le départ afin d'obtenir son approbation et son soutien.

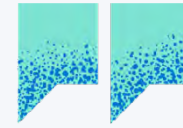
## 2 S'agit-il d'un problème que vous pouvez résoudre avec une base de connaissances ?

Une base de connaissances est une collection de contenus informatifs, par exemple des articles du support technique ou des processus internes.



Réfléchissez aux contenus dont vous disposez et déterminez si ces contenus peuvent servir à résoudre votre problème avec efficacité. Sont-ils suffisants pour automatiser et personnaliser des réponses ?



**Par exemple :** Vous avez constaté que les employés passaient énormément de temps à chercher des informations relatives aux RH. Il n'y a pas d'expert RH dans votre équipe. De ce fait, les employés sont redirigés vers l'intranet de l'entreprise, où ils se retrouvent à devoir lire un nombre incalculable de documents pour savoir, par exemple, combien de jours de vacances il leur reste dans l'année. Pour résoudre ce problème, vous avez besoin d'une base de connaissances (dans le cas présent, il s'agit des documents de politique RH), ainsi que d'un accès aux données personnelles des employés pour personnaliser la réponse à leur fournir.



**"Quand vous tirez un fil, une centaine d'autres viennent aussi."**

**Baha Azarmi,**  
VP de l'ingénierie client au  
niveau mondial chez Elastic

L'étape 0 consiste à comprendre quels sont les processus qui freinent la productivité. En réduisant les tâches sans intérêt, vous donnez plus de temps à vos employés pour qu'ils expriment leur créativité. Une mise en œuvre bien exécutée est une mise en œuvre responsable. **Le fait d'investir dans vos employés, de perfectionner leurs compétences et de planifier des flux de travail et des processus ajustés offre l'assurance d'un mariage réussi avec l'IA générative.**



Au moment de définir votre base, **restez aussi simple et aussi précis que possible**. Pour pouvoir mettre en place l'IA générative, la toute première étape, et non des moindres, consiste à identifier le cas d'utilisation auquel vous souhaitez vous attaquer en premier. Par la suite, des projets de plus petite envergure vous ouvriront la voie vers une mise en œuvre efficace.

Par exemple, dans le scénario RH précédent, l'IA générative pourrait être utile dans de nombreux cas d'utilisation :

1

## Découverte

Un employé saisit une requête dans l'interface : combien de jours de vacances me reste-t-il pour l'année ? Pour répondre, l'IA doit effectuer une recherche et présenter les documents qui concernent la requête, en s'appuyant sur les politiques des RH et le dossier de l'employé.

2

## Synthèse

Si l'on pousse un peu les choses, l'IA générative est capable d'analyser les documents et de les synthétiser pour l'employé sous forme de réponse conversationnelle. "Il vous reste 10 jours de vacances pour cette année et quatre congés mobiles que vous pouvez utiliser. Consultez la page intranet sur les politiques de congés payés pour plus de détails."

3

## Création et automatisation

Un chatbot pourrait faire gagner du temps aux responsables en créant une réponse pour approuver ou refuser une demande de vacances, et même expliquer la décision. Il pourrait aussi créer des invites de calendrier et consigner la demande de congés payés dans le système.



# Adaptation en fonction de votre secteur

Copilote, assistant, robot... L'IA générative prend de nombreuses "formes" et fournit des services précieux qui boostent la productivité dans de nombreux domaines et secteurs. En tant que copilote de sécurité ou d'observabilité, l'IA générative aide les entreprises à gagner en efficacité, à renforcer leurs initiatives de sécurité, à améliorer l'expérience client et à se démarquer plus facilement de leurs concurrents. C'est le cas également lors de l'utilisation d'applications internes et externes.

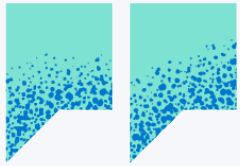
En se servant des données pour permettre à l'IA de fournir des réponses appropriées et pour maîtriser la puissance de la technologie de recherche sous-jacente, **les entreprises peuvent réduire le temps passé sur les tâches répétitives, raccourcir les temps de réponse, et par-dessus tout, améliorer la productivité.** En ajoutant la RAG à l'équation, les entreprises peuvent puiser dans leurs données propriétaires afin que l'IA générative s'en serve dans ses réponses, et ce, d'une manière totalement sécurisée qui respecte les permissions au niveau des documents et au niveau des utilisateurs.

Désormais, vous visez un niveau de rapidité et de pertinence supérieur, tout comme les clients dont les connaissances techniques ne cessent de se renforcer. Vous devez donc vous assurer que vos services répondent à ces attentes. De même, il est essentiel d'utiliser l'IA générative des façons les plus pertinentes qui soient pour l'expérience utilisateur. Il n'y a rien de pire que d'investir dans une technologie tendance que personne n'utilise.

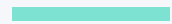
En général, l'IA générative est intégrée dans l'infrastructure informatique d'une entreprise en tant qu'assistant d'IA ou comme copilote de sécurité et/ou d'observabilité.







L'IA générative offre de multiples opportunités aux entreprises, parce qu'il s'agit d'une approche fondamentalement plus humaine et plus intuitive d'obtenir des informations à partir des systèmes.



**Ash Kulkarni**

PDG d'Elastic





## Assistants d'IA

Tirez le meilleur parti des compétences conversationnelles de l'IA générative dans le cadre des interactions des employés et des clients avec des applications internes et externes. Les assistants d'IA fournissent une aide flexible, adaptable et personnelle à chaque utilisateur. Ils peuvent faire office d'experts pratiques, de shoppers personnels, voire même de modules de rappel de l'agenda.



## Copilote de sécurité et d'observabilité

Boostez vos capacités d'observabilité et de sécurité avec les copilotes d'IA générative. Conçus pour travailler en tandem avec les équipes informatiques, les copilotes d'IA générative fonctionnent comme des partenaires experts dans la résolution des problèmes. Par exemple, vous pouvez inviter votre copilote à vous fournir une description détaillée des raisons pour lesquelles une alerte de sécurité a été déclenchée et à vous recommander une procédure pour catégoriser et résoudre l'attaque (par rapport aux attaques précédentes similaires que votre entreprise a rencontrées). Ce type d'invite peut générer un runbook dynamique pour une entreprise.

Grâce à ces intégrations, les entreprises, tous secteurs confondus, peuvent améliorer leur potentiel de personnalisation, d'automatisation et de productivité. On aboutit ainsi à **trois cas d'utilisation majeurs en matière d'IA générative** :

### Amélioration de la résilience opérationnelle

La résilience opérationnelle est essentielle pour garantir le bon fonctionnement des systèmes. Grâce à l'aide de l'IA générative, les équipes informatiques peuvent accélérer l'analyse de la cause première, établir plus de corrélations entre les données sur tous les environnements afin de repérer plus rapidement les problèmes, et avoir un outil de découverte dédié à disposition pour accélérer la résolution. Tous ces éléments mis bout à bout contribuent à maintenir la continuité des opérations.

### Amélioration de l'expérience client

La satisfaction client se trouve au cœur de chaque entreprise. L'IA générative fournit à vos équipes les outils nécessaires pour résoudre les problèmes plus rapidement et récupérer les informations dont elles ont besoin, tout en accordant à vos clients une attention personnalisée et un accès rapide aux informations pertinentes. Résultat ? L'expérience client s'en retrouve améliorée, tout comme les résultats de votre entreprise.

### Atténuation des risques de sécurité

Alors que le monde numérique évolue à un rythme effréné, de nouvelles menaces de sécurité sophistiquées font leur apparition. Pour pouvoir les gérer, il faut des mesures dynamiques et proactives, ainsi que l'expertise nécessaire pour les contrecarrer. L'IA générative vient non seulement soutenir votre équipe de sécurité et vos opérations, mais aussi automatiser les alertes et maintenir une posture proactive.

Dans tous les secteurs, l'IA générative permet de renforcer l'expérience des employés et des clients en proposant des réponses personnalisées, pertinentes et normatives à leurs requêtes. Quel que soit le secteur dans lequel vous travaillez, il y a forcément un moyen d'exploiter l'IA générative pour propulser votre recherche et décupler le potentiel de vos données.



## Télécommunications

Pour les compagnies de télécommunications, on projette que l'IA générative entraînera la création d'une valeur économique de plus de 60 milliards de dollars.<sup>4</sup> Elle permettra aux employés et aux clients d'interroger le site web des compagnies ou leur base de connaissances interne pour obtenir rapidement des réponses personnalisées et pertinentes. Résultat ? Le service client s'en trouve amélioré, tout comme la productivité.

<sup>4</sup> McKinsey, *Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom*, (2024).

### Expérience client

Augmentation du chiffre d'affaires et de la rentabilité

Expérience de recherche renforcée



Recommandations de produit en fonction des requêtes individuelles

- Amélioration de l'expérience client
- Amélioration de la satisfaction
- Augmentation du nombre de transactions par client

Service client automatisé



Support continu en libre-service par chatbots

- Amélioration de la qualité globale du service
- Réduction du temps de réponse aux clients
- Amélioration de la satisfaction et de la rétention

Assistant de base de connaissances



Récupération et synthèse d'informations avec l'IA générative

- Accélération de la prise de décision
- Réduction du temps passé sur les tâches manuelles
- Synthèse et extraction rapides d'informations

Assistant d'IA pour le réseau



Recommandation et résolution proactives des problèmes de réseau

- Amélioration de l'efficacité opérationnelle avec les recommandations de l'IA générative
- Réduction de l'indisponibilité du réseau
- Réduction des coûts de réparation d'urgence



## Services financiers

Grâce à l'IA générative, les entreprises de services financiers peuvent personnaliser davantage l'expérience de leurs clients et de leurs employés. On projette que les améliorations apportées à l'expérience client, à la prévention de la fraude et à l'automatisation devraient entraîner la création d'une valeur économique de plus de 250 milliards de dollars pour le secteur des services financiers.<sup>5</sup>

<sup>5</sup>McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

### Expérience client

Augmentation du chiffre d'affaires et de la rentabilité

Assistant de services bancaires pour le commerce de détail



Récupération et synthèse d'informations avec l'IA générative

- Meilleure visibilité sur les finances personnelles
- Offres sur mesure pour augmenter le taux de conversion
- Amélioration de la satisfaction et de la rétention

Meilleur service client



Recommandation et résolution proactives des problèmes de réseau

- Amélioration de la qualité globale du service
- Réduction du temps de réponse aux clients
- Amélioration de la rétention

Synthèse de la détection de la fraude



Synthèse de la détection des anomalies/des transactions et action recommandée

- Amélioration de la précision et de la vitesse de détection de la fraude
- Réduction des coûts grâce à l'automatisation des tâches
- Réduction des pertes financières

Assistant virtuel



Récupération et synthèse d'informations avec le NLP

- Accélération de la prise de décision
- Réduction du temps passé sur les tâches manuelles
- Synthèse et extraction rapides d'informations

### Expérience employé

Amélioration de la productivité et réduction des coûts et des risques



## Commerce de détail

Pour le secteur du commerce de détail, l'IA générative promet de renforcer la rétention des clients en améliorant la pertinence de la recherche, en recommandant des produits supplémentaires et en envoyant des communications de suivi personnalisées sur les différents canaux. Avez-vous déjà reçu un e-mail du type "Vous n'avez pas validé votre panier" ? L'IA peut automatiser et améliorer ces communications pour proposer des recommandations optimales et permettre une découverte de produits plus personnalisée.

Qu'il s'agisse de mettre en place une expérience client nouvelle génération pour stimuler les ventes en ligne ou de donner plus de moyens aux employés avec les dernières technologies pour améliorer la productivité, on projette que l'IA générative entraînera la création d'une valeur économique de plus de 240 milliards de dollars pour les détaillants.<sup>5</sup>

<sup>5</sup>McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

### Expérience client

Augmentation du chiffre d'affaires et de la rentabilité

Personnalisation de la recherche de produits et de la découverte



Réponse aux questions, expérience de recherche personnalisée

- Amélioration des taux de conversion des sites web
- Augmentation du nombre de transactions par client
- Amélioration de la satisfaction

Meilleur service client



Interactions en libre-service via les chatbots

- Réduction du temps de réponse aux clients
- Amélioration du service, réduction de la perte de clientèle
- Amélioration de la rétention

Meilleur service client



Amélioration de l'expérience des agents et des interactions

- Résolution dès le premier contact
- Intégration plus rapide
- Réduction de la rotation des agents

Maintenance prédictive



Évaluation de l'intégrité des systèmes stratégiques pour hiérarchiser les tâches de maintenance prioritaires

- Réduction de l'indisponibilité des équipements et du système
- Réduction des coûts de réparation d'urgence
- Amélioration de l'efficacité opérationnelle

### Expérience employé

Amélioration de la productivité et réduction des coûts

# Étude de cas : HSE

HSE est l'une des premières marques du secteur européen de live commerce.<sup>6</sup>

**"Pour Home Shopping Europe [HSE], la réussite commerciale commence par la personnalisation et la pertinence du site web."**

**Peter Strasser**

Développeur logiciel chez HSE



## L'opportunité

Comme pour n'importe quelle entreprise d'e-commerce, la fonctionnalité de recherche est essentielle à l'expérience client ainsi qu'aux ventes. HSE doit prendre en compte les parcours client émanant de nombreux canaux, ce qui donne lieu à divers termes de recherche qui reflètent l'endroit où le client a découvert le produit.

HSE s'est servi de l'IA générative et des LLM pour extraire le sens sémantique d'une requête client et générer des résultats qui complètent la correspondance par mot-clé traditionnelle.



## Le résultat

HSE a constaté une **amélioration de 4 % du taux de clic** et une **amélioration de 8 % de la satisfaction client** grâce à une meilleure précision et une plus grande pertinence des résultats de recherche.



## Le conseil

Ciblez un domaine que vous cherchez déjà à améliorer, comme l'expérience de recherche client. Déterminez comment intégrer l'IA générative pour faire passer l'expérience au niveau supérieur grâce à la personnalisation et à la pertinence.

<sup>6</sup> Elastic, "En utilisant Elasticsearch sur AWS, HSE a augmenté la satisfaction des clients tout en réduisant le temps de maintenance de 42 %", (2024).



## Automobile et fabrication

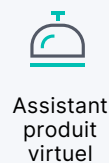
Chaque étape des processus des secteurs de l'automobile et de la fabrication peut être rationalisée avec l'IA. On projette donc la création d'une valeur économique de plus de 170 milliards de dollars.<sup>5</sup> L'IA générative a le potentiel de transformer ces secteurs, depuis la recherche de produit et l'innovation en matière de développement, jusqu'aux stratégies personnalisées de rétention des clients. Qui sait, peut-être qu'un jour les voitures volantes existeront.

<sup>5</sup>McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

### Expérience client

Augmentation du chiffre d'affaires et de la rentabilité

Manuels numériques interactifs



Assistant produit virtuel

- Réponses en temps réel sur les fonctionnalités des produits, la maintenance et le dépannage
- Réduction du nombre de demandes de support technique
- Amélioration de la satisfaction

Meilleur service client



Interactions en libre-service via les chatbots

- Réduction du temps de réponse aux clients
- Amélioration du service, réduction de la perte de clientèle
- Amélioration de la rétention

Optimisation des technologies opérationnelles



Maintenance prédictive : synthèse des problèmes et de leur résolution

- Identification et résolution rapides des problèmes
- Amélioration de l'efficacité opérationnelle et de la prise de décision
- Réduction des coûts de fabrication

Analyse des sentiments concernant les produits



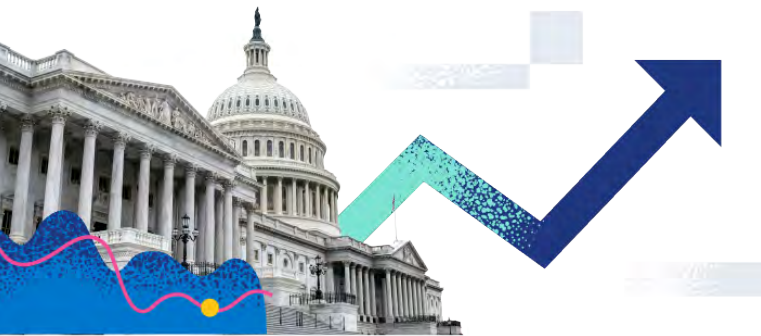
Synthèse de l'appréciation du produit et recommandation d'améliorations

- Amélioration des offres de produits avec les commentaires des clients
- Accélération du retour sur investissement des nouvelles offres de produit

### Expérience employé

Amélioration de la productivité et réduction des coûts





## Secteur public

L'IA générative peut accélérer de manière significative l'obtention de résultats relatifs aux missions menées. Elle peut aussi améliorer les services fournis au grand public et mettre à la disposition des analystes du gouvernement et des professionnels de la sécurité les données appropriées au moment opportun, en reliant l'IA générative aux données des organismes en toute sécurité.



### Réduction des charges de travail

Automatiser les processus et les workflows manuels



### Conformité

Mettre en place un accès aux données basé sur les rôles



### Connaissance de la situation en temps réel

Prendre des décisions plus précises



### Productivité des employés

Trouver les bonnes informations au bon moment



### Expérience du grand public

Établir la confiance via des interactions numériques personnalisées



### Services publics

Améliorer l'accessibilité et les options en libre-service



### Veille dynamique

Accélérer la recherche de mission et les insights



### Cybersécurité

Réaliser une évaluation et une analyse des risques en temps réel

## Applications accessibles au grand public :

- Accès personnalisé aux services publics
- Expérience en ligne rationalisée
- Meilleure accessibilité et options en libre-service

## Applications accessibles aux employés :

- Plus grande précision des renseignements et des investigations
- Amélioration de la productivité grâce à l'automatisation des processus et des workflows manuels
- Plus grande efficacité des processus d'approvisionnement



# Étude de cas : Relativity

Relativity aide les sociétés, les cabinets juridiques et les agences à stocker et à utiliser des données pour la découverte électronique et la recherche juridique.<sup>7</sup>

**"La plus grande problématique à laquelle les clients de Relativity sont confrontés actuellement est l'explosion des données venant de sources hétérogènes. Et elle d'autant plus difficile à appréhender qu'il y a des différences dans les données générées à partir de différents modes de communication."**

**Brittany Roush**  
Directrice des produits



## L'opportunité

Relativity avait besoin de consolider ses données tout en maintenant une approche axée sur la sécurité. Face à l'explosion des données, des sources et de la complexité, la recherche traditionnelle par mots-clés était inefficace. C'est là que la RAG entre en jeu.



## Le résultat

En combinant la RAG et une base de données vectorielle, Relativity a mis en œuvre des expériences de recherche rapides, pertinentes et précises, tirant parti de ses données propriétaires, qu'elle a ensuite pu proposer aux utilisateurs. Sa solution d'IA générative respecte les normes de conformité, telles que PCI, DSS, SOC2 et HIPAA.



## Le conseil

Gardez systématiquement votre échelle en tête. En commençant petit, vous pourrez vous familiariser avec les capacités de l'IA générative et déterminer les applications les plus pertinentes. Une fois que vous aurez identifié le point de démarrage idéal, vos possibilités seront illimitées.

<sup>7</sup>Elastic, Relativity uses Elasticsearch and Azure OpenAI to build futuristic search experiences, today (2024).

Vous comprenez maintenant le potentiel économique immense que représente l'IA générative pour tous les secteurs. Vous avez peut-être même déjà réfléchi à des cas d'utilisation potentiels. Vous savez peut-être aussi par où commencer.



Il n'en reste pas moins que la mise en œuvre de l'IA générative peut sembler ardue et révolutionnaire. Elle suscite des inquiétudes concernant la confidentialité, elle implique de revoir la conformité et elle vient bouleverser la façon dont les gens font leur travail. Une exploitation responsable nécessite une formation, un perfectionnement des compétences et une réorganisation partielle des effectifs.

Malgré ces défis, la valeur que l'IA générative peut apporter à votre entreprise est indéniable. Pour rester dans la course, votre entreprise n'a pas le choix : elle doit la mettre en œuvre. La bonne nouvelle ? Même si vous n'êtes pas tout à fait prêt à passer en production, tous les tests que vous effectuez génèrent de la valeur : voyez-les comme un investissement pour l'avenir de vos systèmes d'IA. En d'autres termes, c'est le moment de vous lancer. Alors, foncez !

# Deuxième partie :

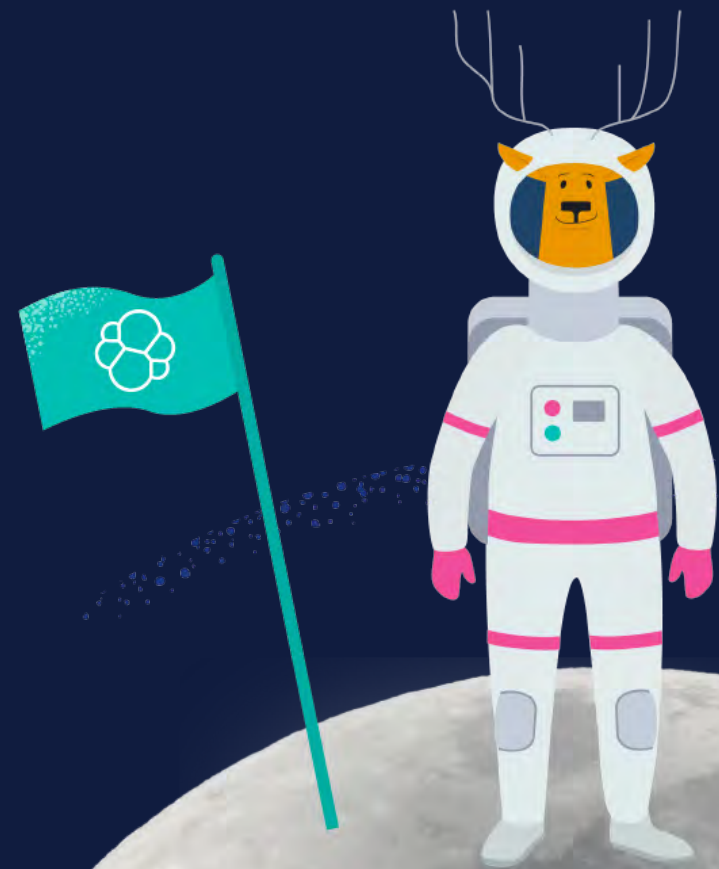
Mise en place de l'IA générative

## Un petit pas pour la machine, un grand pas pour votre entreprise

La mise en place de l'IA générative ne se fait pas en un claquement de doigts. Il s'agit d'une approche itérative qui nécessite une planification et un objectif clair. En commençant avec un projet simple d'IA générative, votre équipe entame son inévitable courbe d'apprentissage et peut travailler sur les processus, affiner les technologies et pallier ses inquiétudes, ce qui lui permet de prendre la voie du succès, et avec elle, votre entreprise tout entière.

**Voici comment.**

Avançons grâce  
à la RAG !



## Étape 1 Définir les résultats que vous voudriez atteindre dans l'idéal

Vous avez un problème que vous souhaitez régler. Vous cherchez à optimiser un processus inefficace. Vous devez à présent réfléchir à la façon dont vos utilisateurs interagissent avec votre solution. Faut-il renforcer une application de recherche ou un chatbot ? Cherchez-vous un nouveau moyen d'interagir avec vos équipes ou vos clients ?

Votre processus de réflexion devrait s'apparenter à ce qui suit :

- ||→ Vous cherchez à renforcer la rétention des clients.
- ||→ Vous avez décidé de mettre en œuvre une application personnalisée de recherche de produits et de découverte.  
  
Vous créez des indicateurs de réussite. C'est ce qui viendra étayer votre base.
- ||→ Vous souhaitez utiliser l'IA générative. Pourquoi ? Pour personnaliser la recherche de produits et la découverte. Pourquoi ? **C'est la réponse à cette question qui correspond aux résultats que vous voudriez atteindre dans l'idéal :** En interagissant de cette façon inédite avec vos données, les clients trouveront les produits dont ils ont besoin sans effort. Ils découvriront aussi des produits qui peuvent les intéresser, en fonction de leur historique de recherche et de leur géolocalisation.  
Résultat : la rétention des clients augmentera.
- ||→ Vous êtes maintenant prêt à mettre en place votre premier projet d'IA générative.



### Posez-vous les questions suivantes :

Quelles actions et quels résultats est-il possible de générer grâce à cette nouvelle façon d'interagir avec vos données ?

La réponse à cette question vous aidera à définir le cadre de vos objectifs. En définissant les résultats que vous voudriez atteindre dans l'idéal, vous pouvez déterminer à partir de quel stade vos performances sont "bonnes" en ce qui concerne votre projet, et de façon plus générale, votre entreprise.

## Étape 2 Déterminer l'impact et mesurer le succès

Pour réussir la mise en place de l'IA générative, vous devez établir un ensemble de KPI qui vous permettront de mesurer ce qu'est une "bonne" performance. Le gain de productivité dû à l'IA générative n'est que l'un des nombreux indicateurs qui sont à votre disposition.

D'autres peuvent s'intéresser à l'augmentation de la satisfaction client, mesurée à l'aide des commentaires dans le cadre du support client, la diminution du nombre de tickets ou encore l'accélération des temps de résolution. Selon le cas d'utilisation que vous testez, vous devez établir les indicateurs de performance correspondants. Il est crucial de les intégrer à chaque étape de votre processus de test pour pouvoir comprendre les progrès que votre équipe et vous réalisez.

### Indicateurs de performance de base

1

#### Impact de la productivité

Mesurez les changements de productivité découlant de votre cas d'utilisation. Comparez le temps nécessaire pour exécuter certaines tâches avec et sans l'utilisation de l'IA générative.

2

#### Scalabilité

Évaluez la capacité qu'a le modèle de scaler par rapport aux augmentations de l'utilisation et de la demande. Ses performances sont-elles toujours aussi fiables et précises ?

3

#### Résultat net

Évaluez l'impact que la mise en œuvre de l'IA générative a eu concernant les coûts. Si vous le souhaitez, vous pouvez y inclure certains indicateurs opérationnels, comme le nombre de plaintes enregistrées ou toute fluctuation des ventes.

4

#### Conformité

Exercez un monitoring continu pour vous assurer que l'IA générative respecte bien les réglementations en matière de confidentialité des données à tout moment.

5

#### Satisfaction client

Examinez les indicateurs opérationnels, tels que la perte de clientèle, l'augmentation des ventes et la fidélité continue à la marque, et étudiez les commentaires des clients.

Servez-vous de ces indicateurs pour déterminer si un projet est réalisable, exploitable, scalable et abordable. Ces indicateurs vous aideront à déterminer votre retour sur investissement et peuvent être élargis par la suite au fur et à mesure que vous étendez vos cas d'utilisation.

## Étape Sélectionner un modèle

3

Comment concevoir une architecture d'IA générative qui répond aux besoins de votre entreprise ? De nombreux facteurs vont influencer vos choix : le coût, la langue, votre écosystème informatique, votre capacité et votre calendrier de déploiement, les réglementations en matière de confidentialité des données et la gouvernance. C'est pourquoi il est préférable de commencer petit, avec un cas d'utilisation simple et spécifique.

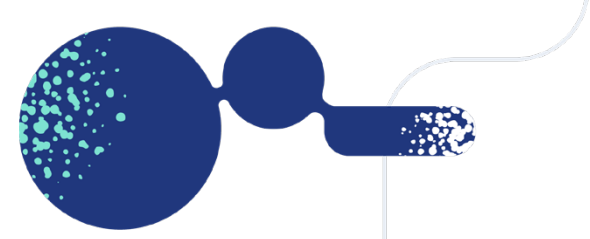
### Pour mettre en place l'IA générative, vous aurez besoin des composants suivants :

- ||→ **Une infrastructure cloud totalement gérée** augmentera l'agilité, optimisera les coûts et diminuera le gaspillage de ressources. Les composants et le matériel évoluent à un rythme effréné. Si vous investissez dans la construction de votre propre data center, celui-ci risque de devenir obsolète d'ici quelques mois.
- ||→ **Un LLM** est le socle permettant à l'IA générative de communiquer en langage naturel et de le comprendre.
- ||→ **Une plateforme de données** incluant la recherche vectorielle, la recherche hybride et la recherche traditionnelle par mots-clés peut être utilisée pour enrichir le LLM avec un contexte approprié issu de vos données propriétaires.
- ||→ **Des API complètes** vous permettent d'enrichir vos données et de les transmettre au LLM et à votre moteur de recherche.

### Les ingrédients nécessaires pour une recherche adossée à l'IA



Vous souhaitez affiner votre propre modèle, utiliser votre propre base de données vectorielle, appliquer votre propre modèle, voire les trois ? Dans tous les cas, la façon dont vous combinez ces différents éléments va déterminer votre schéma de mise en œuvre, et donc influencer sur votre calendrier, sur la complexité de votre test et sur la nécessité (ou non) de renforcer votre équipe.



## Pré-entraîner un LLM

Cette approche mobilise beaucoup de ressources et implique de partir de zéro en entraînant un grand modèle de langage sur un ensemble volumineux de données.

## Affiner un modèle

Cette approche utilise un LLM existant avec votre moteur de recherche et une base de données vectorielle afin d'exploiter vos données en contexte.

## RAG

Ce processus utilise un LLM pré-entraîné existant et différentes techniques pour affiner le modèle afin qu'il réponde à vos besoins.

### Coût

\$\$\$\$

\$\$\$

\$\$

### Durée du déploiement

Déploiement long sur plusieurs mois

Déploiement modéré en quelques semaines

Déploiement rapide en quelques jours

### Confidentialité des données

Avez-vous un ensemble de données suffisamment volumineux pour entraîner le LLM de manière significative ? Si ce n'est pas le cas, vous devrez vous servir de données publiques. Associez-vous données publiques et données privées ?

Avez-vous un ensemble de données suffisamment volumineux pour entraîner le LLM de manière significative ? Si ce n'est pas le cas, vous devrez vous servir de données publiques. Associez-vous données publiques et données privées ?

Cette approche vous permet de préserver le caractère confidentiel de vos données privées.

### Précision et pertinence

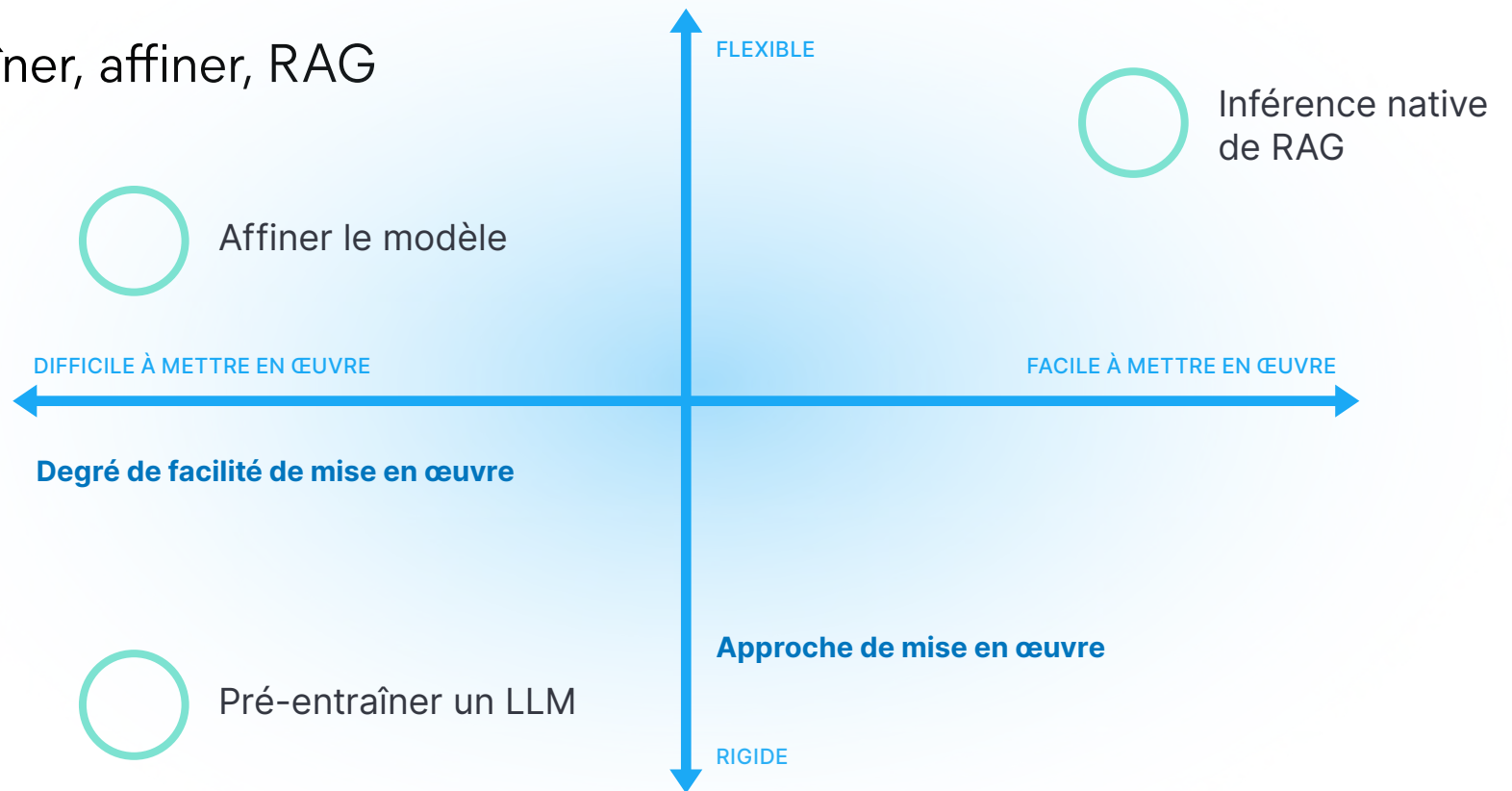
Il est difficile d'assurer la précision et la pertinence.

Il est plus facile d'assurer la précision et la pertinence sur des tâches spécifiques pour lequel le modèle a été affiné.

Le grand avantage de la RAG est sa capacité à "endiguer" les hallucinations en citant les sources sur lesquelles elle se base ou en informant l'utilisateur lorsqu'elle n'a pas la réponse.

Évaluez ces options :

Pré-entraîner, affiner, RAG





## Choisir la bonne méthode pour aller de l'avant

Il n'y a pas une seule façon de procéder. Le plus important, c'est que vous preniez des décisions par rapport aux objectifs définis aux étapes précédentes, et que ces objectifs tiennent compte des besoins des parties concernées. Il vous faut aussi une feuille de route claire à présenter aux parties prenantes et à votre équipe.

À moins que vous ne vouliez procéder à un projet de remaniement conséquent à plus large échelle, la conception d'un LLM à partir de zéro et l'affinement demandent trop de ressources. Vous serez submergé de questions : avez-vous besoin d'une base de données vectorielle pour compléter votre moteur de recherche ? Pouvez-vous mettre à niveau votre moteur de recherche, créer et y stocker des plongements, et établir une logique pour continuer à optimiser votre recherche ? Comment étendre tout cela à un système de recommandation ?

Pour les solutions basées sur le cloud avec la recherche hybride et la recherche sémantique, c'est assez simple. En utilisant un LLM existant, vous pouvez tirer parti de la RAG pour créer une expérience de recherche plus pertinente pour vos clients.

## Voici les aspects à prendre en compte :

||→ Faites le point sur ce que vous avez déjà dans votre environnement informatique. Parfois, il est inutile de remanier l'infrastructure.

||→ Étudiez les solutions prêtes à l'emploi, comme un LLM prêt à l'emploi, une base de données vectorielle prête à l'emploi ou un package tout compris prêt à l'emploi.

**Pour :** les technologies prêtes à l'emploi vous permettent d'être opérationnel rapidement.

**Contre :** votre scalabilité est limitée étant donné que ces technologies sont peu personnalisables.

||→ Trouvez un produit complémentaire qui vous offre de la flexibilité tout en limitant les perturbations. Nous vous conseillons de mesurer la pertinence et la performance de votre recherche, et éventuellement de changer de modèle pour déterminer celui qui vous convient le mieux.

## Étape 4 Faire des essais et apprendre de ses échecs

Étant donné que l'écosystème numérique évolue rapidement, il y a obligatoirement de nombreuses pièces mobiles dans un projet d'IA générative. Le contrôle que vous pouvez exercer sur un LLM pré-entraîné reste limité, tout comme la flexibilité dont vous disposez pour manipuler votre architecture.

L'heure est venue d'appliquer une approche itérative : vous avez votre cas d'utilisation, vous avez établi les résultats que vous vouliez obtenir, vous avez défini des KPI et vous avez réfléchi à la façon de mettre en œuvre votre projet d'IA générative.



### Rappel

La mise en place de l'IA générative a un objectif bien précis : que vous puissiez obtenir des réponses à partir de vos données. Néanmoins, la conformité doit rester une priorité. Le test que vous préparez risque-t-il de compromettre vos politiques en matière de confidentialité ? Ou présente-t-il peu de risques ?

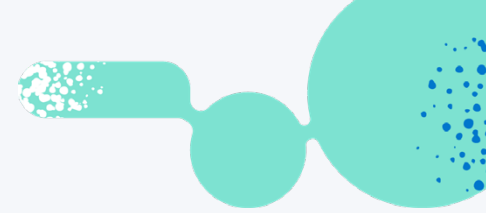


## À ce stade, vous cherchez à...

- ||→ **Mettre en place une boucle de feedback :** définissez qui fait remonter quoi à qui, et identifiez les parties prenantes du projet.
- ||→ **Enrichir votre LLM :** assurez-vous que votre LLM a accès aux informations appropriées stockées dans une base de données vectorielle. Une base de données vectorielle vous permet de fournir les informations les plus pertinentes pour enrichir votre LLM.
- ||→ **Affiner l'expérience utilisateur :** travaillez sur une interface utilisateur conviviale et faites des tests. Le but ultime de l'IA générative est de servir vos employés et vos clients. Il est donc essentiel de mettre en place une interface qui soit adaptée à l'application et aux utilisateurs pour garantir une mise en œuvre réussie du projet d'IA générative et pour en assurer la scalabilité.
- ||→ **Établir une architecture de référence qui peut scaler :** pendant que vous testez votre projet d'IA générative, gardez un œil sur la vue d'ensemble. À quoi votre architecture ressemblera-t-elle lorsque vous scalerez le projet et que vous l'étendrez à d'autres cas d'utilisation ?

Par exemple, si vous trouvez qu'il serait trop lourd de créer une base de données vectorielle de toutes pièces, sélectionnez-en une qui soit téléchargeable (eh oui, ça existe !). Grâce à cette base de données vectorielle, vous pouvez passer à l'étape suivante : la recherche hybride. Servez-vous de la recherche sémantique dans vos applications de recherche pour tester votre prototype de projet d'IA générative nouvelle génération. Grâce à cet exemple, nous avons vu comment il était possible de commencer petit et de procéder ensuite par itération.





## Étape Gouvernance et opérations.

5

Les projets d'IA générative s'accompagnent de leur lot de défis : entre la confidentialité des données, la conformité, les considérations éthiques, le contrôle de la qualité et la gestion des risques, il n'est pas toujours simple de s'y retrouver. Vous devez anticiper les obstacles potentiels et vous assurer que le projet coïncide avec les objectifs de votre entreprise.

### Concernant la gouvernance et les opérations, plusieurs éléments sont à prendre en compte :

- ||→ **Gestion des coûts** : la facturation se fait sur la base de mille tokens. Une partie des coûts est imputable aux invites, l'autre, aux réponses.
- ||→ **Logging** : vous devez consigner chaque réponse pour pouvoir voir la communication entre votre modèle et votre client, et de là, contrôler la qualité.
- ||→ **Sentiment de la réponse** : déterminez le sentiment des réponses du LLM, pour vous assurer qu'elles soient conformes au ton employé par l'entreprise (il s'agit là d'une autre étape importante de contrôle de la qualité).
- ||→ **Monitoring des hallucinations** : les hallucinations peuvent être des informations incorrectes ou erronées, ou encore des discours haineux ou un comportement antisocial émanant d'un chatbot.
- ||→ **Signalement des réponses peu concluantes** : dans le cadre du contrôle qualité, il est essentiel de monitorer la qualité et la pertinence des réponses. C'est là l'occasion de déterminer les applications qui nécessitent une intervention humaine plus importante que d'autres, et de planifier en conséquence le moment approprié pour scaler.

## IA et préjugés

Les modèles d'IA générative s'appuient sur les données à partir desquelles ils ont été entraînés. Si les données d'entraînement contiennent des préjugés ou des limitations, ceux-ci seront reflétés dans les sorties.

Les organisations peuvent atténuer ces risques en choisissant avec soin les données qui serviront à entraîner leurs modèles ou en utilisant des modèles spécialisés et personnalisés adaptés à leurs besoins. Ceci étant dit, les personnes qui programment cette technologie ou organisent les données sur lesquelles le modèle est entraîné ont elles aussi des préjugés.

Les préjugés sont difficiles à éradiquer, et ce, quel que soit le contexte. Cela ne signifie pas pour autant que les organisations doivent ne rien faire pour pallier cette problématique, bien au contraire. Elles doivent sensibiliser les utilisateurs sur le sujet et les inviter à faire preuve d'esprit critique.

Attendez-vous également à ce que votre équipe juridique participe au processus, et veillez à prendre en compte son travail dans votre démonstration de faisabilité. Même si l'implication de cette équipe peut vous sembler être un frein à la phase de test, il est primordial de mettre en place des processus d'examen rigoureux et efficaces pour que la mise en œuvre se fasse de manière responsable, éthique et conforme.

## Ce qu'il faut savoir à propos de la sécurité des données

Chaque jour, de nombreuses menaces planent sur les entreprises. La sécurité des données est donc essentielle. Vos clients vous font confiance dans la gestion de leurs données. C'est pourquoi de nombreuses entreprises ont opté pour un framework Zero Trust. Ce dernier se base sur le principe selon lequel les utilisateurs et les dispositifs ne doivent jamais être automatiquement ou implicitement approuvés, que ce soit à l'intérieur ou à l'extérieur du périmètre réseau d'une entreprise.



## Pour optimiser la sécurité, vous pouvez :

1. **Appliquer la RAG** : les modèles RAG tirent parti des mécanismes de récupération pour mieux comprendre le contexte de l'invite d'entrée, et ainsi, proposer des réponses appropriées qui tiennent compte du contexte tout en omettant les détails sensibles. En utilisant la RAG avec une plateforme qui dispose d'une sécurité au niveau des documents et basée sur les rôles, les permissions sont à coup sûr respectées.
2. **Investir dans une solution d'observabilité ou étendre celle dont vous disposez** : gérez la question de la confiance. Suivez la piste des données grâce aux fonctionnalités de monitoring, et monitoriez les réponses produites par votre projet d'IA générative. Où vont vos données et que dit l'IA générative à vos clients ?

Pour finir, la mise en place de l'IA générative dans votre écosystème implique que vous établissiez de nouveaux protocoles opérationnels, et par conséquent, de nouvelles politiques. Grâce à une meilleure efficacité des processus et une augmentation du chiffre d'affaires, le temps économisé sur les tâches peu stratégiques peut être réinvesti dans ces efforts.

## Étape Établir une chronologie et des jalons

6

Définissez une chronologie. Par exemple, un trimestre. Sur ce laps de temps, établissez des jalons à 30 jours et à 90 jours. Tirez parti de cette période pour prouver la valeur de votre cas d'utilisation boosté par l'IA générative.



**Au 30e jour, votre objectif est d'avoir lancé votre premier test. Comment cela se présente-t-il ?**

- Vous avez choisi un cas d'utilisation.
- Vous avez affecté une petite équipe à la tâche.
- Vous avez organisé des sessions de formation selon les besoins.
- Vous avez établi les résultats à atteindre.
- Vous avez conçu un prototype d'interface.



**Au 90e jour, vous êtes prêt à lancer votre premier cas d'utilisation. Comment cela se présente-t-il ?**

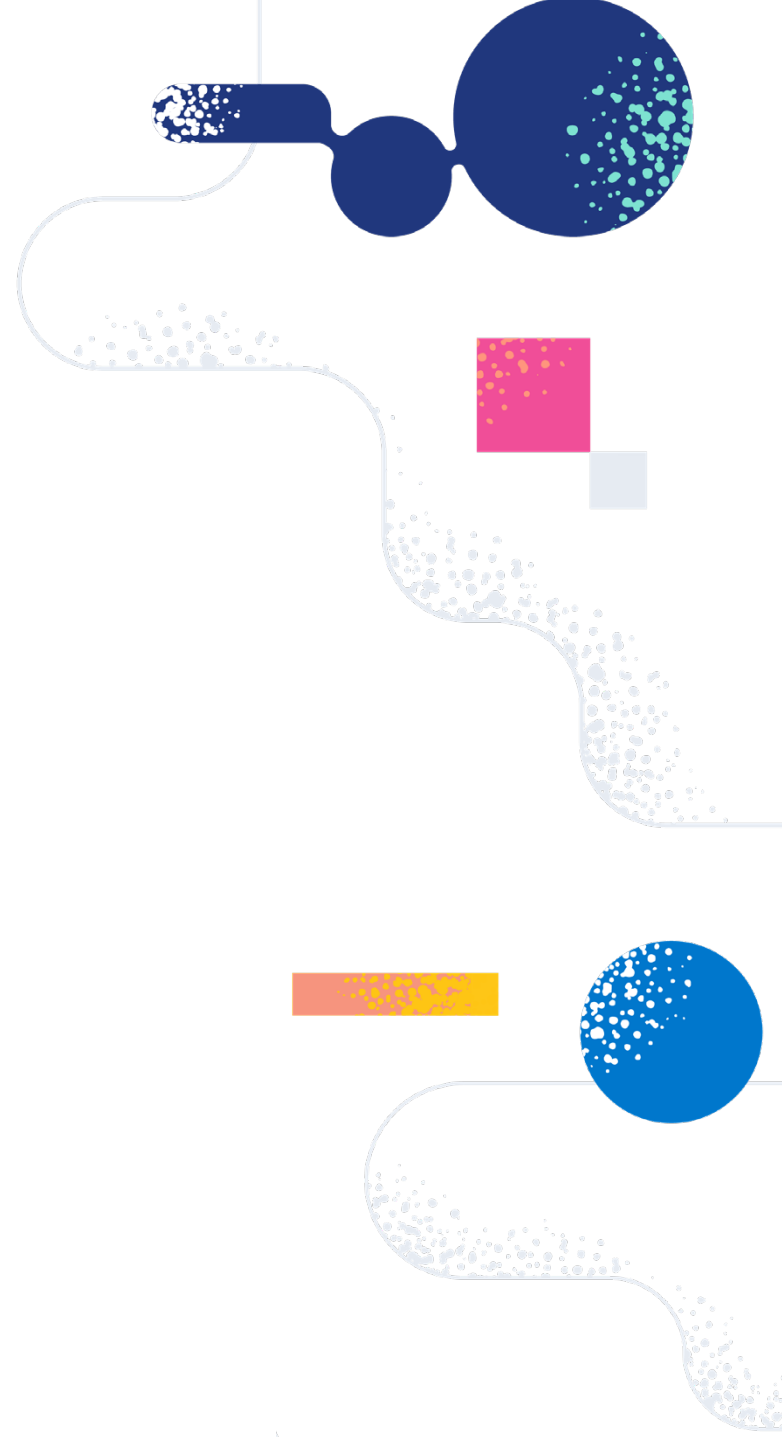
- Vous avez rendu le test accessible à quelques personnes en interne.
- Vous avez testé, affiné et mesuré les sorties générées.
- Vous avez monitoré en continu la façon dont les utilisateurs interagissent avec l'interface.
- Vous avez établi un ensemble de directives concernant les sorties de qualité.
- Vous avez collecté des données sur certains indicateurs clés de performance.
- Vous avez mesuré la valeur de l'initiative.

Ces tâches devraient servir de référence. Les besoins propres à votre entreprise, c'est-à-dire la constitution de votre équipe et les technologies avec lesquelles elle travaille ou qu'elle ajoute à votre suite, ont une incidence sur la vitesse avec laquelle vous pouvez déployer votre premier cas d'utilisation et collecter des insights.

## À ce stade, prenez en compte les éléments suivants :

1. **Taux d'erreur** : mesurez le taux d'erreur. L'IA générative produit-elle des sorties correctes et pertinentes ? Selon la réponse à cette question, affinez l'IA générative en conséquence.
2. **Durée d'entraînement et coût** : mesurez la durée et les ressources nécessaires pour entraîner votre modèle. Cela vous permettra d'optimiser votre période de test, et de là, mettre le projet en place plus rapidement.
3. **Intervention humaine nécessaire** : l'IA générative fonctionne-t-elle uniquement avec un humain dans la boucle ? Quel est le degré de supervision nécessaire pour maintenir la fiabilité et la précision ?
4. **Temps de réponse et qualité des sorties** : mesurez la rapidité avec laquelle l'IA générative fournit des sorties et comparez la qualité des sorties par rapport à un ensemble de règles ou de directives établies.

**Et en procédant ainsi, vous jetterez les bases de votre réussite et vous pourrez la décupler.**





# Le début d'une nouvelle ère

De nombreux leaders du secteur récoltent déjà les fruits de l'IA générative. Et bien d'autres essayent de reproduire ce succès et de garder le rythme face à l'évolution des attentes des clients. L'innovation en matière d'IA générative avance rapidement. Mais vous n'irez nulle part si vous n'avez pas les bases.

Pour tirer le meilleur parti de vos données, vous devez déterminer les moyens qui correspondent le mieux à votre stratégie pour mettre en œuvre l'IA générative. Ainsi, vous ne vous laisserez pas distraire par des innovations qui, certes, peuvent paraître excitantes, mais qui ne vous seront d'aucune utilité. Pour mettre en place l'IA générative de manière optimale, le mieux est d'effectuer une mise en œuvre par étapes en y consacrant du temps et des ressources spécifiques. La clé pour obtenir un retour sur investissement maximal est d'intégrer les nouvelles technologies à bon escient. Qui plus est, en personnalisant et en adaptant les outils d'IA en fonction de vos besoins opérationnels, vous êtes sûr de bénéficier d'une pertinence et d'une efficacité optimales. Ce qui est, au final, le but premier de l'IA générative.

Gardez à l'esprit qu'il est nécessaire de procéder à une mise en œuvre responsable de l'IA générative, c'est-à-dire qui protège la confidentialité des données et la sécurité, qui gère la sensibilité et qui soit éthique. En plus d'ajouter des milliers de milliards de dollars à l'économie mondiale, l'IA générative offre l'opportunité à tout un chacun de l'utiliser pour réaliser des tâches et de perfectionner les compétences des travailleurs. Positionnez-vous en tant que pionnier dans l'utilisation de l'IA générative, mais aussi en tant que pionnier dans le développement de nouveaux processus opérationnels pour votre entreprise.





# À vos marques, prêt, partez !

Pour identifier votre premier cas d'utilisation exploitant l'IA générative, ce sont toutes vos équipes qui doivent mettre main à la pâte. Il s'agit d'un effort collectif. Vous aurez besoin de votre équipe de sécurité, de votre équipe informatique, de votre équipe de développement et de votre équipe du secteur, et ce, dès le départ. Mais vous et vos équipes n'êtes pas seuls. **Elastic est là pour vous aider.**

## Avec votre équipe de sécurité

Vous pouvez booster la productivité des professionnels et réduire les risques. Pour mettre en place l'IA générative dans vos cas d'utilisation de sécurité, il vous faut en premier lieu [une approche unifiée sur une plate-forme ouverte](#). Vous pouvez tirer parti de la puissance de l'IA générative avec Elastic Security et créer une expérience adaptée aux besoins de votre équipe de sécurité.

Découvrir Elastic Security

## Avec votre équipe SRE et informatique

Mettez à la disposition de vos SRE et de vos ingénieurs une interface interactive de tchat en langage naturel sur laquelle s'appuyer pour repérer rapidement les informations les plus pertinentes. Découvrez comment associer l'IA conversationnelle au Machine Learning avancé et à Elastic Observability pour permettre des [conversations par tchat interactives et contextuelles](#) basées sur vos runbooks et vos données propriétaires.

Découvrir Elastic Observability

## Avec votre équipe de développement

Vous pouvez améliorer la boîte à outils de votre équipe de développement afin qu'elle puisse fournir un support client au moyen d'options en libre-service, par exemple avec des chatbots hautement personnalisés avec Elastic Search. Équipez vos agents du service client avec les mêmes outils de recherche efficaces pour qu'ils puissent résoudre les cas facilement. Mettez notamment l'IA générative à leur disposition pour qu'ils puissent trouver des réponses parmi des sources de données disparates. Découvrez comment [mettre en place une recherche puissante dans votre base de connaissances](#).

Découvrir Elastic Search

