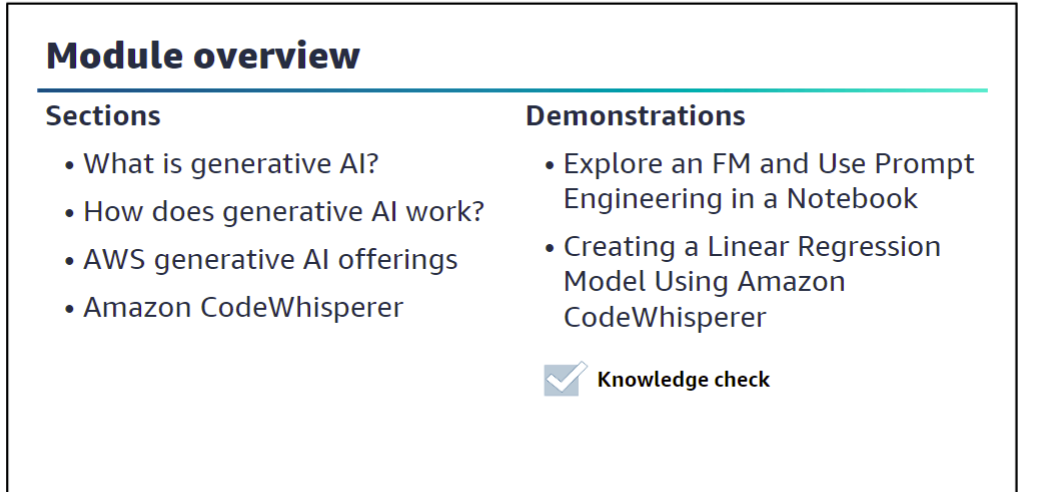
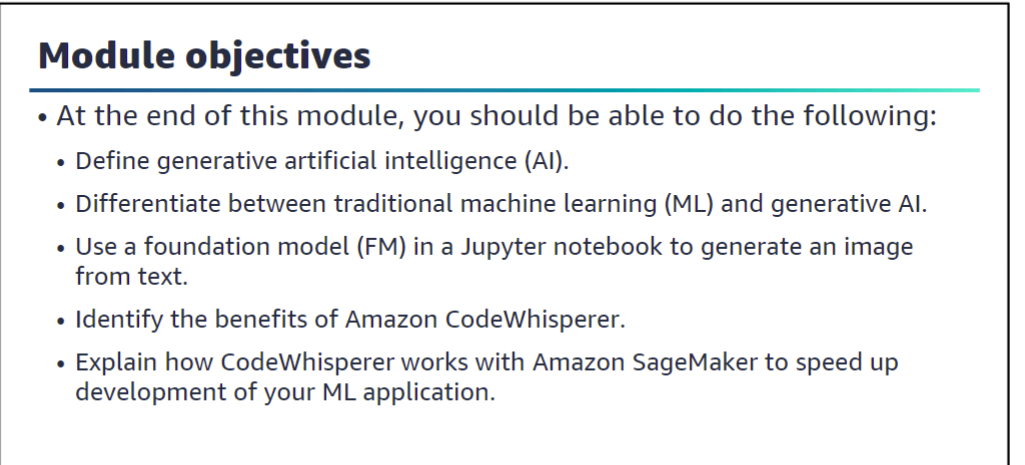
**Module 7: Introduction to Generative AI.**

****

Ce module présente l'IA générative. Les sujets suivants sont abordés dans ce module : • Qu'est-ce que l'IA générative ? • Comment fonctionne l'IA générative ? • Offres d'IA générative d'AWS • Amazon CodeWhisperer

Il vous sera demandé de compléter une évaluation des connaissances qui testera votre compréhension des concepts clés abordés dans ce module.



À la fin de ce module, vous devriez être capable de faire ce qui suit :

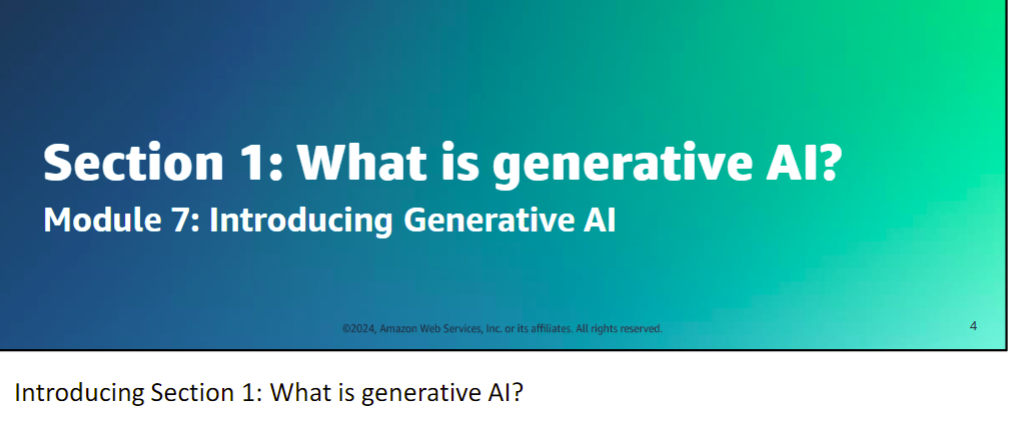
• Définir l'intelligence artificielle (IA) générative.

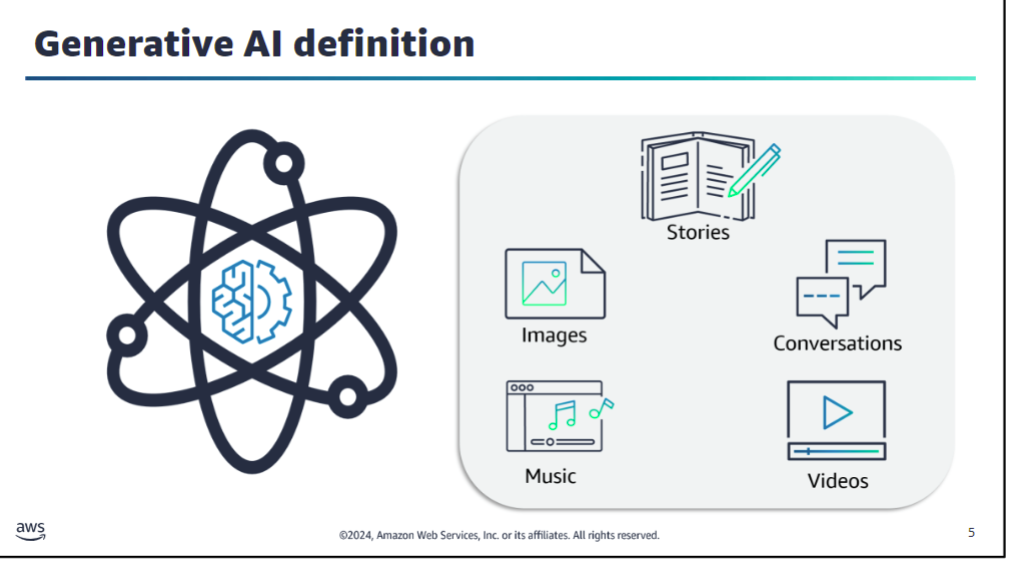
• Différencier l'apprentissage automatique (ML) traditionnel et l'IA générative.

• Utiliser un modèle de base (FM) dans un notebook Jupyter pour générer une image à partir d'un texte.

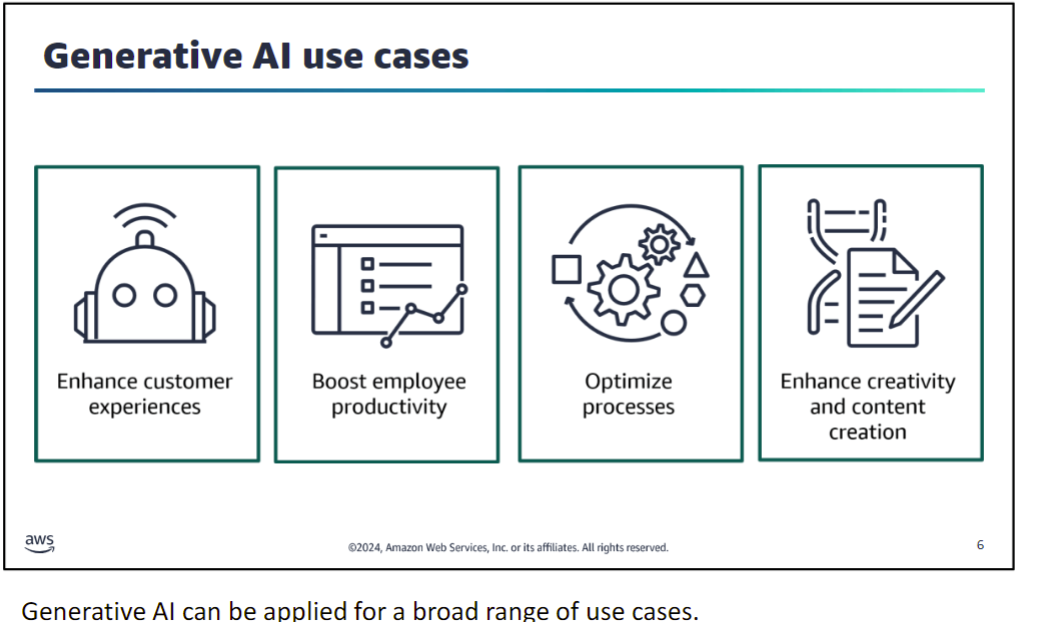
• Identifier les avantages d'Amazon CodeWhisperer.

• Expliquer comment CodeWhisperer fonctionne avec Amazon SageMaker pour accélérer le développement de votre application ML.





L'IA générative est un type d'intelligence artificielle qui peut créer de nouveaux contenus et idées, notamment des conversations, des histoires, des images, des vidéos et de la musique. Les générateurs d'IA sont alimentés par des modèles d'apprentissage automatique préentraînés appelés modèles de base (FM). Ces modèles sont capables de produire du contenu à votre place. Ce contenu généré par l'IA peut être modifié, vous permettant ainsi de faire les ajustements nécessaires pour répondre à vos besoins.

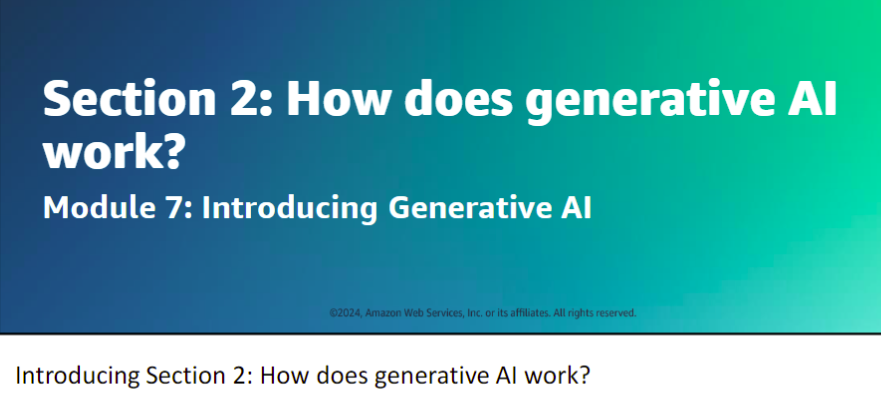


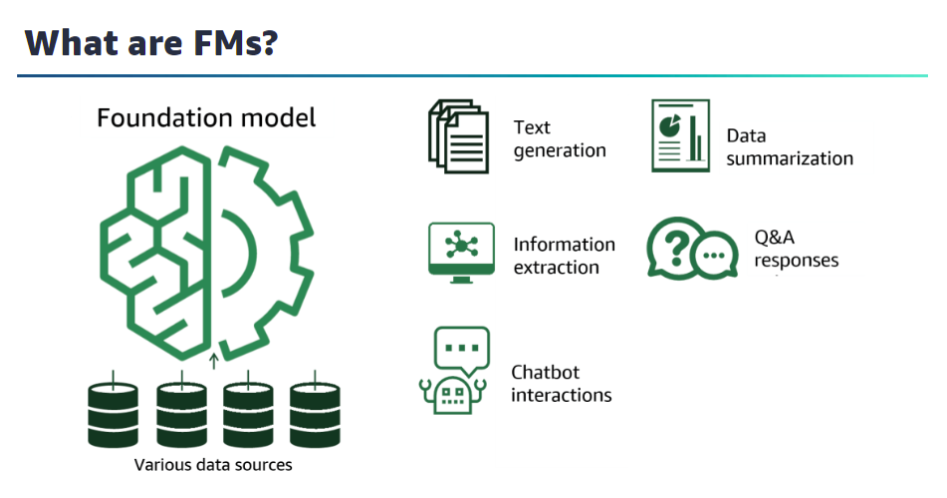
**Améliorer l'expérience client** : Les chatbots et les assistants virtuels peuvent améliorer les expériences de service client en simplifiant le libre-service client et en réduisant les coûts opérationnels. L'IA générative peut également aider à résumer et analyser les conversations avec les clients, ainsi qu'à créer des expériences client plus personnalisées.

**Augmenter la productivité des employés** : L'IA générative peut améliorer la productivité des employés en accélérant le développement d'applications, en automatisant la génération de rapports et en améliorant les recherches de contenu via une interface conversationnelle.

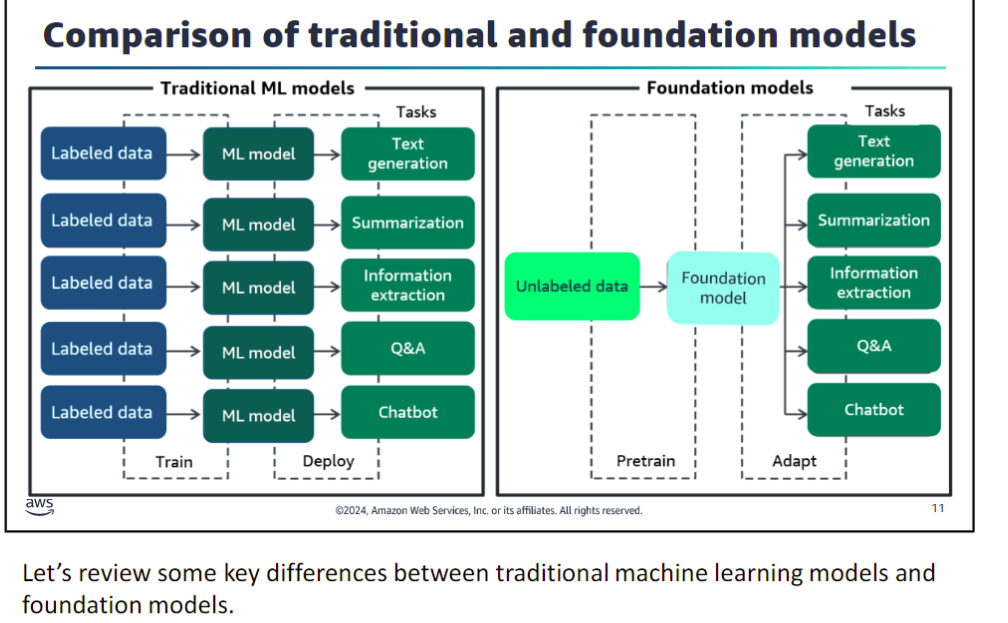
**Optimiser les processus** : Les opérations commerciales peuvent être optimisées grâce au traitement intelligent des documents, qui extrait et résume les données via des questions-réponses alimentées par l'IA générative. L'IA générative peut également aider à améliorer la logistique de la chaîne d'approvisionnement et à augmenter les données en générant des données synthétiques pour entraîner les modèles d'apprentissage automatique.

**Améliorer la créativité et la création de contenu** : Les médias générés par l'IA peuvent renforcer la créativité et la création de contenu. Le contenu marketing créé à l'aide de l'IA générative, comme les articles de blog ou les mises à jour sur les réseaux sociaux, peut faire gagner du temps et des ressources. Le contenu de vente, les conseils et les outils de soutien peuvent également être générés pour correspondre au profil et au comportement d'un prospect, améliorant ainsi les taux de réponse. Les développeurs de produits peuvent utiliser l'IA générative pour générer plusieurs prototypes de conception basés sur certains paramètres et contraintes, accélérant ainsi la phase d'idéation.

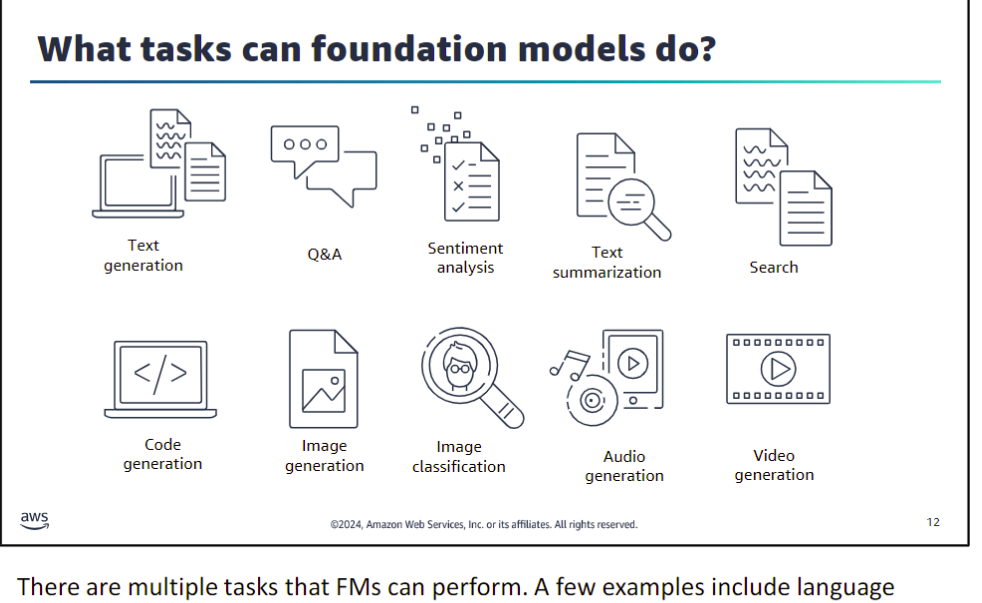




Au cœur de chaque générateur d'IA se trouve un modèle de base. Les modèles de base sont une classe de modèles d'apprentissage automatique puissants qui se distinguent par leur capacité à être préentraînés sur de vastes quantités de données pour accomplir un large éventail de tâches en aval. Ces tâches incluent la génération de texte, la synthèse de données, l'extraction d'informations, les réponses aux questions et les interactions avec des chatbots. Ils peuvent également générer des images, de la musique, des vidéos ou du code. En revanche, les modèles d'apprentissage automatique traditionnels sont entraînés pour accomplir une tâche spécifique à partir d'un ensemble de données.

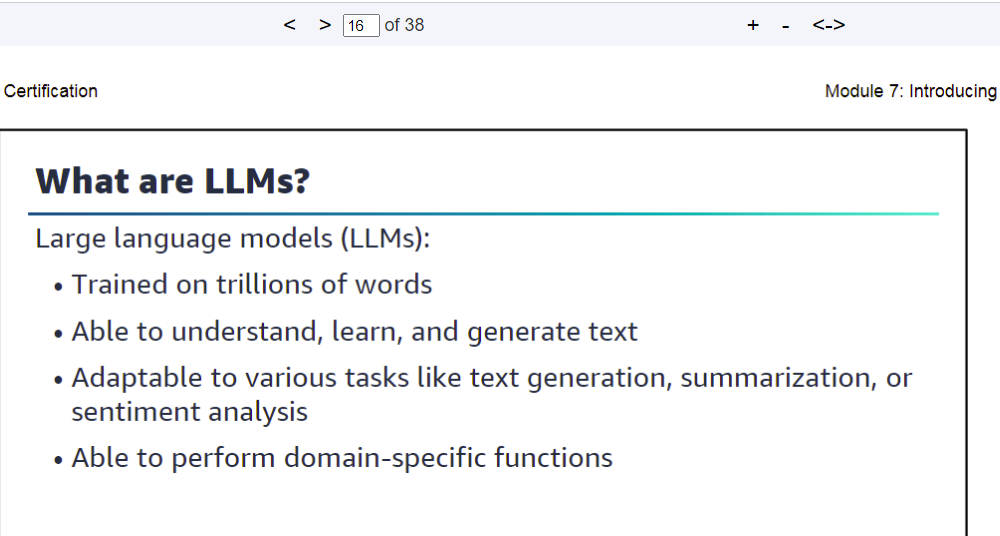


Examinons certaines des principales différences entre les modèles d'apprentissage automatique traditionnels et les modèles de base. Les modèles de ML traditionnels sont généralement isolés pour accomplir des tâches spécifiques, comme l'analyse de texte pour le sentiment, la classification d'images ou la prévision de tendances. Pour accomplir chaque tâche, les clients doivent recueillir des données étiquetées, entraîner un modèle et déployer ce modèle. La taille et la nature polyvalente des modèles de base les rendent très différents des modèles de ML traditionnels. Avec les modèles de base, au lieu de rassembler des données étiquetées et d'entraîner plusieurs modèles, vous utilisez le même modèle de base préentraîné pour adapter plusieurs tâches. Les modèles de base peuvent également être personnalisés pour exécuter des fonctions spécifiques à un domaine qui différencient les entreprises. Cela utilise seulement une petite fraction des données et de la puissance de calcul requises pour entraîner un modèle à partir de zéro.

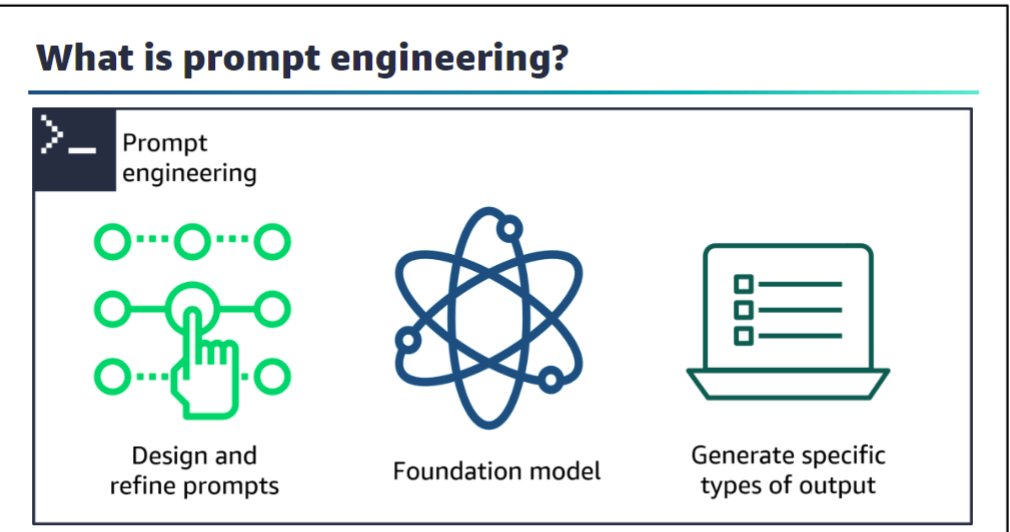


Il existe plusieurs tâches que les modèles fondamentaux (MF) peuvent accomplir. Quelques exemples incluent le traitement du langage, la compréhension visuelle et la génération de code. Un exemple de modèle fondamental capable de traiter des tâches en langage naturel est Amazon Titan. Stable Diffusion est un autre exemple de MF qui peut effectuer des tâches de compréhension visuelle. Pour plus d'informations sur ce que les modèles fondamentaux peuvent faire et pour d'autres exemples, consultez Qu'est-ce qu'un modèle fondamental ?at

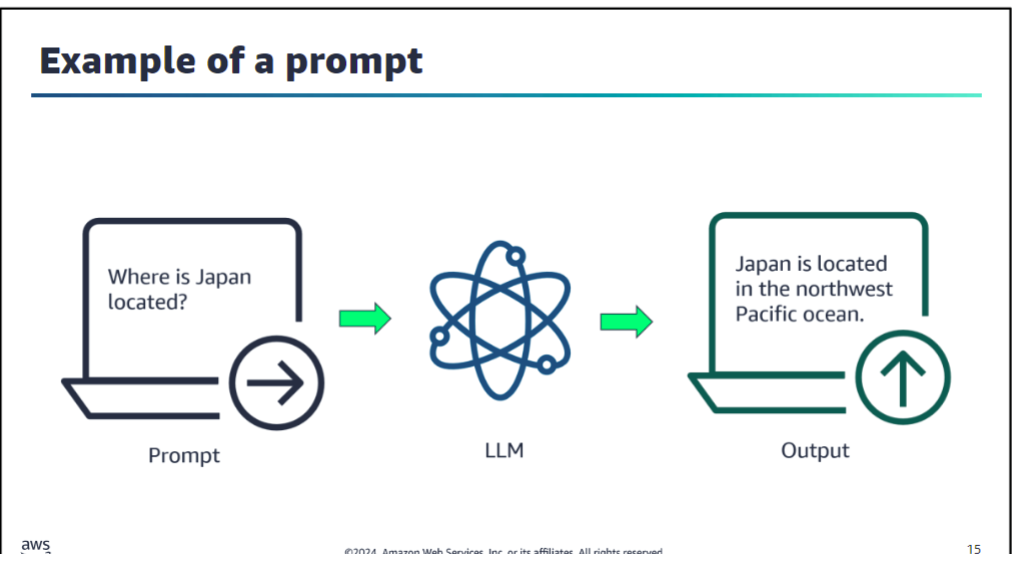
<https://aws.amazon.com/what-is/foundation-models/>



Un sous-ensemble de modèles de base, appelés modèles de langage de grande taille (LLM), sont entraînés sur des milliers de milliards de mots à travers de nombreuses tâches en langage naturel. Ces LLM peuvent comprendre, apprendre et générer du texte qui est presque indiscernable de celui produit par des humains. Avec les LLM, au lieu de rassembler des données étiquetées pour chaque modèle et d'entraîner plusieurs modèles, les clients peuvent utiliser le même LLM préentraîné pour adapter diverses tâches telles que la génération de texte, la synthèse ou l'analyse de sentiment. Les LLM peuvent être personnalisés pour ajouter des connaissances spécifiques à un domaine, comme des documents internes pour une entreprise particulière.



Les architectures d'IA générative utilisent l'ingénierie de prompt pour initier une action à partir d'un modèle de base. L'ingénierie de prompt est le processus de conception et de raffinement des prompts ou des stimuli d'entrée pour qu'un modèle de langage génère des types spécifiques de sortie. L'ingénierie de prompt consiste à sélectionner des mots-clés appropriés, à fournir du contexte et à façonner l'entrée de manière à encourager le modèle à produire la réponse souhaitée. C'est une technique essentielle pour façonner activement le comportement et la sortie des modèles de base.



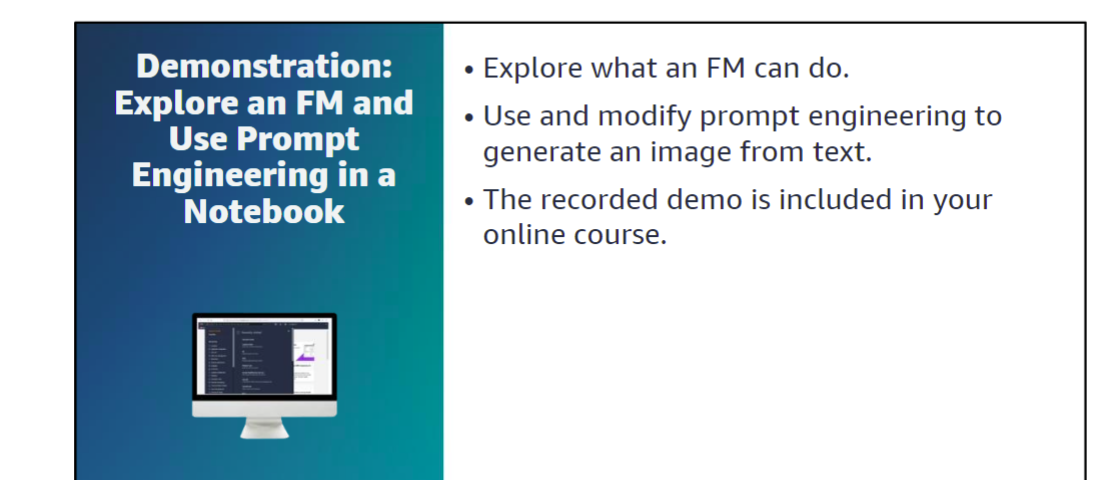
Un **prompt** est le texte que vous saisissez dans le modèle. Le modèle génère ensuite une réponse au prompt. Cela est appelé **inférence**. Le résultat de l'inférence est affiché sous forme de sortie. Dans cet exemple, votre prompt est : « Où se trouve le Japon ? ». L'inférence du modèle sur le prompt renvoie une sortie qui dit : « Le Japon se trouve dans le nord-ouest de l'océan Pacifique. »



La structure d'un **prompt** dépend de la tâche que vous souhaitez que le modèle exécute. Un prompt est composé des éléments suivants :  
• **Instructions** : Il s'agit d'une tâche que le modèle de langage étendu doit accomplir. Cela fournit une description de la tâche ou des instructions sur la manière dont le modèle va la réaliser.  
• **Contexte** : Il s'agit d'informations externes pour guider le modèle.  
• **Données d'entrée** : Ce sont les données pour lesquelles vous souhaitez obtenir une réponse.

Le modèle générera ensuite une sortie.  
• **Sortie** : C'est le type ou le format de la sortie.

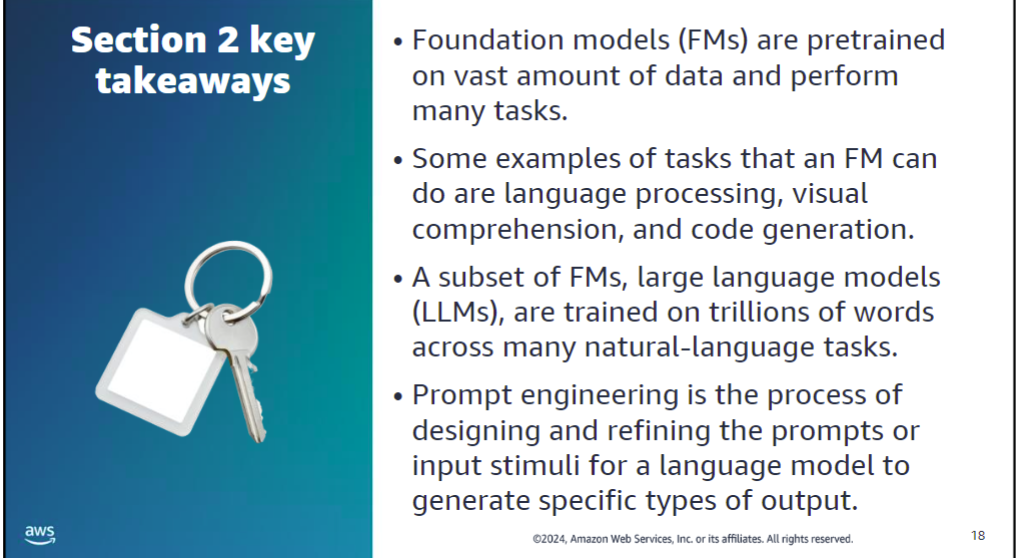
Si le modèle ne vous donne pas la sortie souhaitée, vous devez modifier le prompt ou fournir au modèle des exemples de tâches dans le prompt. Cela s'appelle **l'ingénierie de prompt**



This demonstration explores foundation models in Amazon SageMaker Studio. This

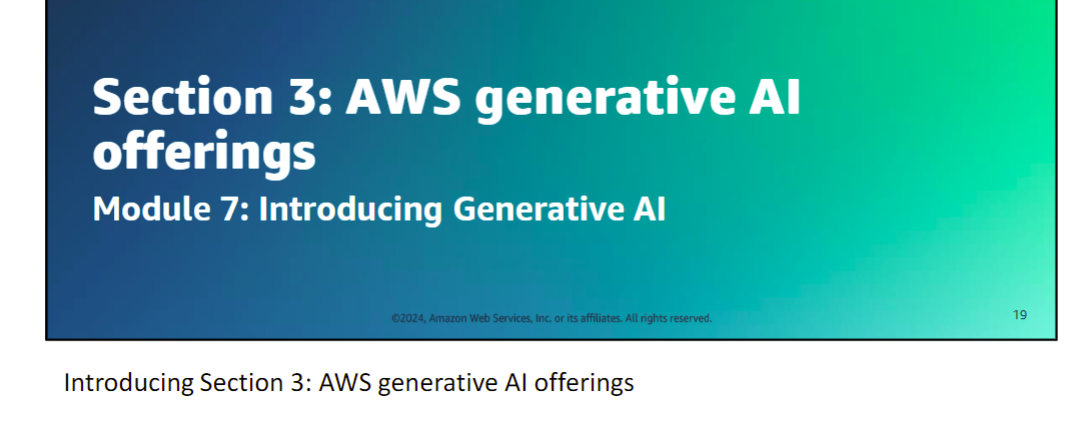
demonstration also highlights how to use generative AI to create images from text

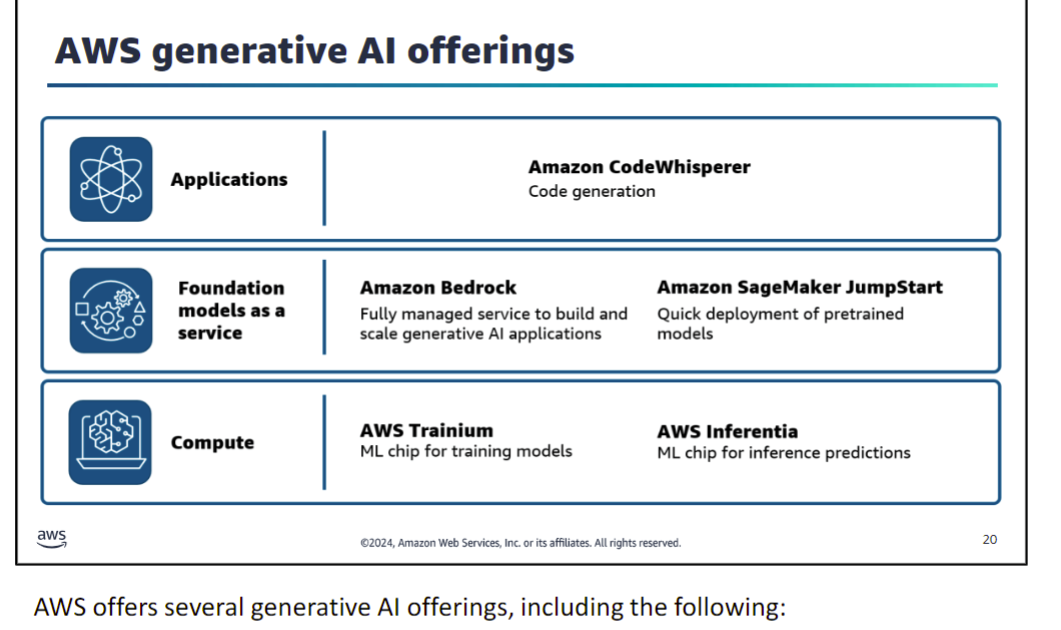
prompts.



Voici quelques points clés à retenir de cette section du module :

• Les **modèles de base** (Foundation Models, FMs) sont pré-entraînés sur une grande quantité de données et peuvent accomplir de nombreuses tâches.  
• Parmi les exemples de tâches qu'un FM peut réaliser, on trouve le traitement du langage, la compréhension visuelle et la génération de code.  
• Un sous-ensemble des FMs, les **grands modèles de langage** (Large Language Models, LLMs), sont entraînés sur des trillions de mots à travers de nombreuses tâches de langage naturel.  
• L'**ingénierie de prompt** est le processus de conception et de raffinement des prompts ou des stimuli d'entrée pour qu'un modèle de langage génère des types spécifiques de sortie.





Amazon CodeWhisperer est un assistant de codage qui s'intègre à votre environnement de développement intégré (IDE). Vous en apprendrez davantage sur CodeWhisperer dans la prochaine section.

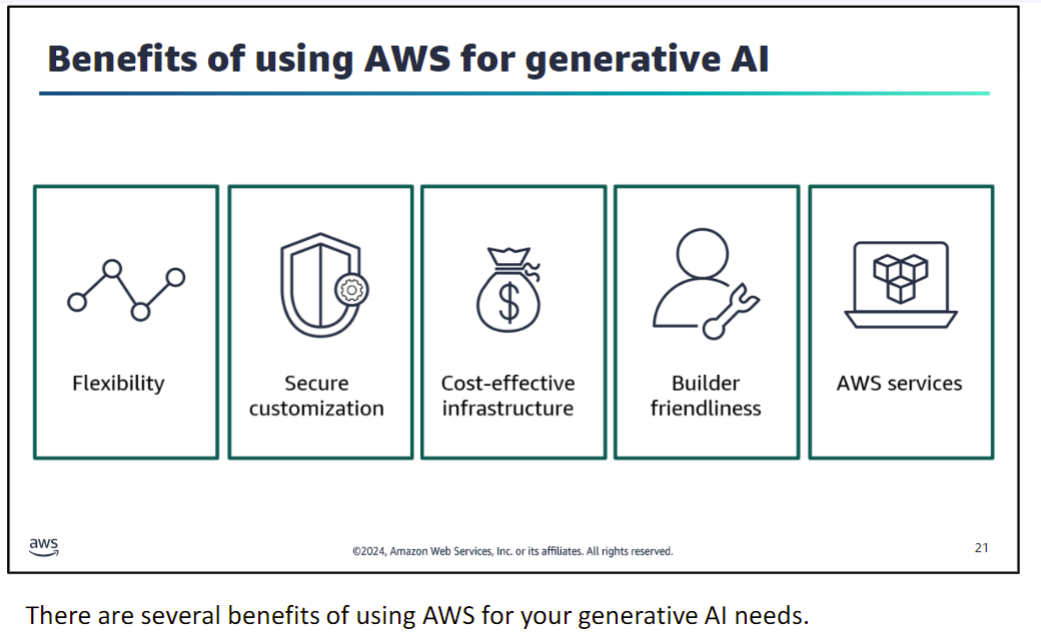
• **Amazon Bedrock** est un service entièrement géré qui met à disposition des modèles de base (FMs) d'AWS et de startups d'IA de premier plan via une API. Cela signifie que vous pouvez choisir parmi différents FMs pour trouver celui qui est le mieux adapté à votre cas d'utilisation. Amazon Bedrock aide les développeurs à créer des applications d'IA générative capables de fournir des réponses à jour basées sur des sources de connaissances propriétaires.

• **Amazon SageMaker JumpStart** vous aide à démarrer rapidement avec le machine learning. Il propose un ensemble de solutions pour les cas d'utilisation les plus courants, qui peuvent être déployées en quelques étapes seulement. Les solutions sont entièrement personnalisables et illustrent l'utilisation des modèles AWS CloudFormation et des architectures de référence, afin d'accélérer votre parcours ML.

• **AWS Trainium** est l'accélérateur de machine learning de deuxième génération qu'AWS a conçu spécifiquement pour l'entraînement de modèles de deep learning de plus de 100 milliards de paramètres.

• **AWS Inferentia** est une puce ML personnalisée conçue par AWS, que vous pouvez utiliser pour des prédictions d'inférence haute performance. Pour utiliser ces puces, configurez une instance Amazon Elastic Compute Cloud (Amazon EC2) et utilisez le SDK AWS Neuron pour invoquer la puce Inferentia ou Trainium.

Pour plus d'informations sur les autres offres d'IA générative d'AWS, consultez la diapositive "Ressources supplémentaires" de ce module.



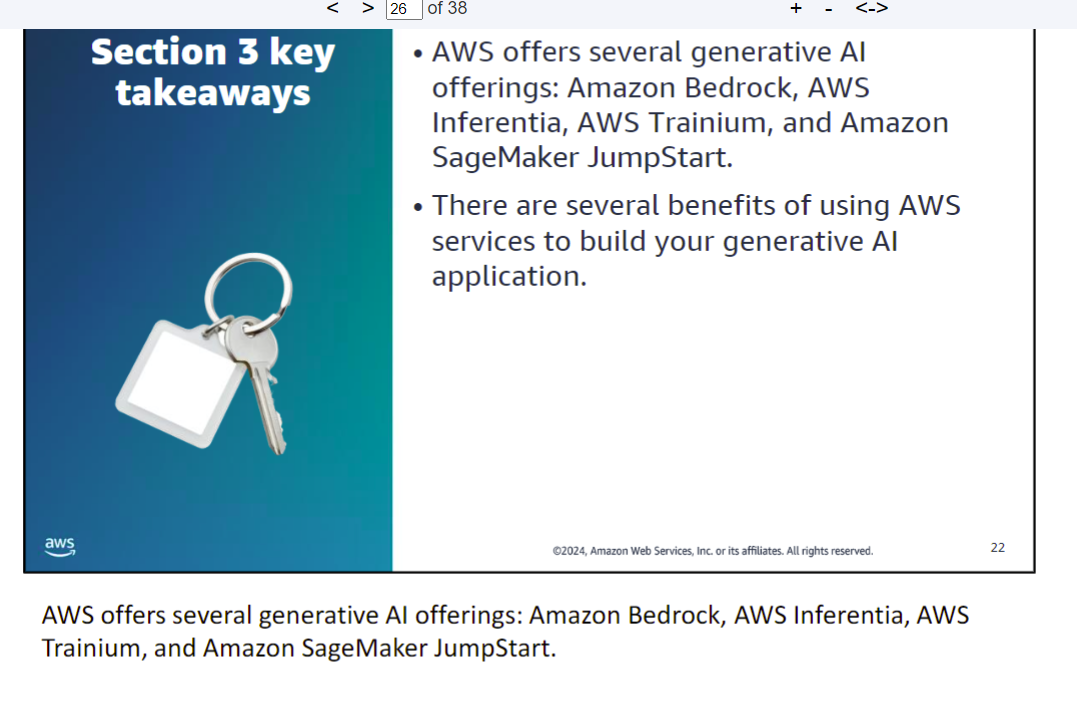
AWS offre de la flexibilité. Les clients ont besoin d'un moyen simple pour trouver et accéder à des modèles de base (FMs) performants, adaptés à leurs besoins professionnels. AWS propose une large sélection de FMs développés par Amazon et des startups d'IA de premier plan, comme AI21 Labs, Anthropic, et Stability AI.

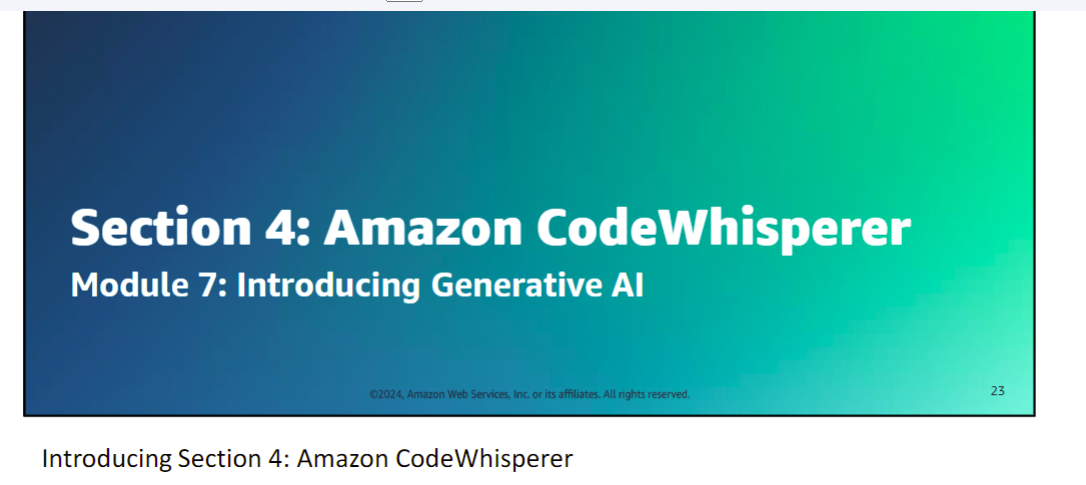
Un autre avantage est la personnalisation sécurisée. Les clients souhaitent créer des applications différenciées en utilisant leurs propres données, tout en assurant que ces données restent totalement protégées, sécurisées et privées. Grâce à Amazon Bedrock, les clients peuvent ajuster les modèles pour une tâche spécifique sans avoir à annoter de grands volumes de données (aussi peu que 20 exemples suffisent). Aucune donnée client n'est utilisée pour entraîner les modèles sous-jacents. Toutes les données sont cryptées et restent au sein du cloud privé virtuel (VPC) du client.

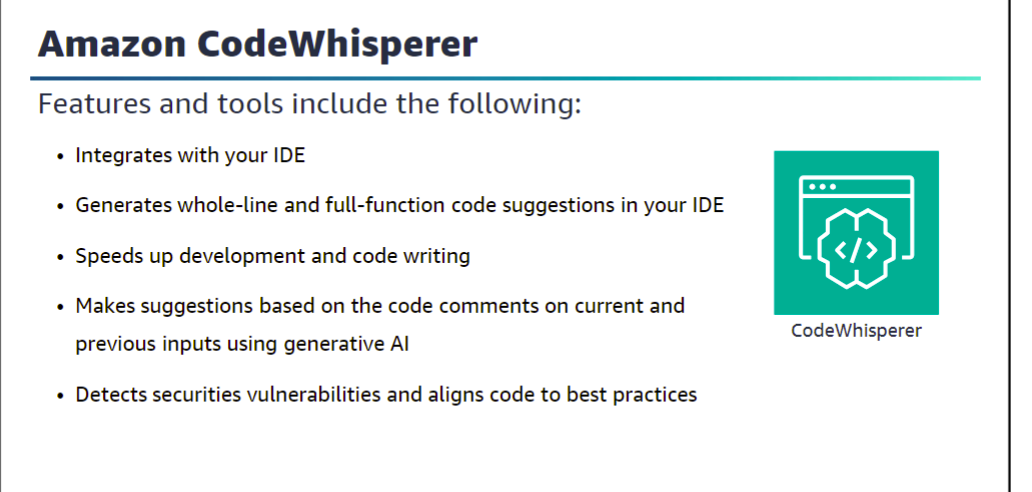
Les clients veulent pleinement utiliser les FMs dans une infrastructure performante et rentable, spécialement conçue pour le machine learning. AWS offre le meilleur rapport prix/performance pour l'IA générative grâce à une infrastructure alimentée par des puces ML conçues par AWS et des GPU NVIDIA. Avec AWS, les clients peuvent faire évoluer leur infrastructure de manière économique pour entraîner et exécuter des FMs contenant des centaines de milliards de paramètres.

Les solutions AWS sont adaptées aux développeurs. Les clients souhaitent intégrer et déployer rapidement des FMs dans leurs applications et charges de travail fonctionnant sur AWS. Avec les capacités d'IA générative d'AWS, les clients n'ont pas besoin d'envoyer leurs données au modèle. Ils peuvent plutôt amener le modèle à leurs données en utilisant des contrôles et des intégrations familières avec des services tels qu'Amazon SageMaker et Amazon Simple Storage Service (Amazon S3).

Les services AWS offrent aux clients des solutions d'IA générative pour améliorer la productivité tout en interagissant de manière transparente avec les applications et systèmes. Grâce à l'IA générative intégrée, des services tels qu'Amazon CodeWhisperer peuvent aider les clients à améliorer leur productivité. Les clients peuvent également déployer des solutions d'IA générative, comme la synthèse d'appels et les réponses à des questions, qui combinent les services d'IA d'AWS avec des FMs de premier plan.







CodeWhisperer

Amazon CodeWhisperer is an AI coding companion that generates whole-line and

full-function code suggestions in your IDE. With CodeWhisperer, ML engineers can

speed up development and code writing. CodeWhisperer will make suggestions

based on the code comments on current and previous inputs using generative AI.

CodeWhisperer is best used for undifferentiated, repetitive, common coding tasks.

Aside from helping to generate code, CodeWhisperer also scans your code to detect

hard-to-find security vulnerabilities and provides code suggestions to remediate

them. It aligns code to best practice for correcting security vulnerabilities outlined by

Open Worldwide Application Security Project (OWASP) and other best practices.

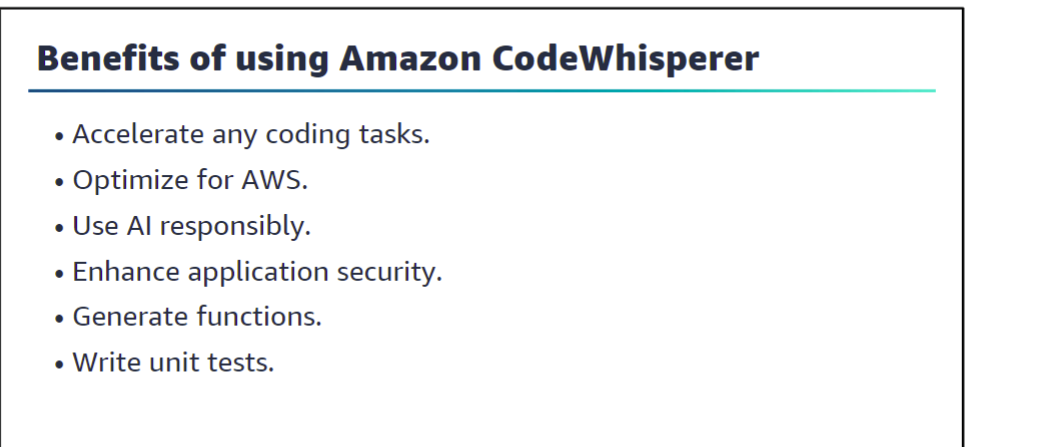
For more information about the languages and platforms supported by Amazon

CodeWhisperer, see the Additional resources slide of this module.

**CodeWhisperer**  
Amazon CodeWhisperer est un assistant de codage alimenté par l'IA qui génère des suggestions de code complètes, allant de lignes entières à des fonctions complètes, directement dans votre IDE. Avec CodeWhisperer, les ingénieurs en machine learning (ML) peuvent accélérer le développement et la rédaction de code. CodeWhisperer fait des suggestions basées sur les commentaires de code et les entrées actuelles et précédentes en utilisant l'IA générative. CodeWhisperer est particulièrement utile pour les tâches de codage courantes, répétitives et non différenciées.

En plus d'aider à générer du code, CodeWhisperer analyse également votre code pour détecter les vulnérabilités de sécurité difficiles à identifier et propose des suggestions de code pour les corriger. Il aligne le code sur les meilleures pratiques pour corriger les vulnérabilités de sécurité, telles que celles décrites par l'Open Worldwide Application Security Project (OWASP) et d'autres meilleures pratiques.

Pour plus d'informations sur les langages et plateformes pris en charge par Amazon CodeWhisperer, consultez la diapositive "Ressources supplémentaires" de ce module



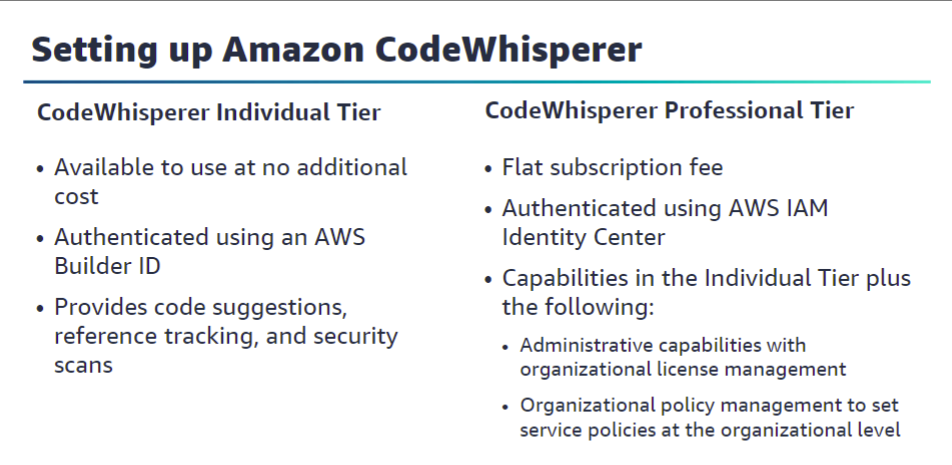
Amazon CodeWhisperer a été entraîné sur des milliards de lignes de code et est principalement utilisé pour des tâches de codage répétitives, courantes et non différenciées. Il aide les développeurs à accélérer n'importe quelle tâche de codage en générant des suggestions de code en temps réel. Il peut suggérer tout, allant de morceaux de code à des fonctions complètes. Les développeurs peuvent obtenir des suggestions de code optimisées pour les API AWS, y compris Amazon EC2, AWS Lambda et Amazon S3.

CodeWhisperer aide les développeurs à utiliser l'IA de manière responsable pour éviter les biais en filtrant les suggestions de code qui pourraient être considérées comme biaisées et injustes. De plus, CodeWhisperer dispose également d'un scanner de sécurité qui aide à atténuer les vulnérabilités de sécurité, protégeant ainsi l'intégrité de la base de code. Lorsqu'il recherche des vulnérabilités de sécurité, CodeWhisperer évalue votre code par rapport à plusieurs ensembles de normes et de meilleures pratiques. Cela inclut les éléments suivants :

• Normes du projet de sécurité des applications ouvertes dans le monde entier (OWASP)  
• Meilleures pratiques des bibliothèques cryptographiques  
• Normes de sécurité AWS

CodeWhisperer génère des fonctions entières et des blocs logiques de code (comportant souvent jusqu'à 10 à 15 lignes de code) directement dans l'éditeur de code de l'IDE. CodeWhisperer peut également écrire des tests unitaires, mettre en œuvre la logique pour les tests unitaires et examiner les suggestions pour des tests unitaires supplémentaires.

Pour plus d'informations sur Amazon CodeWhisperer, consultez la diapositive "Ressources supplémentaires" de ce module.

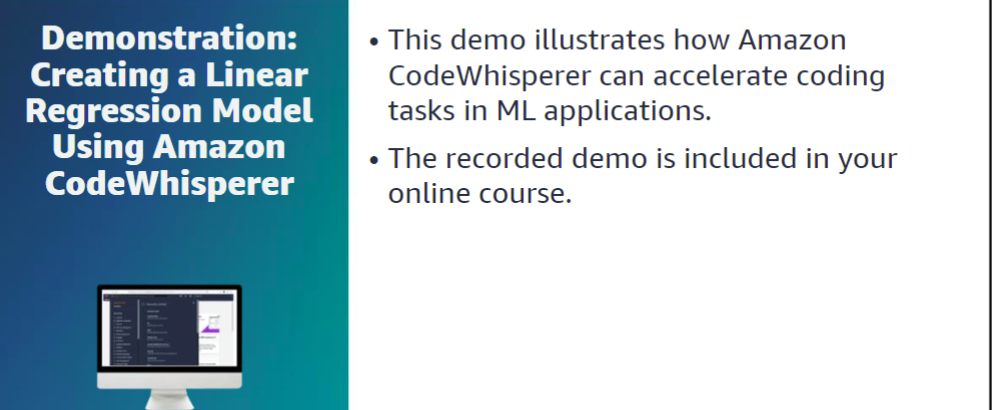


Il existe deux niveaux de service pour Amazon CodeWhisperer.

**Niveau individuel de CodeWhisperer**  
Le niveau individuel de CodeWhisperer est disponible sans coût supplémentaire. Les développeurs individuels peuvent s'inscrire et se connecter en utilisant un AWS Builder ID. Vous avez besoin d'une adresse e-mail pour obtenir un AWS Builder ID. Le niveau individuel fournit des suggestions de code, un suivi des références et des analyses de sécurité.

**Niveau professionnel de CodeWhisperer**  
Le niveau professionnel de CodeWhisperer est disponible pour un tarif d'abonnement fixe pour les professionnels ayant un compte AWS. En plus des capacités offertes dans le niveau individuel, le niveau professionnel de CodeWhisperer propose des fonctionnalités administratives pour les organisations qui souhaitent donner à leurs développeurs un accès à CodeWhisperer. Les administrateurs bénéficient d'une gestion centralisée des licences organisationnelles pour gérer qui, au sein de l'organisation, doit avoir accès à CodeWhisperer. Ils obtiennent également une gestion des politiques organisationnelles pour définir des politiques de service au niveau organisationnel, telles que la possibilité pour les développeurs de recevoir des suggestions de code pouvant être similaires à des données d'entraînement open-source spécifiques.

Pour plus d'informations sur la configuration d'Amazon CodeWhisperer, consultez la diapositive "Ressources supplémentaires" de ce module.

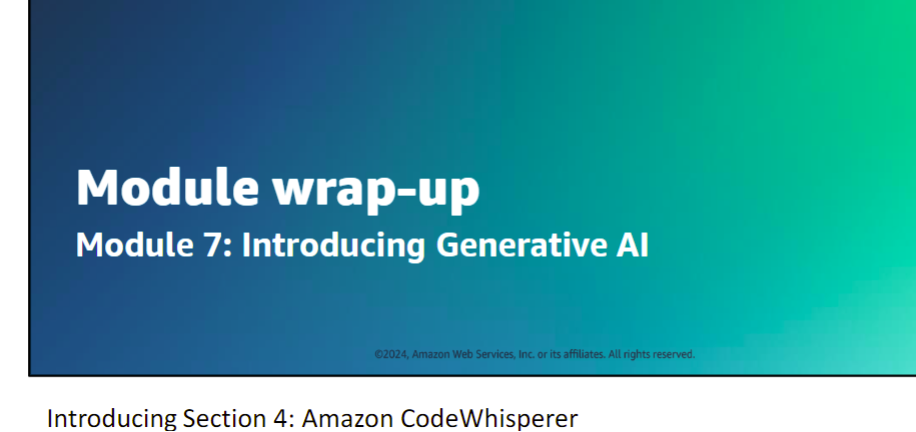


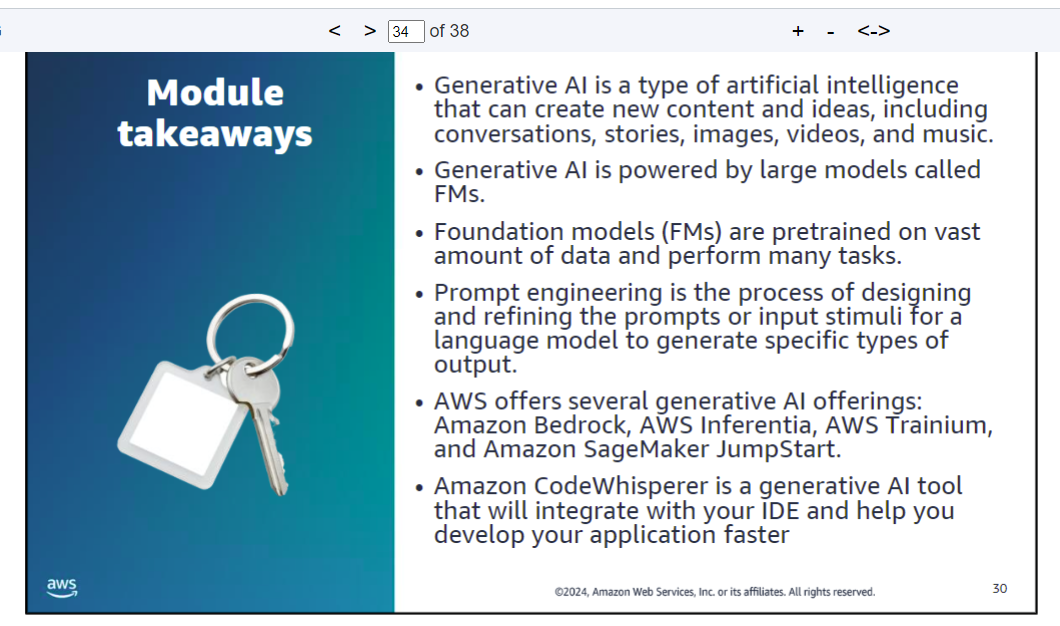
This demonstration highlights how CodeWhisperer increases developer productivity

by suggesting code comments and suggestions in Jupyter notebooks.



**Amazon CodeWhisperer est un outil d'IA générative qui vous aidera à développer votre application plus rapidement**. CodeWhisperer dispose également d'un scanner de sécurité qui aide à atténuer les vulnérabilités de sécurité, protégeant ainsi l'intégrité de la base de code. Il existe deux niveaux de service pour Amazon CodeWhisperer : le niveau individuel et le niveau professionnel.

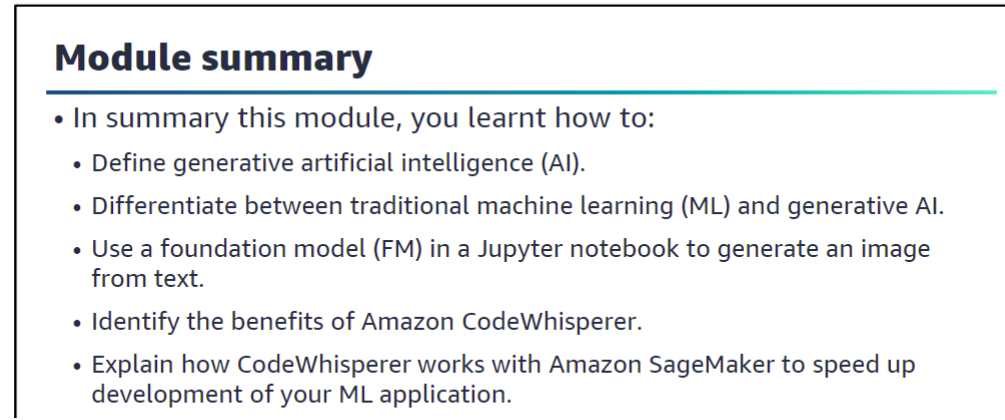




Voici quelques points clés à retenir de cette section du module :

• L'IA générative est un type d'intelligence artificielle capable de créer de nouveaux contenus et idées, y compris des conversations, des histoires, des images, des vidéos et de la musique.  
• L'IA générative est alimentée par de grands modèles appelés modèles de base (FMs).  
• Les modèles de base (FMs) sont pré-entraînés sur de vastes quantités de données et accomplissent de nombreuses tâches.  
• L'ingénierie de prompt est le processus de conception et de raffinement des prompts ou des stimuli d'entrée pour qu'un modèle de langage génère des types spécifiques de sortie.  
• AWS propose plusieurs offres d'IA générative : Amazon Bedrock, AWS Inferentia, AWS Trainium et Amazon SageMaker JumpStart.  
• Amazon CodeWhisperer est un outil d'IA générative qui s'intégrera à votre IDE et vous aidera à développer votre application plus rapidement.

4o mini



**En résumé, dans cette section, vous avez appris à :**

• Définir l'intelligence artificielle (IA) générative.  
• Différencier l'apprentissage automatique (ML) traditionnel de l'IA générative.  
• Utiliser un modèle de base (FM) dans un carnet Jupyter pour générer une image à partir de texte.  
• Identifier les avantages d'Amazon CodeWhisperer.  
• Expliquer comment CodeWhisperer fonctionne avec Amazon SageMaker pour accélérer le développement de votre application de ML.

