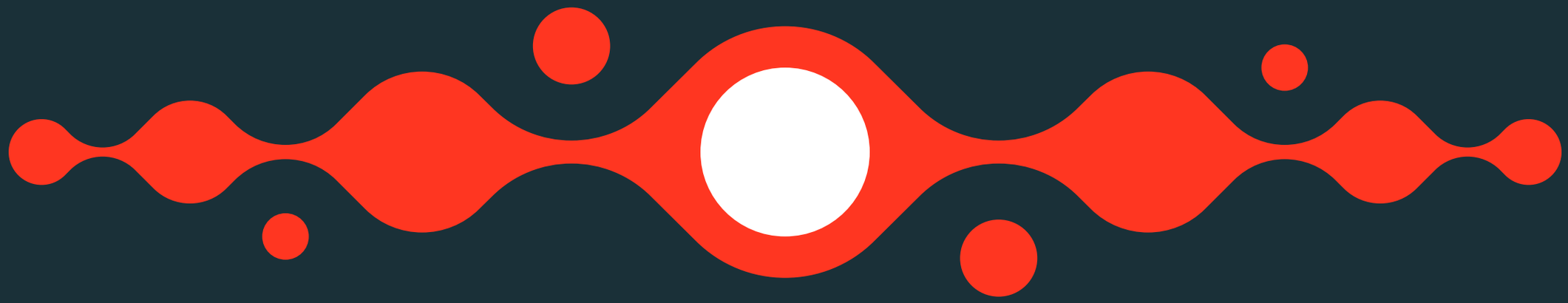
Guide pratique de

GenAI



databricks

TABLE DES MATIÈRES

databricks

Introduction

Des outils innovants et de nouvelles compétences pour une IA générative de qualité production

En transformant les pratiques, la création et les interactions avec les clients, l'IA générative a ouvert tout un monde de possibilités aux entreprises. Selon un récent rapport, « Libérer l'IA en entreprise », 85 % des entreprises mondiales utilisent déjà l'IA générative, et ce chiffre devrait atteindre 99 % d'ici 2027. Pourtant, elles rencontrent fréquemment des difficultés dans la mise en œuvre de leurs projets. Ce rapport précise en effet que 22 % seulement des entreprises sont convaincues que leur infrastructure est prête pour l'IA. Et elles ne sont que 37 % à considérer leurs modèles d'IA générative véritablement prêts pour la production. Pour beaucoup d'entre elles, les difficultés surviennent lors du passage en production de ces applications. Pour atteindre le niveau de qualité attendu de la part d'applications destinées aux clients, les résultats de l'IA doivent être précis, contrôlés et sûrs.

Mosaic AI : unifier le cycle de vie de l'IA, des données au déploiement

Mosaic AI propose une suite complète d'outils pour relever les défis inhérents au déploiement d'applications d'IA générative de qualité production. Mosaic AI permet une gestion transparente à chaque étape du cycle de vie de l'IA : collecte et préparation des données, développement, déploiement, surveillance et gouvernance des modèles. En juin 2024, les clients de Mosaic AI avaient créé plus de 200 000 modèles d'IA personnalisés au cours de l'année précédente. Ils ont exploité l'infrastructure et la technologie qui alimentent les modèles de pointe de Databricks.

L'infrastructure de donnée doit évoluer pour prendre en charge les applications reposant sur l'IA générative

Pour entrer dans l'ère de l'IA générative, il ne suffit pas de déployer un chatbot : il faut engager une transformation des pratiques fondamentales de gestion des données. Le data lakehouse, fer de lance de la « pile de données moderne », est au cœur de cette transformation. Il fournit aux entreprises des solutions plus rapides et économiques pour la gestion des données. Ces architectures sophistiquées permettent aux organisations d'exploiter efficacement tout le potentiel de l'IA générative en intégrant et en gouvernant les données à grande échelle dans un système unifié.

Les entreprises s'appuient de plus en plus sur des outils et des applications intégrant l'IA générative pour renforcer leur compétitivité. Pour prendre en charge ces technologies avancées de façon sûre et évolutive, elles doivent faire évoluer leur infrastructure de données.

Où que vous en soyez dans le déploiement de l'IA générative, la qualité de vos données est déterminante.

Pour obtenir des résultats de qualité production avec l'IA générative, les entreprises ont besoin d'outils robustes pour comprendre la qualité de leurs données et des résultats de leurs modèles. Cette démarche consiste à combiner et optimiser tous les aspects du processus d'IA générative : préparation des données, entraînement des modèles (à l'aide de modèles de récupération, de modèles de langage et de pipelines de classement et de post-traitement), l'ingénierie de prompt et la personnalisation des modèles avec les données de l'entreprise.

Sur la plateforme unifiée de Mosaic AI, les data scientists, les ingénieurs et les utilisateurs métier travaillent ensemble. Ils sont assurés d'obtenir une traçabilité complète des données et des modèles, des données brutes jusqu'aux déploiements en production. Quant à l'intégration de la plateforme avec Unity Catalog, Lakehouse Monitoring et le service de modèles, elle garantit aux entreprises la gouvernance et l'optimisation de chaque étape du workflow d'IA générative.

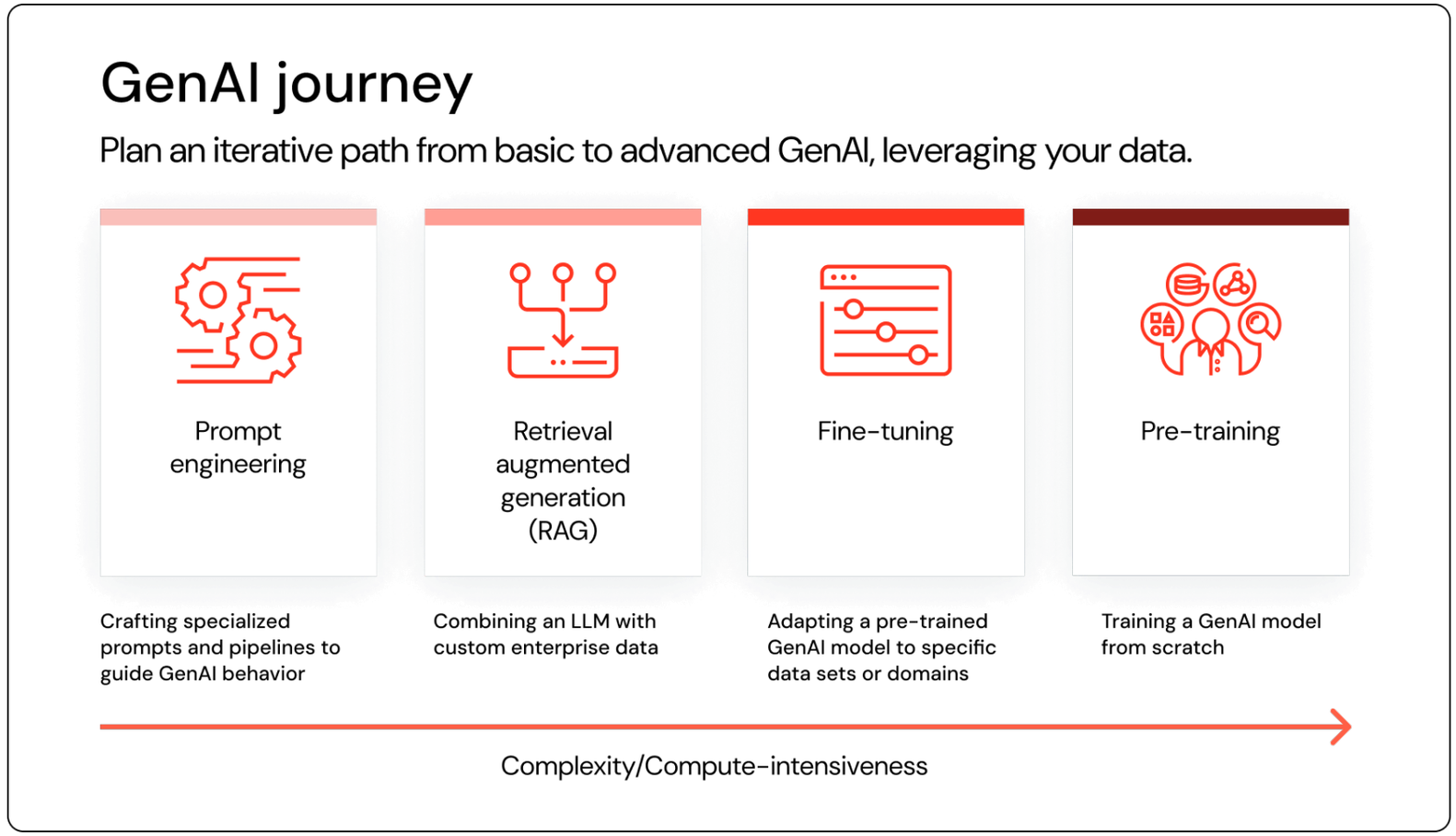
Dans ce guide, vous apprendrez :

- Les secrets des prompts efficaces et des techniques avancées pour exploiter les grands modèles de langage (LLM) dans des applications d'IA générative
- Les fondamentaux de la création de systèmes d'agents IA orientés tâches et de l'intégration de la génération augmentée par récupération (RAG) pour des résultats plus fins
- Comment évaluer et améliorer les performances des applications d'IA générative et des processus d'ajustement et de préentraînement des modèles
- Comment déployer des LLM personnalisés en production et les superviser à l'aide d'une plateforme unifiée
- Les bonnes pratiques pour étendre et optimiser les performances et la qualité de vos applications d'IA générative tout en maîtrisant les coûts

Les cas d'usage de l'IA générative abordés :

- Utiliser les LLM pour obtenir des informations exploitables à partir d'avis sur des produits
- Améliorer la qualité des résultats des chatbots avec la RAG
- Entraîner des modèles d'IA générative personnalisés sans dépasser les budgets
- Surveiller les modèles d'IA générative déployés et suivre leurs performances

Chapitre 1 : Les fondamentaux de l'IA générative



Qu'est-ce que l'IA générative ?

L'IA générative est un domaine de l'intelligence artificielle qui vise à créer des modèles capables de générer du contenu inédit, en particulier du texte, des images, du code et des données synthétiques. Contrairement aux modèles d'IA traditionnels, conçus pour classer, analyser ou prédire en fonction d'associations prédéfinies entre entrées et sorties, l'IA générative a pour but de produire des sorties qui ressemblent à du contenu existant ou qui l'enrichissent. Les modèles d'IA générative ont donc la particularité de pouvoir produire des combinaisons entièrement nouvelles et uniques, en suivant des schémas appris à partir de vastes datasets.

Ces modèles se répartissent généralement en deux catégories : grands modèles de langage (LLM) et modèles de fondation. Les LLM forment une classe de modèles génératifs spécialisés dans les tâches linguistiques. Les modèles de fondation, quant à eux, sont de grands modèles préentraînés et destinés à être affichés ou adaptés en vue d'une application spécifique. Une fois entraînés, ces modèles produisent des sorties qui imitent la structure, le ton et l'intention du langage naturel. Ils peuvent ainsi générer du texte, réaliser des traductions et des résumés, et répondre à des questions.

L'IA générative s'appuie également sur des techniques avancées, et notamment sur l'**ingénierie de prompt** et la **génération augmentée par récupération (RAG)**. L'ingénierie de prompt est l'art d'élaborer des prompts capables de susciter un certain comportement chez un modèle. La RAG, en revanche, associe au LLM la consultation de connaissances externes, ce qui permet au modèle d'enrichir le contenu généré avec des informations pertinentes et à jour.

La flexibilité et la puissance de l'IA générative sont sensibles dans un large éventail d'applications :

- **Génération de documents :** rédaction automatique de rapports, de contrats et de propositions commerciales selon des besoins spécifiques, dans le but d'améliorer la productivité en réduisant les efforts manuels
- **Génération d'images :** production de nouvelles images à partir d'une entrée ou d'une transformation de style
- **Tâches liées à la parole :** transcription, traduction et interprétation de l'intention grâce au traitement du langage naturel
- **Agents :** systèmes autonomes capables d'accomplir des tâches complexes en interagissant avec des utilisateurs ou d'autres systèmes, souvent pilotés par des modèles génératifs
- **Ajustement :** utilisation de datasets ciblés pour personnaliser des modèles préentraînés en vue de tâches plus spécifiques et spécialisées

Si de nombreux modèles d'IA générative intègrent des protections, ils peuvent tout de même produire des résultats inexacts ou nuisibles. Il faut impérativement accompagner leur utilisation d'une grande prudence.

À propos des grands modèles de langage

Les grands modèles de langage sont des modèles d'apprentissage profond qui excellent dans les tâches linguistiques, en particulier dans l'interprétation et la génération du langage naturel. Ces modèles ont été développés à partir de grandes quantités de données, généralement du texte issu d'un large éventail de sources : livres, sites web et réseaux sociaux. Ces modèles sont entraînés à partir d'énormes datasets afin qu'ils identifient des schémas et des associations. Ils apprennent ainsi à prédire le mot ou l'expression qui suit dans une phrase.

Une fois qu'un modèle a été entraîné, il peut générer des phrases cohérentes et contextualisées en exploitant les schémas qu'il a appris. Ces modèles exploitent des probabilités statistiques dérivées des données d'entraînement, ce qui leur permet d'imiter un texte de type humain avec une grande efficacité. Les LLM peuvent effectuer diverses tâches linguistiques, notamment générer, traduire et résumer du texte.

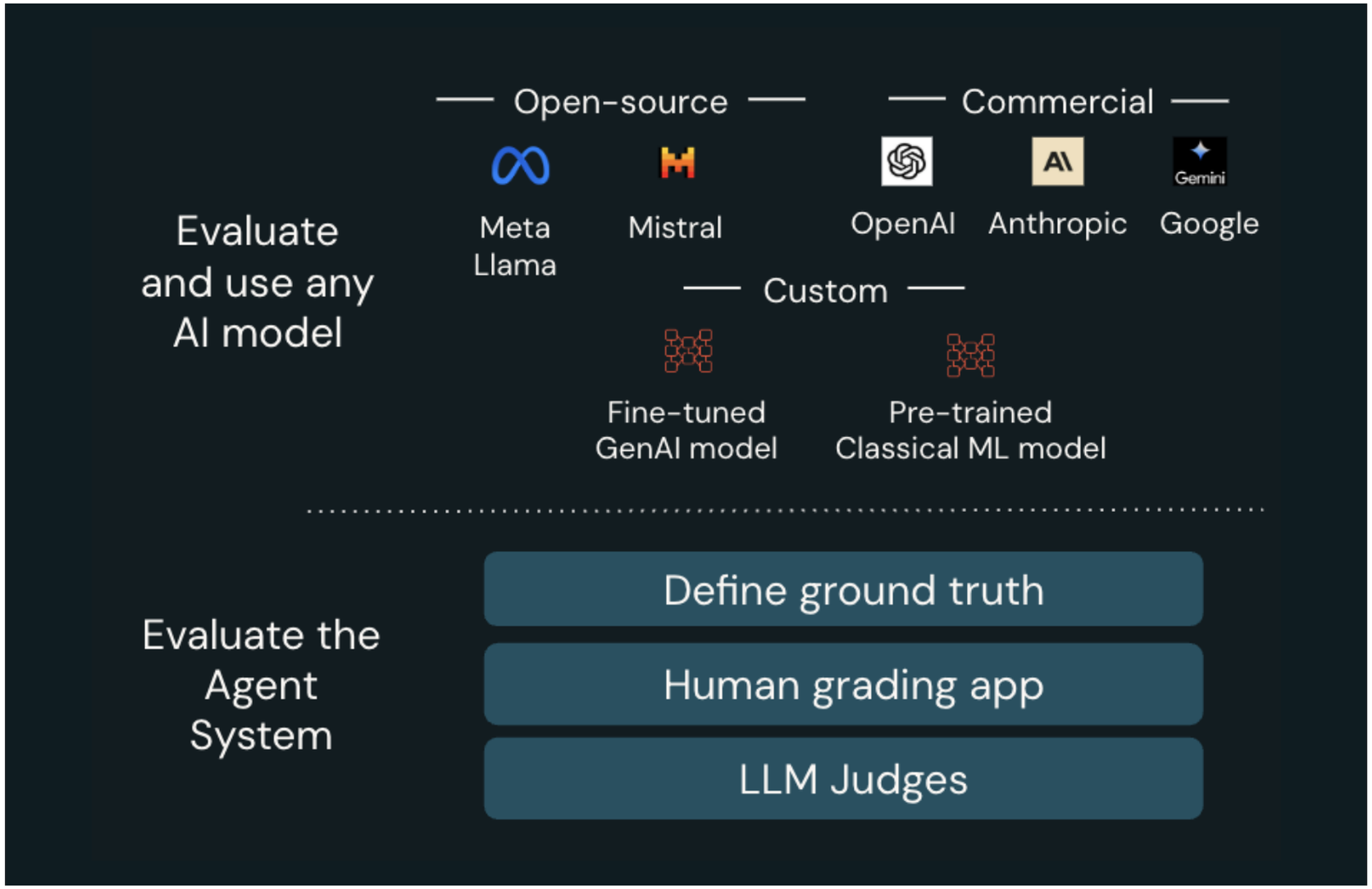
Entraînement et adaptation :

1. **Préentraînement :** les LLM sont d'abord préentraînés sur des quantités massives de données textuelles pour acquérir une connaissance générale de la langue. Cette étape leur permet d'apprendre les structures linguistiques de base et les associations de mots typiques. Le préentraînement se fait généralement à l'aide de techniques d'apprentissage non supervisées : le modèle prédit les mots manquants dans des phrases ou apprend à modéliser des formulations à partir de zéro.
2. **Ajustement :** une fois le préentraînement terminé, les LLM sont généralement affinés à l'aide de datasets propres à un domaine. Cette étape permet d'adapter un modèle généraliste aux besoins de secteurs ou d'applications spécifiques, comme la santé, la finance ou le support client. L'ajustement consiste à modifier les paramètres du modèle pour optimiser ses performances dans des tâches ou des datasets ciblés, souvent en utilisant des ensembles de données plus petits et constitués avec soin.

Les LLM peuvent être combinés à la **génération augmentée par récupération**. Celle-ci améliore leurs capacités en puisant des informations pertinentes dans des bases de données ou de connaissances externes. Grâce à cette approche qui associe la puissance de la génération à des informations factuelles, le modèle se tient informé sur des sujets spécifiques et produit des résultats plus précis.

On peut encore ajouter une couche de sophistication aux LLM en déployant des **agents**. Ces systèmes sont capables d'interagir avec les utilisateurs et de prendre des décisions en toute autonomie. Les agents alimentés par des LLM savent exécuter des tâches complexes : exécution de requêtes, gestion de workflow, échanges conversationnels, etc. Ils peuvent encore être affinés grâce à des réglages précis, ce qui maximise leur efficacité dans des tâches clés.

En utilisant ces modèles comme éléments de base, il devient possible d'adapter l'IA générative à diverses utilisations. Celles-ci vont des chatbots et assistants virtuels aux applications plus spécialisées nécessitant de synthétiser des connaissances de différents domaines.



Modèles open source et modèles propriétaires : comparer et sélectionner des LLM

Il est important de choisir un grand modèle de langage en fonction de l'application, car il exerce un impact direct sur les performances, le coût et l'évolutivité du projet. Il faut bien souvent choisir entre des modèles open source et propriétaires. Chacun d'entre eux présente des avantages et des inconvénients propres. Les critères essentiels sont les performances, l'adéquation avec le cas d'usage, le coût et la sécurité.

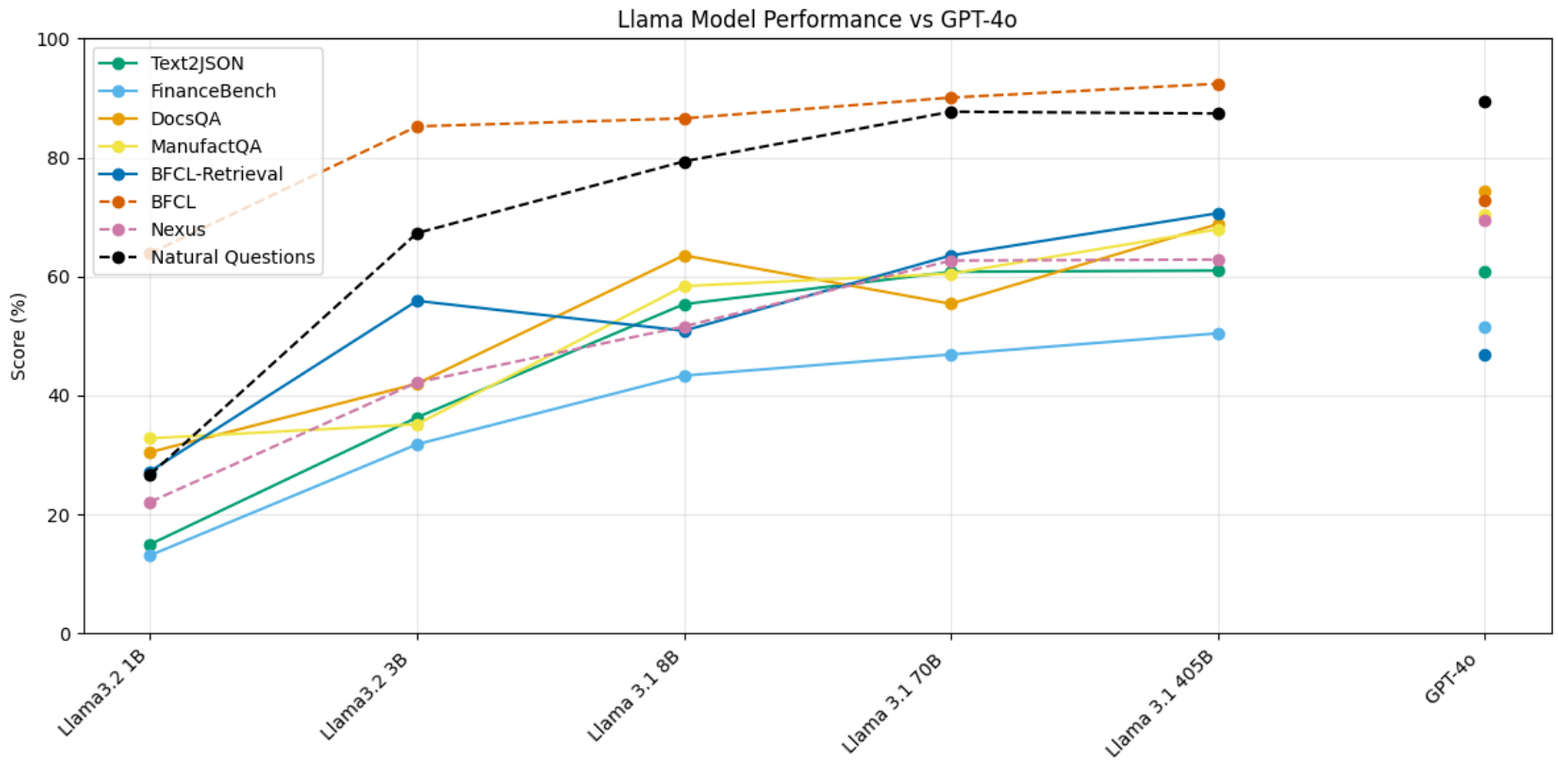


Figure 1 : Les performances de Llama 3.1 405B sont comparables à celles de GPT-4o dans de nombreuses tâches. En revanche, nous observons qu'elles tendent à baisser lorsque la taille du modèle diminue, bien que Llama 3.2 3B reste au-dessus de la moyenne sur une partie des critères académiques évalués.

ANALYSE COMPARATIVE DES PERFORMANCES

L'analyse comparative permet d'évaluer les modèles en fonction de paramètres tels que la précision, la latence, le débit et la rentabilité. Ces indicateurs doivent correspondre aux exigences de votre application :

- **Les applications en temps réel** ont avant tout besoin d'une faible latence
- **Les environnements caractérisés par un grand nombre de transactions** exigent évolutivité et débit
- **Les industries réglementées** ont besoin de modèles respectant des normes strictes de sécurité et de conformité

Les modèles propriétaires intègrent souvent des fonctionnalités de sécurité et de conformité. Les organisations qui optent pour des modèles open source doivent souvent gérer elles-mêmes ces aspects, ce qui nécessite une expertise et des ressources adaptées.

MODÈLES OPEN SOURCE

Les modèles open source, à l'instar de Llama de Meta, sont à la fois économiques et hautement personnalisables. Du fait de leur transparence et de leur flexibilité, ils conviennent parfaitement aux organisations ayant besoin de contrôler leur architecture et leur comportement, en particulier dans les secteurs réglementés. Ils bénéficient du soutien de communautés dynamiques qui sont des moteurs d'innovation, et reçoivent des mises à jour fréquentes.

Ils pèchent toutefois par manque d'optimisation spécialisée, et leur déploiement en production peut nécessiter une maintenance importante. Les organisations qui déploient des modèles open source doivent donc s'attendre à relever des défis supplémentaires en termes de sécurité et d'évolutivité.

MODÈLES PROPRIÉTAIRES

Les modèles propriétaires, qui proviennent de fournisseurs comme OpenAI ou Google, sont souvent optimisés pour des tâches spécifiques, offrant dès le départ des fonctionnalités avancées. Ils s'accompagnent d'un support de niveau professionnel et d'une documentation solide. Ils sont en outre conformes aux normes telles que HIPAA ou SOC 2. Ils répondent ainsi aux besoins des organisations soumises à des exigences réglementaires strictes.

Leur principal inconvénient réside dans leur coût : les frais de licence et d'utilisation peuvent être élevés. Il faut aussi savoir que les organisations n'ont qu'un contrôle limité sur ces modèles, ce qui peut limiter les possibilités de personnalisation et soulever des problèmes de confidentialité des données.

Créer et adapter des LLM pour des cas d'usage spécifiques

Le développement et l'adaptation de grands modèles de langage impliquent généralement trois étapes clés : le **préentraînement**, l'**ajustement** et le **préentraînement continu (CPT)**. Chacune de ces étapes joue un rôle distinct au service d'un objectif : créer un modèle maîtrisant un large éventail de tâches, tout en restant capable de s'adapter à des domaines et des usages spécifiques. Ces étapes forment ensemble un cadre robuste pour la création de solutions d'IA performantes et spécialisées.

PRÉENTRAÎNEMENT

Le préentraînement est le processus fondamental au cours duquel un modèle est exposé à des datasets massifs contenant des milliards de jetons et provenant de sources diverses – livres, articles, sites web, etc. Au cours de cette phase, le modèle acquiert les schémas linguistiques généraux, la grammaire et les faits. Il peut ensuite générer un texte cohérent et accomplir différentes tâches linguistiques. Grâce à ces connaissances généralistes, les modèles préentraînés excellent dans de nombreuses applications sans avoir été explicitement formés pour chaque tâche.

Une remarque toutefois : si le préentraînement confère aux modèles une compréhension globale du langage, il ne leur donne aucune expertise spécialisée. Sans adaptation supplémentaire, les modèles préentraînés peuvent avoir du mal à réaliser certaines tâches de niche, comme l'interprétation de documents juridiques ou l'analyse de données cliniques. Rappelons aussi que le préentraînement mobilise d'importantes ressources, une grande puissance de calcul et des datasets d'un volume considérable. Pour cette raison, cette phase est généralement menée par des organisations qui disposent de ressources conséquentes.

AJUSTEMENT

L'ajustement affine le modèle préentraîné à l'aide de datasets plus réduits et spécialisés. Au cours de cette étape, les organisations adaptent les modèles à des applications spécifiques : automatisation du service client, synthèse de contenu ou diagnostic médical, par exemple. L'ajustement est beaucoup moins gourmand que le préentraînement et permet d'adapter rapidement le modèle à de nouveaux cas d'usage.

L'ajustement fait intervenir plusieurs aspects clés :

- **Qualité des données :** la qualité du dataset d'ajustement est essentielle. Le modèle a besoin de données de haute qualité et bien étiquetées pour acquérir les subtilités propres à sa tâche.
- **Approche itérative :** l'ajustement est souvent itératif : on mène des expériences initiales sur de petits datasets pour valider les performances du modèle. On emploie ensuite des datasets plus volumineux pour gagner davantage en précision.
- **Données synthétiques :** Lorsque les données réelles sont rares, il est possible de générer des données synthétiques pour enrichir le dataset d'entraînement.
  Cette approche permet d'améliorer la généralisation sans procéder à une grande collecte de données.

Grâce à l'ajustement, les organisations peuvent obtenir des modèles hautement spécialisés dans leur domaine, et bénéficier ainsi de résultats plus précis et pertinents.

PRÉENTRAÎNEMENT CONTINU (CPT)

Le préentraînement continu, également appelé préentraînement d'adaptation au domaine, est une étape intermédiaire entre le préentraînement et l'ajustement. Contrairement à ce dernier qui vise à adapter le modèle à une tâche, le CPT consiste à soumettre le modèle préentraîné à un entraînement supplémentaire à l'aide d'un grand dataset propre à un domaine ou à une langue en particulier. Le modèle acquiert ainsi des connaissances approfondies sur un sujet spécifique avant d'être ajusté en vue de tâches précises.

Le CPT est utilisé pour différentes raisons :

- **Connaissances spécialisées :** les organisations qui exercent dans des domaines spécialisés tels que la santé, la finance ou le droit peuvent utiliser le CPT pour amener le modèle à mieux comprendre leur domaine.
- **Spécialisation linguistique :** les performances des modèles généraux peuvent être médiocres dans les langues rares. Le CPT peut améliorer la compétence du modèle dans une langue spécifique en l'exposant à des corpus importants.
- **Langages de programmation :** le CPT peut également servir à améliorer les compétences d'un modèle dans des langages de programmation spécifiques et sous-représentés lors de la phase de préentraînement initiale.

Le CPT consomme généralement moins de ressources que le préentraînement initial, mais davantage que l'ajustement. La quantité de données nécessaires varie en fonction du domaine et de la taille du modèle. En règle générale, le CPT implique plusieurs milliards de jetons, pour un volume souvent équivalent à environ 1 % de la taille du dataset de préentraînement d'origine.

En combinant CPT et ajustement, les organisations peuvent obtenir des modèles hautement spécialisés offrant à la fois une compréhension linguistique approfondie et une expertise hautement spécialisée dans un domaine. Cette combinaison est particulièrement efficace pour créer des applications qui exigent un haut niveau de précision, de spécialisation et de fiabilité.

Rentabilité et stratégie à long terme

Le coût reste un facteur crucial dans le choix d'un modèle. Parce qu'ils sont exempts de frais de licence, les modèles open source sont souvent plus abordables au départ. En revanche, les coûts de maintenance, d'ajustement et de mise à l'échelle peuvent rapidement s'envoler. Les modèles propriétaires sont plus chers, mais ils peuvent offrir davantage de **valeur** du fait de leur optimisation préconfigurée, de leur sécurité et de leur évolutivité, en particulier dans les cas d'usage les plus demandés.

En fin de compte, le choix entre modèle open source et modèle propriétaire dépend de plusieurs facteurs : exigences de performance, de personnalisation et de sécurité, contraintes budgétaires et support à long terme. Les organisations ont donc tout intérêt à procéder à analyse comparative approfondie pour sélectionner le meilleur LLM en phase avec leurs objectifs, tout en trouvant le meilleur compromis entre coût, sécurité et performance.

Le plus souvent, le processus de debugging implique de développer et d'ajuster :

- **L'infrastructure :** Mosaic AI Model Training gère la plupart des problèmes d'infrastructure pour vous. Le service crée automatiquement des points de contrôle pour reprendre l'entraînement en cas de défaillance des GPU, du réseau ou d'un autre aspect de l'infrastructure. Il est toutefois préférable de garder la consommation sous surveillance, en particulier lors de l'utilisation de configurations non standard.
- **La progression de l'apprentissage :** il est nécessaire de surveiller les indicateurs de pertes et d'autres métriques propres aux données d'entraînement et d'évaluation pour détecter les problèmes de données et de configuration. Les pics de pertes et les divergences, notamment, sont à surveiller en priorité. Dans Mosaic AI Training, nous recommandons de journaliser les expérimentations MLflow par défaut, à des fins de supervision en direct comme de revue post-hoc.
- **Les configurations :** des configurations mal définies provoquent généralement ces problèmes dès les premiers temps de l'entraînement. Le rythme d'apprentissage est la configuration qui nécessite le plus souvent un ajustement.
- **Les données :** pour prendre l'exemple d'un problème d'entraînement courant, les ensembles de données mal mélangés provoquent souvent des pics de perte. Mosaic AI Training simplifie le mélange via la bibliothèque Mosaic Streaming, mais ce processus a un coût. Mosaic Streaming propose donc plusieurs paramètres de mélange correspondant à différents compromis qualité-coût. Si vous observez des pics de perte, il est possible que des paramètres de mélange plus intensifs dans Mosaic Streaming y remédient. Par exemple, si vos données proviennent de différents buckets (domaines, langues, etc.) et qu'elles ne sont pas correctement mélangées, les pics de perte seront plus fréquents.

Chapitre 2 : L'art de l'ingénierie de prompt

À l'heure où l'IA générative évolue et transforme tous les secteurs d'activité, l'ingénierie de prompt devient un domaine clé pour les organisations. Cette compétence est essentielle, car elle veille à ce que les modèles d'IA, grands modèles de langage en tête, comprennent l'intention humaine et fournissent des résultats précis, utiles et créatifs. Aux premiers stades du déploiement de l'IA ou comme pour le perfectionnement de systèmes d'IA avancés, l'ingénierie de prompt est indispensable pour exploiter tout le potentiel de l'IA générative.

Ce chapitre offre une description détaillée de l'ingénierie de prompt pour mieux comprendre son importance, et il explique comment l'utiliser efficacement pour optimiser les performances de vos applications d'IA.

Qu'est-ce que l'ingénierie de prompt ?

L'ingénierie de prompt est un domaine dédié à l'élaboration, au perfectionnement et à l'optimisation des entrées en langage naturel (prompts) qui guident les modèles d'IA dans l'exécution de leurs tâches : génération de texte, écriture de code et création d'images, par exemple. L'ingénierie de prompt a pour principale fonction de traduire l'intention humaine sous une forme intelligible par l'IA, pour lui permettre de répondre avec précision.

L'avènement de l'IA générative a fait de l'ingénierie de prompt une discipline cruciale, surtout dans le contexte des progrès rapides de ChatGPT d'OpenAI, Llama de Meta et Gemini de Google, des modèles particulièrement puissants. Ils sont en effet parfaitement capables de générer du texte similaire à une production humaine, mais le résultat dépend en grande partie de la qualité des prompts qu'ils reçoivent. Des prompts mal rédigés peuvent produire des résultats hors sujet ou erronés. À l'inverse, une instruction bien pensée peut améliorer la précision, la créativité et la pertinence de la sortie.

Imaginons qu'une entreprise de commerce en ligne mette en place un chatbot alimenté par un modèle d'IA générative. Ce robot doit répondre aux diverses questions des clients concernant le statut de leur commande ou les politiques de retour. Dans ce cas, les ingénieurs de prompt peuvent élaborer des instructions spécifiques pour aider le chatbot à traiter ces requêtes, à demander des informations nécessaires (le numéro de commande, par exemple), à donner des réponses utiles et à transmettre les problèmes complexes à des agents humains au besoin. Plus les prompts sont précis, plus le chatbot devient efficace. Les clients sont satisfaits et la pression sur les opérations diminue.

L'ingénierie de prompt exploite plusieurs techniques :

- **Prompt sans exemple (zero-shot) :** demande au LLM d'effectuer une tâche sans lui donner d'exemples, en s'appuyant sur ses connaissances existantes
- **Prompt avec quelques exemples (few-shot) :** inclut un ou plusieurs exemples pour orienter la réponse du modèle
- **Prompt avec chaîne de pensée :** décompose les problèmes complexes en étapes logiques, pour aider le modèle à raisonner pour accomplir ses tâches
- **Prompt avec affinement automatique :** incite le modèle à résoudre un problème, à critiquer sa solution et à l'affiner en fonction d'un feedback
- **Prompt avec stimulation directionnelle :** utilise des indices ou des mots-clés pour orienter la réponse du modèle dans une direction spécifique
- **Prompt itératif :** s'appuie sur les réponses précédentes et des questions de suivi pour approfondir l'exploration du sujet
- **Injection de contexte :** fournit du contexte supplémentaire pour améliorer la pertinence de la réponse
- **Prompt avec jeu de rôle :** attribue une personnalité particulière au LLM pour l'encourager à donner des résultats spécialisés

L'ingénierie de prompt, un outil clé pour améliorer la qualité des résultats des modèles

Une ingénierie de prompt efficace ne consiste pas seulement à créer des entrées pour l'IA : elle vise à les optimiser pour maximiser la qualité des résultats. Elle recourt pour cela à plusieurs stratégies afin d'orienter le comportement de l'IA, de stimuler sa créativité et de réduire le risque d'erreur.

1. CONTEXTE ET SPÉCIFICITÉ

Des prompts bien conçus apportent à l'IA le contexte nécessaire pour mieux comprendre la tâche. Qu'il s'agisse de répondre à une question, de générer du contenu créatif ou de résoudre un problème, plus vous apporterez du contexte, plus le résultat sera pertinent. Par exemple, au lieu de demander : « Parle-moi du marché », dites plutôt : « Résume les principales tendances du marché technologique mondial en 2025. » Cette spécificité aide l'IA à affiner sa réponse et à fournir des informations plus ciblées et utiles.

2. LE RAISONNEMENT GUIDÉ POUR LES TÂCHES COMPLEXES

Lorsque l'IA doit réaliser une tâche complexe, des techniques avancées d'ingénierie de prompt, comme le prompt avec chaîne de pensée, vont décomposer le problème en étapes plus petites et digestes. Cette technique aide l'IA à réfléchir au problème de manière logique pour qu'elle produise des réponses plus claires et plus précises. Par exemple, si l'on demande à une IA de résoudre un problème mathématique, un prompt avec chaîne de pensée va l'inciter à décomposer le problème en plusieurs étapes pour arriver à une solution.

3. RÉDUIRE L'AMBIGUÏTÉ ET LES BIAIS

La clarté est cruciale dans l'ingénierie de prompt. Des prompts ambigus peuvent dérouter l'IA et l'amener à donner des réponses hors sujet. Il faut aussi savoir que des prompts mal conçus peuvent involontairement perpétuer des préjugés liés au genre, à l'appartenance ethnique et à d'autres questions sensibles. En détaillant explicitement le résultat souhaité et en formulant vos prompts avec un langage diversifié et inclusif, vous contribuerez à réduire ces risques et rendre les réponses de l'IA plus équitables.

4. AMÉLIORER LA CRÉATIVITÉ ET LA VARIABILITÉ DES RÉSULTATS

Dans les applications créatives, les prompts peuvent encourager les modèles d'IA à générer des résultats diversifiés et innovants. Les prompts d'écriture créative, par exemple, peuvent préciser le genre, le ton et des éléments de scénario particuliers afin d'influencer le style et l'atmosphère de la création de l'IA. Dans ce domaine, des prompts soigneusement rédigés permettent aux entreprises de produire un contenu à la fois unique et conforme aux directives de la marque ou aux formats voulus.

5. ATTÉNUER LES HALLUCINATIONS ET LES ERREURS

Les modèles d'IA générative sont puissants, mais ils ne sont pas infaillibles. Mal construits, les prompts peuvent provoquer des hallucinations, autrement dit, des réponses contenant des informations incorrectes ou inventées. L'ingénierie de prompt peut minimiser ces risques en circonscrivant nettement la tâche, en fournissant des instructions claires et en tenant compte des capacités de l'IA. Rappelons enfin que les résultats de l'IA doivent toujours être vérifiés, en particulier dans les applications qui exigent un haut niveau d'exactitude.

L'ingénierie de prompt dans le contexte de la création d'agents IA

Avec l'essor d'outils d'IA avancés tels que Databricks Mosaic AI, l'ingénierie de prompt a acquis une importance cruciale dans le développement d'agents IA capables de produire des résultats commerciaux. Databricks Mosaic AI fournit une plateforme complète pour la création d'applications d'IA, l'intégration de modèles de machine learning et la gestion des workflows d'IA à grande échelle. Dans l'écosystème Mosaic AI, l'ingénierie de prompt a principalement pour objet d'optimiser les entrées des modèles de Databricks, pour permettre aux organisations de créer des agents IA plus précis, plus efficaces et plus imaginatifs.

Databricks Mosaic AI simplifie l'intégration des LLM et d'autres modèles d'IA avancés, et propose des outils robustes pour le développement d'agents intelligents capables d'accomplir un large éventail de tâches. En rédigeant soigneusement des prompts axés sur l'objectif de chaque agent, qu'il s'agisse d'automatiser le support client ou de générer des insights sur les marchés, les organisations exploiteront tout le potentiel de leurs systèmes d'IA.

OPTIMISER LES AGENTS IA POUR DES TÂCHES SPÉCIFIQUES

Lorsque vous développez des agents IA sur Databricks Mosaic AI, l'ingénierie de prompt permet de les adapter à des tâches métier spécifiques et aux besoins des clients. Un prompt peut notamment amener l'agent IA à émettre des recommandations personnalisées en s'inspirant des préférences de l'utilisateur, ou à analyser l'historique d'un client pour lui apporter une assistance personnalisée. Des prompts bien pensés aideront l'agent IA à apporter une réponse à la fois précise et contextualisée, avec des effets positifs évidents sur l'expérience utilisateur et l'efficacité opérationnelle.

INTÉGRER LES RETOURS DES UTILISATEURS POUR UNE AMÉLIORATION CONTINUE

L'un des grands avantages de Databricks Mosaic AI réside dans la possibilité de suivre, d'analyser et d'améliorer les performances de vos agents IA en permanence. En testant différentes variantes de prompt, en collectant les commentaires des utilisateurs et en suivant les performances des agents, vous pouvez multiplier les itérations et affiner les prompts afin d'optimiser progressivement les résultats. Ce cycle d'amélioration continu des prompts assure la précision, la pertinence et l'efficacité de vos agents IA face à l'évolution des objectifs commerciaux.

En résumé, l'ingénierie de prompt est la pierre angulaire de la création d'agents IA robustes et fiables sur la plateforme Databricks Mosaic AI. Non seulement elle améliore les résultats des modèles d'IA, mais elle aide les organisations à développer des systèmes plus intelligents, flexibles et capables de s'adapter aux fluctuations du paysage numérique.

Chapitre 3 : Créer des systèmes d'agents IA

Les systèmes d'agents IA sont des applications avancées qui combinent différentes technologies d'IA pour accomplir des tâches complexes en toute autonomie, éclairer la prise de décision et améliorer l'efficacité opérationnelle. Les systèmes d'agents IA excellent dans de nombreux cas d'usage : automatisation du support client, recommandations personnalisées, détection de fraude et prévision des tendances. Ils intègrent l'analyse de données structurées à l'interprétation du langage naturel pour délivrer des informations précises et data-driven, tout en conservant la souplesse nécessaire pour traiter les informations non structurées. Ils se montrent ainsi extrêmement précieux dans tous les secteurs, aidant les entreprises à rationaliser leurs opérations, à renforcer l'engagement des utilisateurs et à déployer à grande échelle des solutions basées sur l'IA.

Avec sa plateforme unifiée conçue pour l'intégration et la gouvernance de nombreux composants d'IA, Databricks Mosaic AI permet de créer des systèmes d'agents IA de grande qualité. Reposant sur l'architecture du data lakehouse, Mosaic AI veille à ce que les systèmes d'agents puissent accéder en toute sécurité aux données de l'entreprise et les utiliser à des fins de personnalisation. Mosaic AI prend aussi bien en charge les modèles open source que propriétaires, pour permettre aux organisations de combiner les solutions les mieux adaptées à leurs besoins. Mosaic AI intègre un cadre de gouvernance rigoureux qui veille à ce que les systèmes d'agents IA restent performants, conformes et en phase avec les objectifs commerciaux. Ce cadre apporte une visibilité complète et un contrôle total, de l'importation des données au déploiement.

Qu'est-ce qu'un agent IA ?

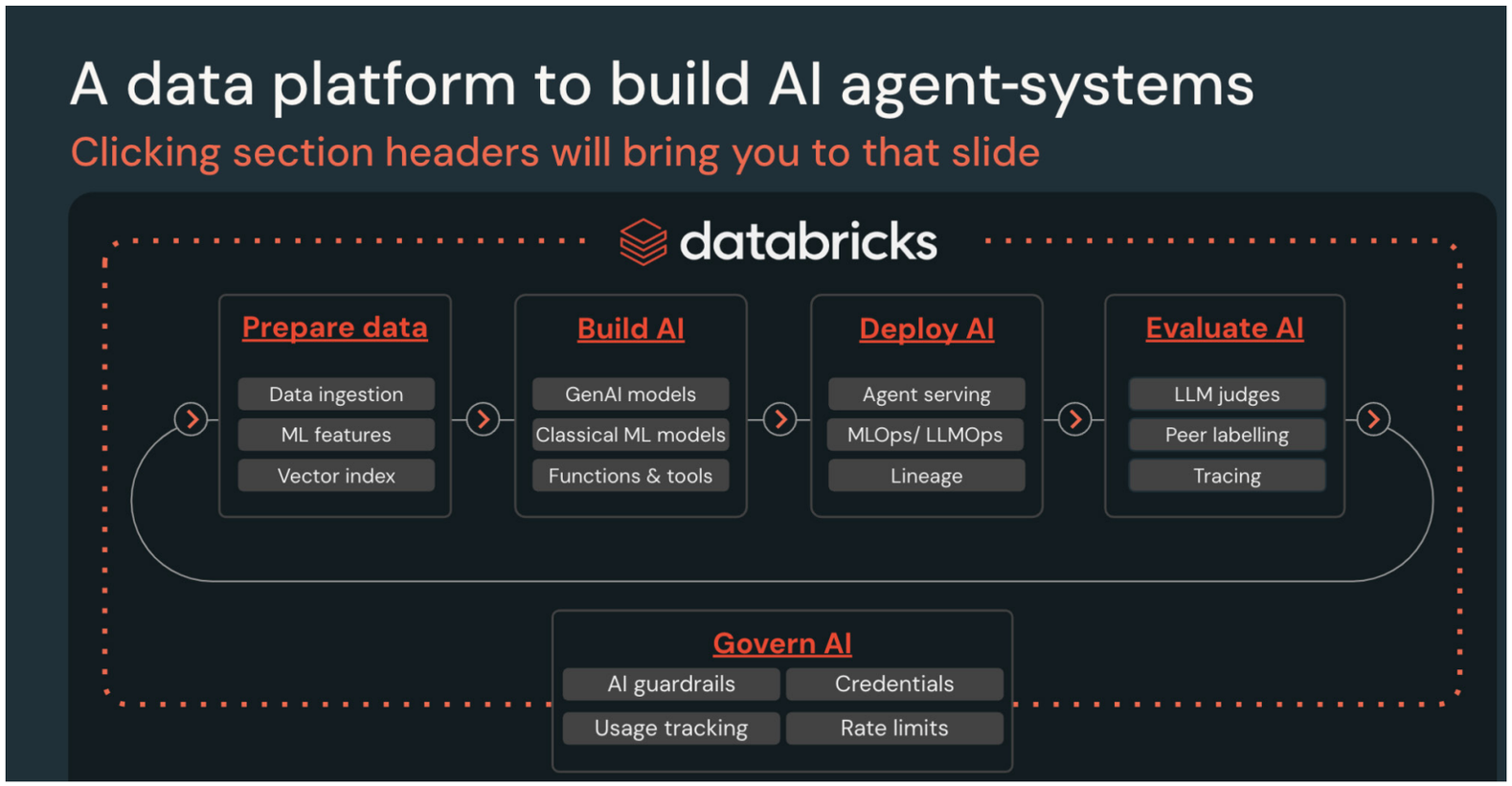
Qu'on veuille déployer un agent IA isolé ou plusieurs agents en interaction, il est indispensable de s'appuyer sur un système d'agents IA. Les agents IA sont des applications intelligentes conçues pour automatiser des tâches afin de soutenir la productivité humaine. Ces agents sont capables d'analyser les informations, de prendre des décisions et d'accomplir des actions pour atteindre des objectifs spécifiques. Ils libèrent ainsi les équipes qui peuvent consacrer davantage de temps et de ressources à des initiatives stratégiques. Quelques exemples :

- **Agent de service client :** interagit avec les clients, comprend leur demande et apporte des réponses pertinentes, afin de résoudre leurs problèmes et leur donner satisfaction.
- **Agent de génération de campagne :** analyse les données, identifie les publics cibles et génère des campagnes marketing personnalisées répondant à des objectifs précis, tels que l'amélioration de la visibilité de la marque ou l'augmentation des ventes.
- **Agent de génération de code :** assiste les développeurs en rédigeant, en déboguant et en optimisant du code en fonction des spécifications données, dans le but de fluidifier le processus de développement logiciel.

Qu'est-ce qu'un système d'agents IA ?

Un système d'agents IA permet de concevoir et d'opérationnaliser des agents intelligents capables d'exécuter des tâches complexes. Contrairement aux modèles autonomes, ces systèmes intègrent de multiples composants : grands modèles de langage, modèles de machine learning classiques, données d'entreprise et outils externes. C'est cette combinaison qui permet d'atteindre efficacement des objectifs spécifiques. Les systèmes d'agents IA intègrent également des techniques d'évaluation et des mécanismes de gouvernance qui garantissent la qualité, la traçabilité et le respect des normes organisationnelles.

Les étapes clés du développement d'un système d'agents IA



- **Préparation des données :** les données sont organisées et prétraitées pour qu'elles soient propres, accessibles et pertinentes à des fins de prise de décision et d'interaction avec les agents
- **Développement des agents :** des modèles d'IA générative, des modèles ML classiques et différents outils sont mobilisés pour créer des agents spécialisés.
- **Déploiement des agents :** les agents sont packagés pour que d'autres utilisateurs et systèmes puissent facilement les utiliser de façon sûre et performante
- **Évaluation des performances :** les résultats des agents sont mesurés et évalués pour vérifier qu'ils répondent bien aux objectifs fixés. Des améliorations itératives sont ensuite apportées sur la base des retours collectés.
- **Gouvernance des opérations :** les normes de sécurité, de conformité et d'éthique sont appliquées, et les activités des agents sont étroitement suivies dans un souci de transparence et de responsabilité

Composants fondamentaux d'un système d'agents IA

Les systèmes d'agents IA rassemblent plusieurs éléments de base pour offrir des fonctionnalités sophistiquées.

LLM/AGENT CENTRAL

Un système d'agent d'IA s'articule autour d'un LLM préentraîné et généraliste qui traite et comprend les données. Ces modèles sont activés à l'aide de prompts rédigés avec soin. Ceux-ci apportent des éléments de contexte indispensable, guident l'interaction, décrivent les objectifs poursuivis et les outils à utiliser.

Les cadres d'agent permettent de personnaliser le modèle afin qu'il adopte une identité distincte et devienne un expert du domaine d'intérêt. Grâce à cette flexibilité, l'agent LLM peut accomplir différentes tâches avec précision, d'où son grand intérêt pour les applications d'entreprise.

PLANIFICATEUR

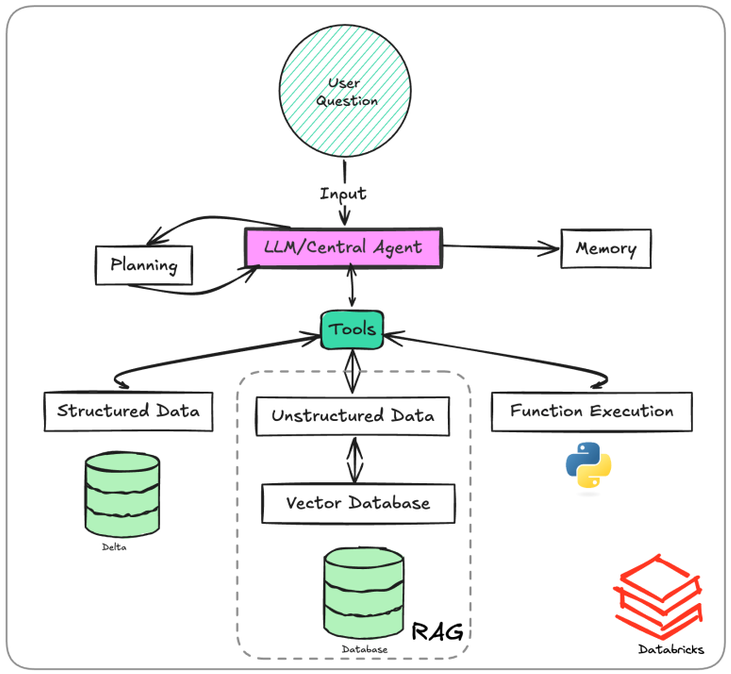
Un composant de planification décompose les tâches complexes en étapes plus simples et coordonne leur exécution. Il détermine la meilleure marche à suivre en employant des techniques de raisonnement comme la chaîne de pensée, ou bien des approches hiérarchiques telles que les arbres de décision.

Une fois le plan créé, il est révisé et affiné à l'aide de mécanismes de rétroaction internes tels que ReAct ou Reflexion, qui améliorent le processus de prise de décision de l'agent par itérations successives. Dans les systèmes à un seul agent, la planification et l'orchestration sont généralement confiées à un seul LLM. Les frameworks multiagents, en revanche, peuvent déléguer ces fonctions à des agents spécialisés.

OUTILS

Les outils sont les composants opérationnels d'un système d'agent IA. Ils accomplissent des tâches spécifiques en suivant les instructions du LLM central. Les outils prennent des formes très diverses : appels API, scripts Python, requêtes SQL, recherches web, et même des fonctionnalités d'entreprise personnalisées comme Databricks AI/BI Genie.

En intégrant des outils, les agents LLM peuvent exécuter des workflows, recueillir des observations et réaliser des tâches secondaires avec une grande efficacité. Toutefois, lors du développement de ces applications, il est essentiel de prendre en compte la durée des interactions. Les longues conversations peuvent épuiser la limite de contexte d'un LLM qui risque alors d'oublier des informations. Il faut donc gérer correctement le flux de contrôle – monothread, multithread ou en boucle – pour prendre en charge des chaînes de décision de plus en plus complexes.



Dans la figure, la prise de décision repose sur un seul LLM hautement performant. Il interprète la question de l'utilisateur, puis détermine le chemin à suivre pour acheminer le flux décisionnel. Il a plusieurs outils à sa disposition pour accomplir certaines opérations, stocker des résultats intermédiaires en mémoire, planifier les étapes suivantes et renvoyer un résultat à l'utilisateur.

MÉMOIRE

La mémoire peut considérablement améliorer l'architecture d'un agent IA en lui permettant de stocker et de mobiliser des informations à des fins de prise de décision.

- **Mémoire à court terme :** stockage temporaire destiné au contexte immédiat ; il est effacé une fois la tâche terminée.

- **Mémoire à long terme :** stockage persistant, souvent hébergé dans des bases de données externes (des bases de données vectorielles, bien souvent). Il aide l'agent à reconnaître des modèles et à tirer des enseignements des interactions précédentes, et informe les décisions qu'il prend par la suite.

En combinant les deux types de mémoire, les moyens sont donnés aux agents de fournir des réponses personnalisées et de s'adapter aux préférences des utilisateurs au fil du temps. Il faut faire une distinction entre la mémoire d'un agent et la mémoire conversationnelle d'un LLM, car elles n'ont pas le même usage.

Cas d'usage des systèmes d'agents IA

Financial Services	Healthcare & Life Sciences	Comm, Media & Entertainment	Retail & Consumer Goods	Manufacturing	Public Sector
Fraud monitoring and predictions	Biomedical literature summarization & discovery	Hyper-personalization for customer experience (CX)	Try before you buy with virtual fitting rooms	Delightful, personalized customer experiences	Analysis of open-source Intelligence
Automating compliance data gathering	Clinical trial optimization	Enhancing customer support and self-service	Optimizing demand prediction and inventory	Increasing productivity and efficiency in operations	Modernizing legacy code bases
Accelerate underwriting and claims processing in insurance	Health insurance claim processing	Intelligent content creation and curation	Generate innovative product designs	Prescriptive and proactive field service	Regulatory compliance assistance

Les systèmes d'agents IA sont des solutions polyvalentes mises au service d'un large éventail de fonctions métier pour en tirer des avantages mesurables en termes d'efficacité, de précision et de productivité. Vous trouverez ci-dessous quelques exemples pratiques d'utilisation de systèmes d'agents IA en environnement réel :

AUTOMATISATION DU SUPPORT CLIENT

Les systèmes d'agents IA peuvent fluidifier les interactions avec les clients en déployant des chatbots intelligents ou des assistants virtuels capables de donner rapidement des réponses précises aux questions de routine. Cette approche allège la charge de travail des agents humains et leur permet de se concentrer sur les cas plus complexes. Un système d'agents IA va, par exemple, gérer les questions courantes de suivi des commandes ou de dépannage. Il saura aussi transmettre automatiquement les problèmes complexes aux humains.

APPUI DES ÉQUIPES DE VENTES ET DE MARKETING

Les systèmes d'agents IA qui automatisent les tâches répétitives, comme la qualification des prospects, les e-mails de suivi et la saisie de données, peuvent être très utiles aux équipes de ventes et de marketing. Ils peuvent également analyser les données clients pour produire des insights et des recommandations personnalisées qui vont améliorer l'efficacité des campagnes. Un système d'agents IA peut, par exemple, trier les prospects à contacter en priorité en fonction d'indicateurs d'engagement. Ou bien recommander des promotions personnalisées en fonction des préférences des clients.

ANALYSE DE DONNÉES ET REPORTING

Les systèmes d'agents IA excellent dans l'automatisation de la collecte de données, de l'analyse et du reporting. Capables de traiter de grandes quantités d'informations en temps réel, ils permettent aux entreprises d'obtenir des renseignements pertinents à la volée, sans intervention manuelle. Une entreprise de commerce de détail peut ainsi extraire des données d'une variété de sources pour alimenter des rapports de ventes hebdomadaires ou des tableaux de bord de performance, sans agréger manuellement les données.

RECOMMANDATIONS PERSONNALISÉES

En analysant le comportement et les préférences des utilisateurs, les systèmes d'agents IA peuvent émettre des recommandations finement personnalisées, ce qui est particulièrement prisé dans le commerce en ligne, le divertissement et le tourisme, notamment. Un détaillant peut exploiter un système d'agents IA capable d'analyser l'historique de navigation et d'achat d'un client pour lui recommander des produits, tandis qu'une plateforme de streaming peut suggérer des contenus en s'appuyant sur les habitudes de visionnage des utilisateurs.

AUTOMATISATION DES TÂCHES DANS LES OPÉRATIONS

Dans les opérations, les systèmes d'agents IA ont le pouvoir d'automatiser de nombreuses tâches de routine, comme la planification, la gestion des stocks et le traitement des commandes, offrant des gains considérables d'efficacité et de précision. Dans un atelier de fabrication, par exemple, un système d'agents IA peut suivre les niveaux d'inventaire et déclencher automatiquement des commandes en dessous d'un certain seuil, pour éviter à la fois les erreurs humaines et les ruptures de stock.

DÉTECTION ET PRÉVENTION DES FRAUDES

Les systèmes d'agents IA sont très efficaces lorsqu'il s'agit d'identifier et d'atténuer la fraude : ils surveillent les transactions en permanence et détectent les anomalies en temps réel. Une institution financière s'appuiera sur un système de ce type pour être avertie en cas d'activité suspecte (retraits inhabituels ou autres) afin d'enquêter immédiatement. La détection précoce des menaces est un puissant atout pour les entreprises qui cherchent à réduire les risques liés à la fraude.

Cas d'usage spécialisés

Les **systèmes de génération augmentée par récupération**, qui combinent des LLM à des mécanismes de récupération de données pour fournir des réponses plus riches et contextualisées, illustrent parfaitement la puissance des systèmes d'agents IA, mais ils sont loin d'exploiter toutes leurs possibilités. Les systèmes de RAG conviennent particulièrement aux applications qui exploitent des données non structurées, comme le support client et l'extraction de connaissances, mais les capacités d'autres types de systèmes dépassent largement ce cadre.

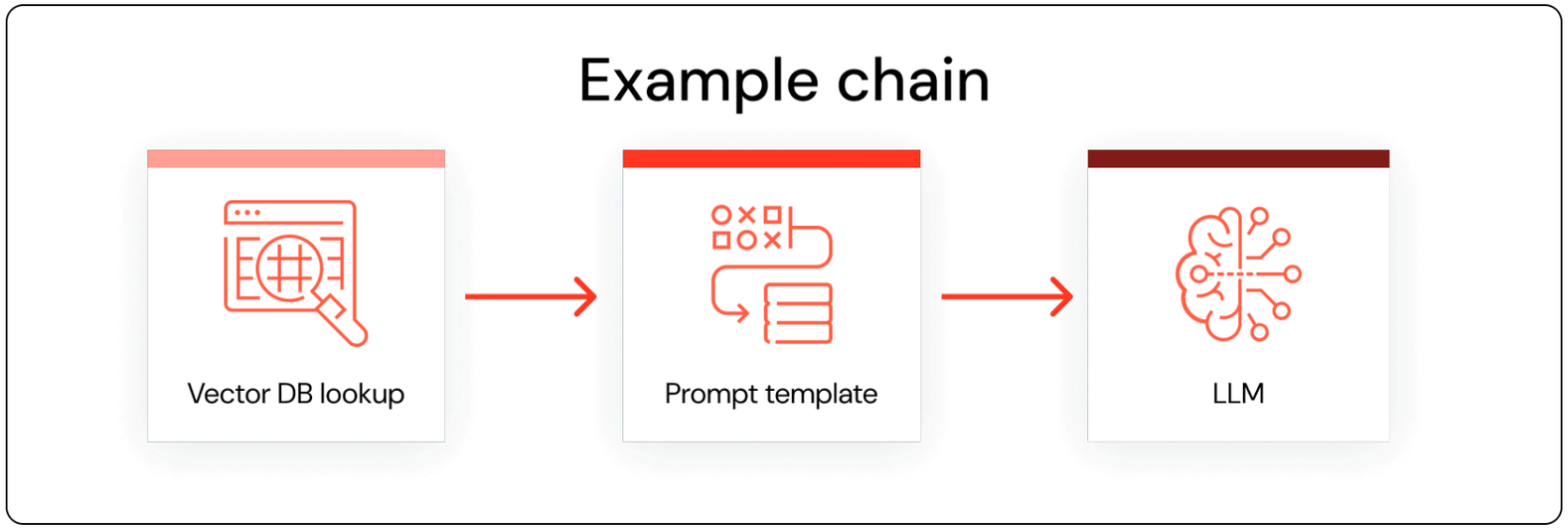
Dans les **environnements de données structurées**, les modèles de machine learning classiques occupent un rôle central. Associés aux LLM, ces modèles améliorent la prise de décision en apportant des informations précises et étayées. Dans la détection des fraudes, par exemple, les modèles ML classiques vont analyser les journaux de transactions structurés pour identifier des schémas suspects, tandis qu'un agent basé sur un LLM se chargera des interactions client liées aux transactions signalées.

On peut aussi prendre l'exemple des **systèmes multiagents**, dans lesquels plusieurs agents spécialisés collaborent pour accomplir des tâches complexes. Ces systèmes comprennent souvent des agents spécialisés de planification, d'exécution et de récupération de données. En orchestrant ces agents, les entreprises peuvent créer des solutions d'IA d'une grande souplesse, capables de gérer des workflows complexes afin d'optimiser la chaîne d'approvisionnement ou l'engagement client multicanal.

Chapitre 4 : Associer des bases de connaissances aux LLM

La génération augmentée par récupération permet aux organisations d'enrichir les systèmes d'IA prêts à l'emploi en intégrant des connaissances externes. Les systèmes gagnent alors en précision et en pertinence sans que leur comportement sous-jacent soit altéré. En combinant les prompts utilisateur et l'extraction d'informations pertinentes, la RAG génère des prompts enrichis grâce auxquels les grands modèles de langage comme GPT-4 ou Llama 3 délivrent des réponses plus précises et plus intéressantes.

Contrairement à l'ingénierie de prompt seule, la RAG implique la mise en place d'un système de récupération sophistiqué – reposant généralement sur une base de données vectorielle – et l'intégration transparente des données extraites dans le workflow du LLM. Parce qu'elle améliore considérablement la qualité des résultats générés par l'IA, cette couche supplémentaire de complexité s'avère particulièrement rentable.



L'INTÉRÊT DE LA RAG POUR L'ACCÈS AUX INFORMATIONS EXTERNES

La RAG offre plusieurs avantages décisifs pour l'intégration de données externes :

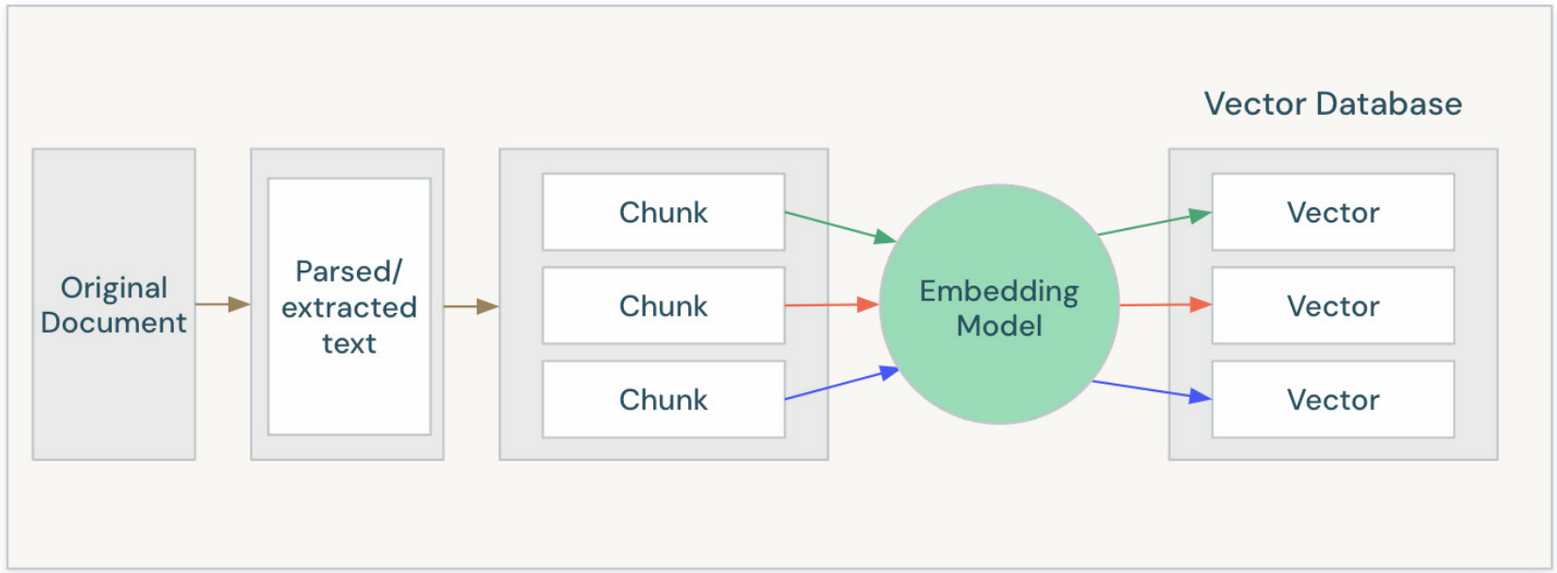
1. **Modularité des données :** il est possible d'ajouter et de supprimer des sources de données sans réentraîner le modèle, ce qui offre une flexibilité maximale.
2. **Contrôle des accès :** les administrateurs peuvent appliquer des autorisations détaillées pour réserver l'accès à des datasets spécifiques aux seuls utilisateurs autorisés.
3. **Flexibilité du choix des modèles :** La RAG permet de comparer facilement différents LLM sans réentraînement intensif avec de nouvelles données.

Les systèmes RAG consomment généralement moins de ressources que le préentraînement et l'ajustement, mais leur coût et leur complexité varient en fonction de l'échelle et de l'architecture du système de récupération. Une base de données vectorielle à faible latence, par exemple, qui est capable de gérer des millions d'enregistrements, coûtera plus cher qu'un système plus réduit, à la latence plus élevée. Malgré ces coûts, la RAG reste un moyen efficace d'enrichir les capacités d'un LLM sans investissement massif au départ.

Améliorer les performances des modèles avec la RAG

La RAG offre plusieurs avantages en termes de performances : les hallucinations sont moins nombreuses, les modèles comprennent mieux le domaine et les réponses sont plus précises. Rappelons toutefois que l'efficacité de la RAG dépend directement des capacités du LLM sous-jacent. Si ses performances ne sont pas à la hauteur, il faudra peut-être envisager des stratégies de personnalisation plus sophistiquées, comme l'ajustement ou le préentraînement continu.

Les organisations ont tout intérêt à acquérir des connaissances fondamentales sur le fonctionnement des LLM et à explorer les possibilités de la RAG. Elles parviendront ainsi à identifier les problèmes de performance et à faire un usage optimal de leurs ressources avant d'opter pour des solutions plus complexes.



SOURCES DE DONNÉES ET WORKFLOWS DE LA RAG

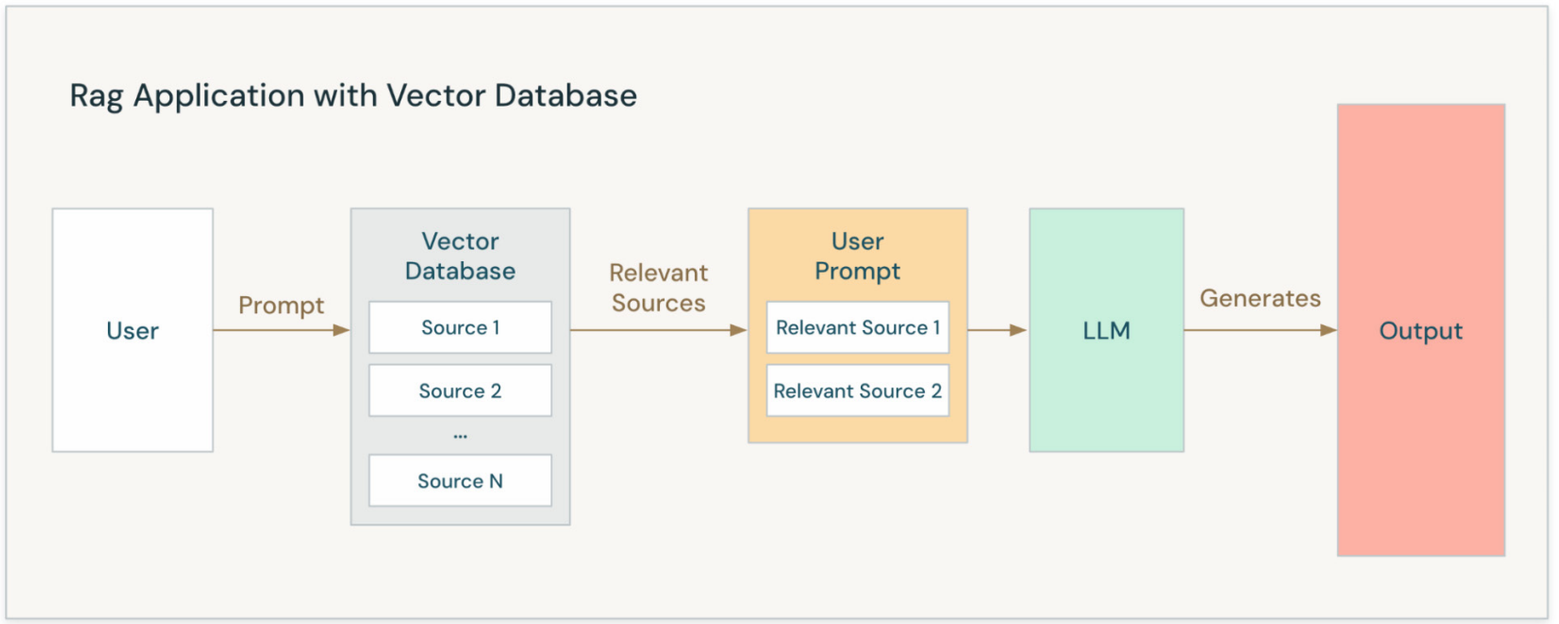
Les systèmes RAG peuvent traiter de nombreux types de données :

- **Texte :** PDF, articles et référentiels de code
- **Multimédia :** podcasts et vidéos
- **Bases de données structurées :** bases de données relationnelles ou NoSQL

Données non structurées et bases de données vectorielles

Parce qu'elles sont dépourvues d'organisation a priori, les données non structurées sont difficiles à interroger directement. Les workflows RAG transforment les données brutes et non structurées en blocs discrets et interrogeables en plusieurs étapes :

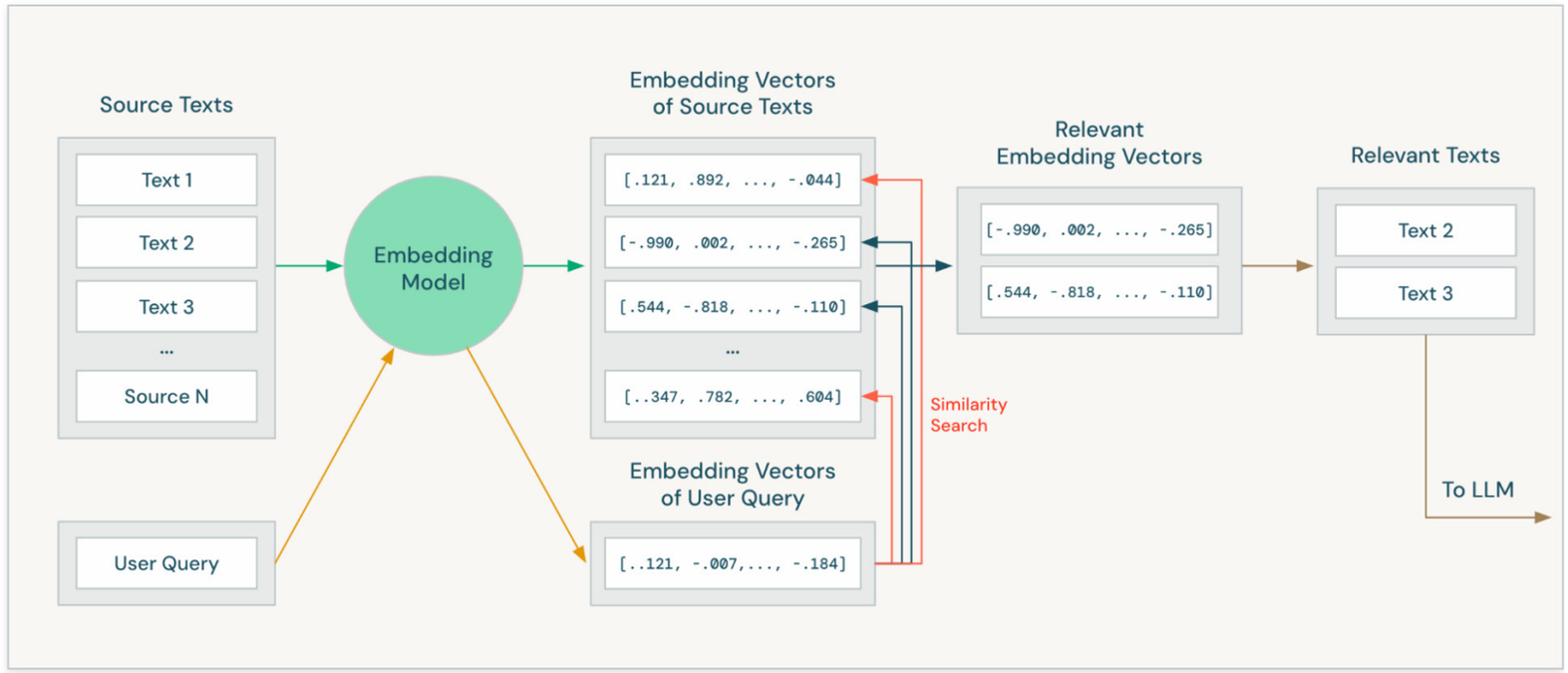
1. **Analyse des documents bruts :** les fichiers PDF, les images ou les fichiers multimédias sont convertis en formats textuels utilisables à l'aide d'outils de reconnaissance optique de caractères (OCR) et d'extraction de texte.
2. **Extraction des métadonnées (facultatif) :** les métadonnées (titres de documents, URL, etc.) sont extraites et exploitées afin d'améliorer la précision des requêtes.
3. **Fragmentation des documents :** les documents analysés sont découpés en petits fragments discrets compatibles avec la fenêtre contextuelle du modèle d'intégration et du LLM.
4. **Intégration des fragments :** des modèles d'intégration sont utilisés pour transformer les fragments en vecteurs numériques de haute dimension qui encodent leur sémantique.
5. **Indexation dans une base de données vectorielle :** les intégrations sont stockées avec le texte correspondant dans une base de données vectorielle optimisée pour une récupération rapide et efficace.



Optimisation de la recherche vectorielle

La recherche vectorielle est au cœur des applications de RAG. Elle identifie les informations utiles en intégrant les requêtes des utilisateurs et en les comparant aux intégrations stockées. L'efficacité de la récupération repose sur plusieurs facteurs :

- **Modèles d'intégration :** les modèles d'intégration utilisés doivent être précis et savoir capturer la sémantique des données.
- **Index vectoriels :** ces index organisent les intégrations de manière à accélérer les calculs de similarité
- **Algorithmes :** les algorithmes de récupération doivent équilibrer précision et efficacité pour produire des résultats rapides et exacts



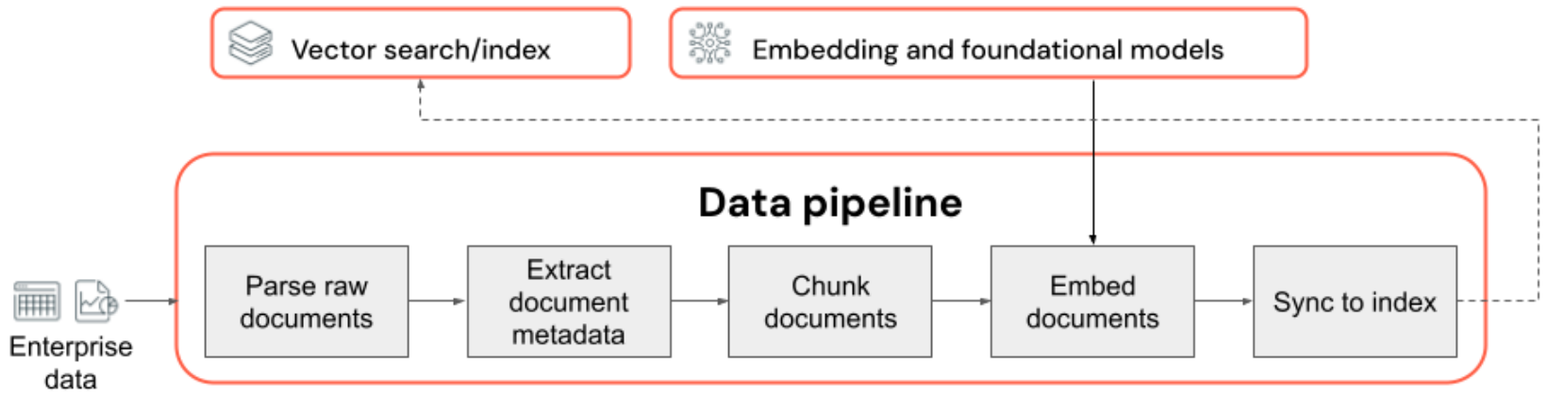
Qualité de la RAG

D'un point de vue conceptuel, il est intéressant d'envisager les curseurs de qualité de la RAG par le prisme de deux problèmes majeurs possibles :

- **Qualité de la récupération :** récupérez-vous réellement les informations les plus utiles pour une requête donnée ? La RAG pourra difficilement produire des résultats de haute qualité s'il manque des informations importantes dans le contexte fourni au LLM, ou au contraire s'il regorge d'informations superflues.
- **Qualité de la génération :** en examinant les informations récupérées et la requête initiale de l'utilisateur, le LLM génère-t-il la réponse la plus précise, cohérente et utile possible ?
  Différents problèmes peuvent se manifester à ce stade : des hallucinations, des résultats incohérents ou une réponse hors sujet.

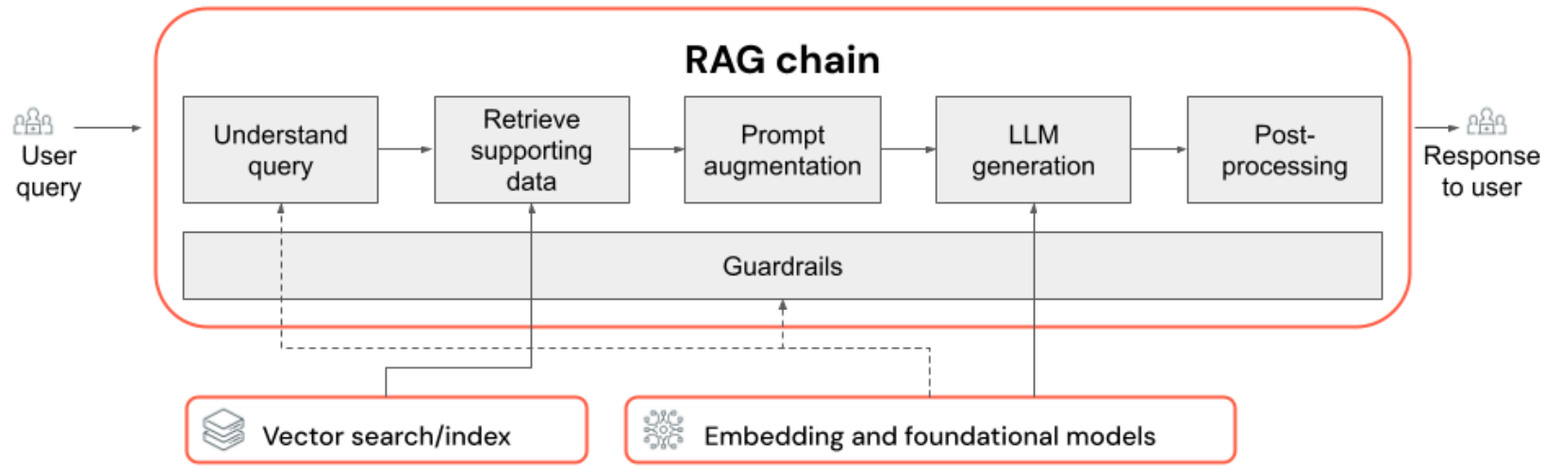
Les applications de RAG comportent deux composants sur lesquels il est possible d'intervenir pour corriger les problèmes de qualité : le pipeline de données et la chaîne. Il est tentant d'établir une distinction claire entre les problèmes de récupération (qui concernent le pipeline de données) et les problèmes de génération (qui touchent à la chaîne RAG). Comme souvent, la réalité est plus subtile. La qualité de la récupération subit à la fois l'influence du pipeline de données (stratégie d'analyse et de fragmentation, exploitation des métadonnées, modèle d'intégration) et celle de la chaîne RAG (transformation des requêtes utilisateur, nombre de fragments récupérés, reclassement). De même, la qualité de la génération sera nécessairement affectée par une récupération médiocre : les informations inutiles et les lacunes vont dégrader la sortie du modèle.

Ce chevauchement souligne la nécessité d'une approche holistique de l'amélioration de la qualité des RAG. Il faut localiser les composants à modifier dans le pipeline de données comme dans la chaîne RAG, puis anticiper l'impact des modifications sur la solution globale, afin d'effectuer des mises à jour ciblées qui amélioreront la qualité des résultats de la RAG.



Les facteurs de qualité du pipeline de données :

- Composition du corpus de données d'entrée
- Méthode employée pour extraire les données brutes et les transformer en un format utilisable (analyse d'un document PDF, par exemple)
- Modalités de fragmentation des documents et de mise en forme des fragments obtenus (stratégie de découpage et taille des fragments)
- Nature des métadonnées (titres de section ou de document) extraites de chaque document et/ou fragment, et incorporation de ces métadonnées dans chaque bloc
- Modèle d'intégration utilisé pour convertir le texte en représentations vectorielles pour la recherche de similarité



- Choix du LLM et de ses paramètres (température et nombre maximum de jetons)
- Paramètres de récupération (nombre de fragments ou de documents récupérés)
- Approche de récupération (recherche par mot-clé, hybride ou sémantique, réécriture de la requête utilisateur, transformation de la requête utilisateur en filtres, ou reclassement)
- Composition du prompt intégrant le contexte récupéré pour amener le LLM à produire un résultat de qualité

Un ensemble d'évaluation vous aidera à mesurer les performances de différents aspects de votre application de RAG :

- **Qualité de la récupération :** les métriques de récupération évaluent la capacité de votre application RAG à extraire les bonnes données à l'appui. À cet égard, la précision et le rappel sont deux indicateurs essentiels.
- **Qualité de la réponse :** les métriques de qualité de la réponse évaluent dans quelle mesure l'application RAG répond à la demande de l'utilisateur. Elles mesurent, par exemple, si la réponse obtenue est exacte par rapport à la vérité attestée, dans quelle mesure la réponse était fondée sur le contexte récupéré (présence ou non d'hallucinations) et si elle était sûre (dépourvue de toxicité).
- **Performances du système (coût et latence) :** les métriques déterminent le coût global et les performances des applications RAG. La performance de la chaîne se mesure notamment en termes de latence globale et de consommation de jetons.

Il est essentiel de collecter des métriques concernant à la fois les réponses et la récupération. Une application RAG peut en effet produire des réponses médiocres tout en extrayant correctement le contexte ; elle peut aussi fournir de bonnes réponses en s'appuyant sur des récupérations erronées.

Techniques de récupération avancées

Au-delà de la recherche vectorielle standard, plusieurs techniques permettent d'améliorer la précision de la récupération :

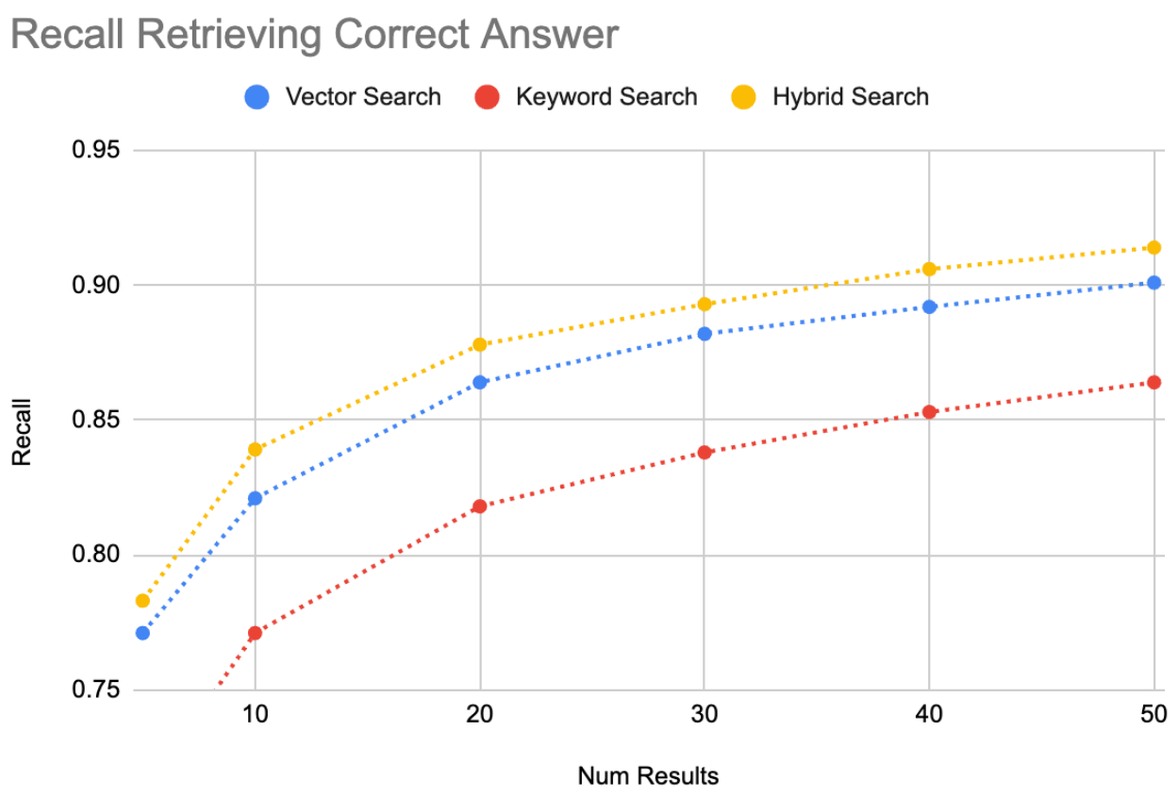
1. **Recherche hybride :** combine la recherche traditionnelle par mots-clés avec la recherche vectorielle. Cette approche est particulièrement utile lorsque les datasets contiennent des mots-clés stratégiques comme des codes de produits ou des termes spécifiques.
2. **Reclassement :** utilise un modèle supplémentaire pour réorganiser les résultats récupérés au départ et mettre en avant les plus pertinents.
3. **Comparaison de textes résumés :** intègre des versions résumées des textes plutôt que du texte brut pour accélérer la reconnaissance.
4. **Récupération de fragments contextuels :** récupère les fragments adjacents (les paragraphes précédents et suivants, par exemple) pour fournir un contexte plus riche.
5. **Affinement des prompts :** utilise un modèle de langage pour affiner le prompt initial de l'utilisateur afin d'améliorer la pertinence de la recherche.
6. **Ajustement propre au domaine :** ajuste les modèles d'intégration pour des domaines spécifiques afin d'améliorer la précision de la récupération dans les applications spécialisées.

IMPLÉMENTATION DE LA RECHERCHE HYBRIDE

La recherche hybride améliore le rappel en ajoutant un index de mots-clés appris à l'index vectoriel. Cet index, entraîné sur le dataset, capture les mots-clés et les identifiants critiques indispensables à une récupération précise. La recherche hybride est particulièrement efficace lorsqu'il est nécessaire de retrouver des mots clés exacts, comme dans le cas de codes de produits ou de termes techniques.

La recherche hybride est simple à mettre en œuvre, car la plupart des systèmes d'indexation actuels la prennent en charge sans configuration supplémentaire. L'index des mots-clés est entraîné sur tous les champs de texte du corpus pour permettre l'interrogation des fragments de texte et des champs de métadonnées.

Dans les applications RAG, le **rappel** est un indicateur clé de performance. Il désigne la proportion de requêtes pour lesquelles le fragment correct est récupéré dans les premiers résultats. La recherche hybride améliore le rappel en réduisant le nombre de fragments que le LLM doit traiter, et donc également la latence et les coûts de traitement. Les benchmarks internes montrent une amélioration de 20 % du rappel ; autrement dit, le nombre de documents requis pour un rappel élevé passe de 50 à 40 dans les datasets classiques.



Notre implémentation de la recherche hybride s'appuie sur la fusion de classement réciproque (RRF) des résultats de la recherche vectorielle et de la recherche par mots-clés. La RRF est paramétrée de manière à renvoyer des résultats de haute qualité pour la plupart des datasets.

Les scores sont normalisés, et le score le plus élevé possible est de 1,0. Cela permet d'identifier facilement dans quels cas les documents sont considérés comme de grande valeur par la recherche vectorielle et la recherche par mots-clés. Des scores proches de 1,0 signifient que les deux méthodes ont trouvé le document très pertinent. Les scores approchant de 0,5 et inférieurs indiquent qu'au moins l'un des outils de recherche estime le document peu utile.

Pour plus d'informations, consultez notre documentation sur la recherche hybride :

- Détails des calculs de recherche de similarité avec la recherche hybride
- SDK Python pour similarity_search

Amélioration continue des systèmes RAG

Les systèmes RAG peuvent être améliorés continuellement grâce à des tests et des ajustements itératifs. Plusieurs aspects se prêtent particulièrement à l'optimisation :

- **Mises à jour des données :** mise à jour régulière de la base de données vectorielles afin d'inclure de nouvelles informations pertinentes
- **Ajustement des paramètres :** ajustement continu de la taille des fragments, des limites de récupération et des modèles d'intégration
- **Surveillance et évaluation :** utilisation d'outils comme Databricks MLflow pour suivre les indicateurs de performance afin de maintenir une qualité constante

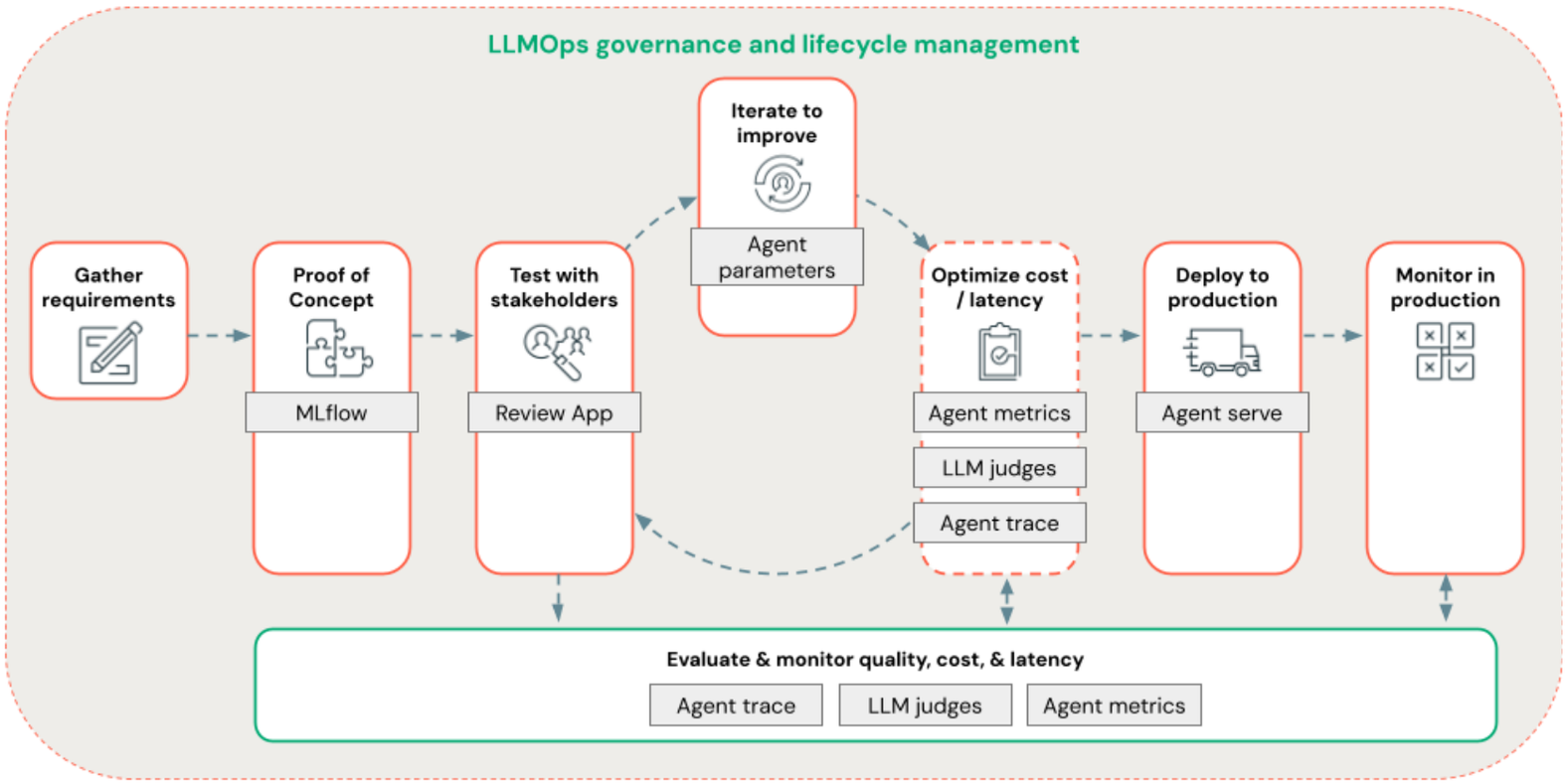
En adoptant une approche systématique de l'amélioration, les organisations préserveront l'efficacité et l'évolutivité de leurs systèmes RAG, qui sauront s'adapter à l'évolution de leurs besoins.

Chapitre 5 : Performances et évaluations de l'IA générative

Mosaic AI Agent Evaluation

Mosaic AI Agent Evaluation aide les développeurs à évaluer la qualité, le coût et la latence des applications d'IA agentiques, catégorie qui inclut les applications et les chaînes RAG. Agent Evaluation a pour but d'identifier les problèmes de qualité et d'en déterminer la cause profonde. Les fonctionnalités d'Agent Evaluation s'intègrent aux phases de développement, de préparation et de production du cycle de vie MLOps. L'ensemble des métriques et données d'évaluation sont consignées dans les cycles MLflow.

Agent Evaluation délivre des techniques d'évaluation sophistiquées et fondées sur la recherche sous la forme d'un SDK doublé d'une interface conviviale, directement intégré à votre lakehouse, à MLflow et aux autres composants de la Databricks Data Intelligence Platform. Développée en collaboration avec la recherche Mosaic AI, cette technologie propriétaire propose une approche complète pour l'analyse et l'amélioration des performances des agents.



databricks

Les applications d'IA agentique sont complexes et reposent sur un grand nombre de composants. Cela explique que leurs performances soient plus délicates à évaluer que celles des modèles ML traditionnels. Les indicateurs qualitatifs et quantitatifs employés pour évaluer la qualité sont intrinsèquement plus complexes. Agent Evaluation mobilise des juges LLM et des métriques propriétaires pour évaluer la qualité de la récupération et des requêtes, ainsi que des indicateurs de performances généraux tels que la latence et le coût en jetons.

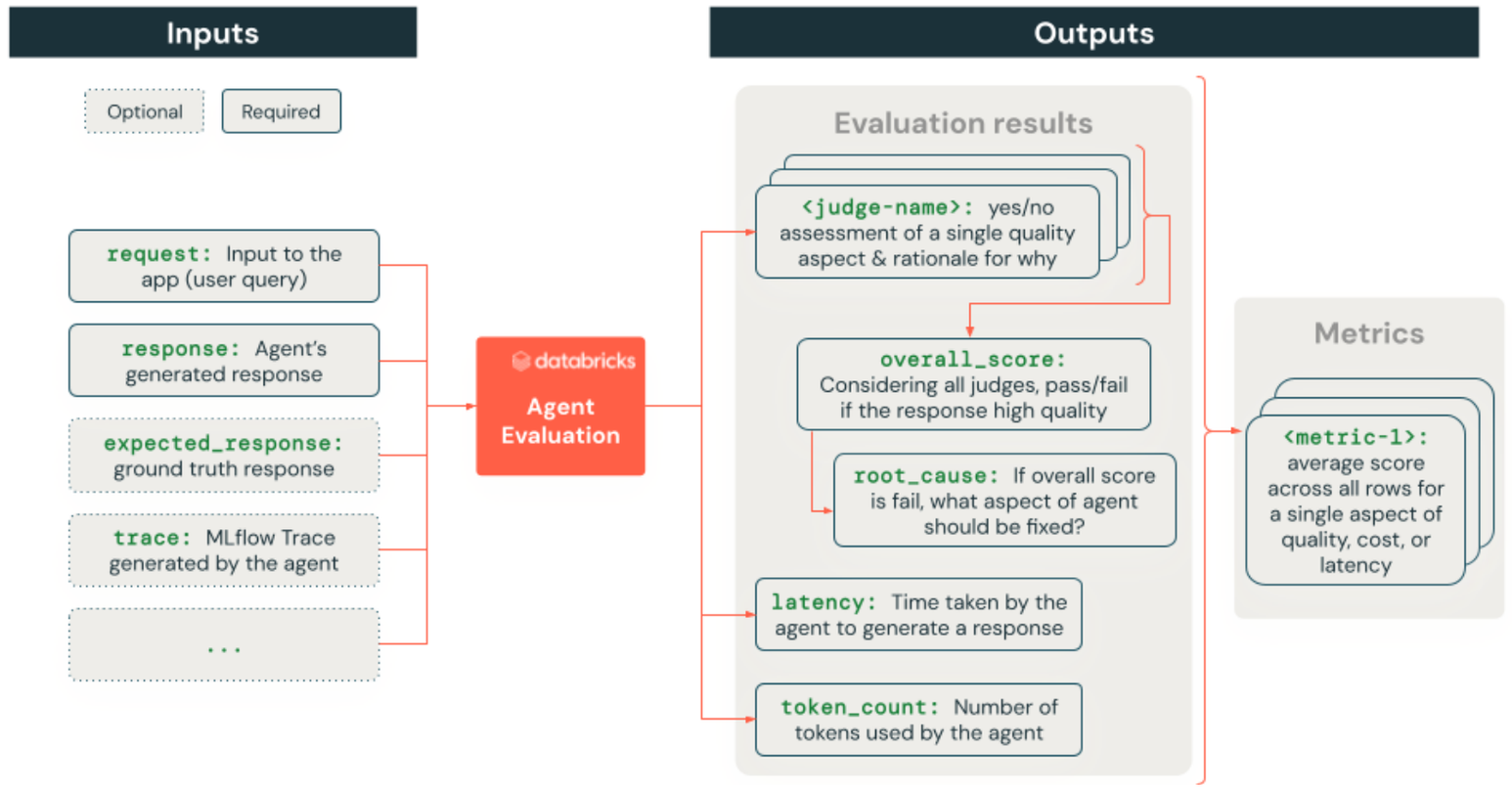
Mosaic AI Agent Framework intègre à Databricks un ensemble d'outils conçus pour aider les développeurs à créer, déployer et évaluer des agents de qualité production, notamment des applications RAG. Il est compatible avec les cadres tiers tels que LangChain et LlamaIndex. Vous pouvez donc utiliser votre framework de développement habituel tout en profitant des avantages du Unity Catalog géré de Databricks, du cadre Agent Evaluation et des autres fonctionnalités de la plateforme.

Accélérez les itérations grâce aux fonctionnalités suivantes :

- Créez et enregistrez des agents en utilisant MLflow et la bibliothèque de votre choix. Paramétrez vos agents pour accélérer l'expérimentation et les itérations au cours du développement.
- Utilisez le traçage des agents pour enregistrer, analyser et comparer les traces de votre code afin de le déboguer et de comprendre comment votre agent répond aux demandes.
- Améliorez la qualité des agents avec DSPy. DSPy peut automatiser l'ingénierie de prompt et l'ajustement pour rehausser la qualité de vos agents d'IA générative.
- Déployez des agents en production grâce à la prise en charge native du streaming de jetons et de la journalisation des demandes/réponses. Et misez sur l'application de feedback intégrée pour recueillir l'avis des utilisateurs de votre agent.

AGENT EVALUATION : ENTRÉES ET SORTIES

Le schéma suivant offre une vue d'ensemble des entrées acceptées par Agent Evaluation et des sorties correspondantes.



Pour plus d'informations sur l'entrée attendue par Agent Evaluation, et pour connaître les noms de champs et les types de données, consultez le schéma d'entrée. Voici une sélection de champs :

- **Requête de l'utilisateur (request) :** entrée de l'agent (question ou requête de l'utilisateur). Par exemple, « Qu'est-ce que la RAG ? »
- **Réponse de l'agent (response) :** réponse générée par l'agent. Par exemple, « La génération augmentée par récupération est… »
- **Réponse attendue (expected_response) :** (facultatif) réponse correcte donnant la vérité attestée.
- **Trace MLflow (trace) :** (facultatif) trace MLflow de l'agent, à partir de laquelle Agent Evaluation extrait des sorties intermédiaires comme le contexte extrait ou les appels d'outils. Vous avez également la possibilité de fournir ces sorties intermédiaires directement.
- **Lignes directrices (guidelines) :** (facultatif) liste de directives que la sortie du modèle est censée respecter.

Sur la base de ces entrées, Agent Evaluation produit deux types de sorties :

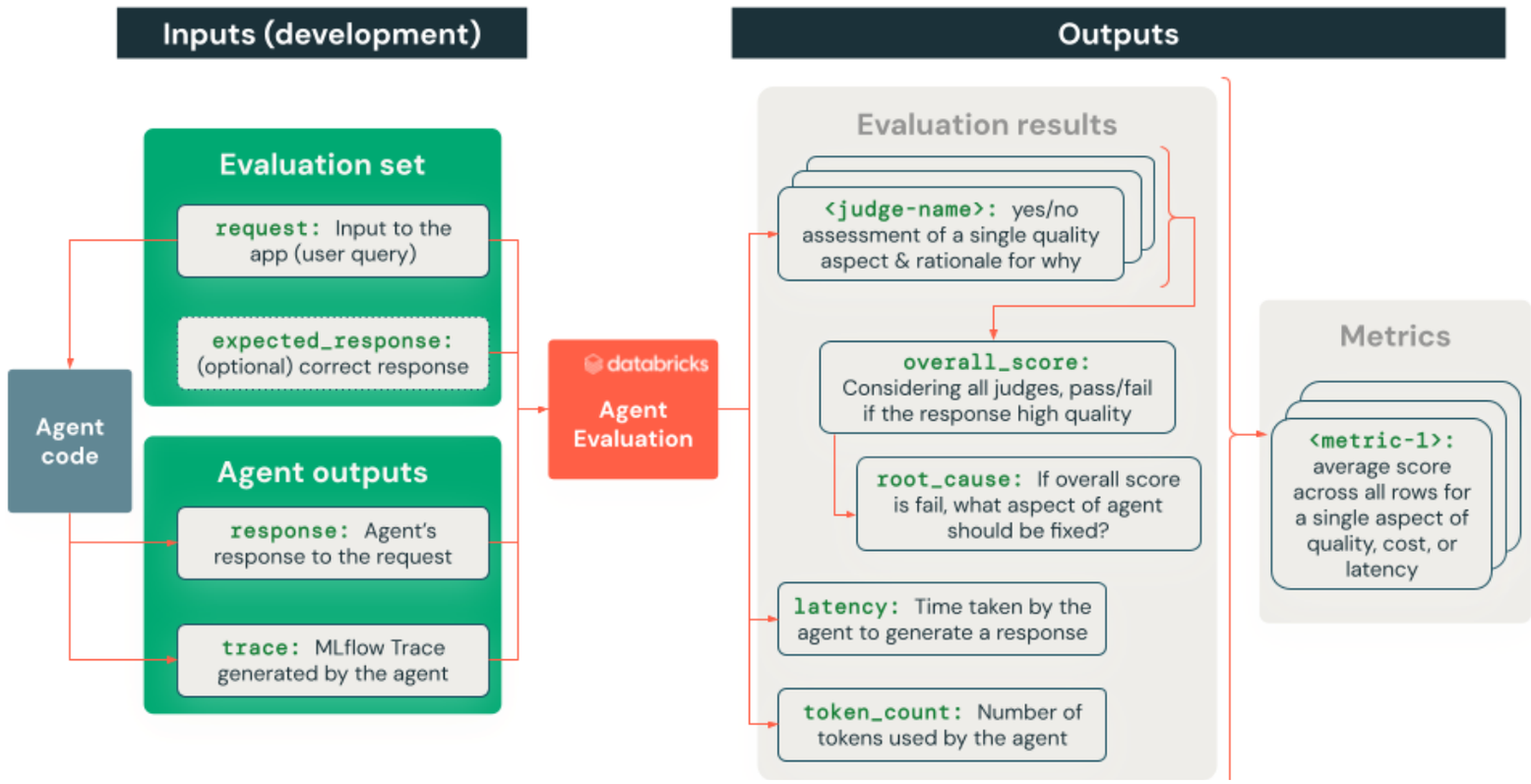
1. **Résultats de l'évaluation (par ligne) :** pour chaque ligne fournie en entrée, Agent Evaluation produit une ligne de sortie correspondante qui contient une évaluation détaillée de la qualité, du coût et de la latence de votre agent.
   - Les juges LLM examinent différents aspects de la qualité, dont l'exactitude et le bien-fondé, puis attribuent une note oui/non accompagnée d'une justification écrite. Pour plus de détails, voir Comment Agent Evaluation évalue la qualité, le coût et la latence.
   - Les évaluations des juges du LLM sont combinées pour produire un score global qui indique si cette ligne « réussit » (sa qualité est élevée) ou « échoue » (elle présente un problème de qualité).
     - La cause profonde est identifiée pour chaque ligne problématique. Elle correspond à l'évaluation d'un juge LLM spécifique : vous pouvez donc utiliser la justification du juge en question pour trouver des pistes de résolution.
   - Le coût et la latence sont extraits de la trace MLflow. Pour plus d'informations, voir Évaluation du coût et de la latence.
2. **Métriques (scores agrégés) :** scores agrégés qui résument la qualité, le coût et la latence de votre agent en tenant compte de toutes les lignes d'entrée. Ces scores mesurent notamment le pourcentage de réponses correctes, le nombre moyen de jetons, la latence moyenne et bien d'autres aspects. Pour plus d'informations, voir Évaluation du coût et de la latence et Agrégation des métriques de qualité, de coût et de latence au niveau d'un cycle MLflow.

DÉVELOPPEMENT (ÉVALUATION HORS LIGNE) ET PRODUCTION (SURVEILLANCE EN LIGNE)

Agent Evaluation couvre de façon cohérente vos environnements de développement (hors ligne) et de production (en ligne). Ce fonctionnement assure un passage en douceur du développement à la production, ce qui vous permet d'accélérer les itérations, puis d'évaluer, déployer et surveiller vos applications agentiques pour atteindre la meilleure qualité.

La principale différence entre le développement et la production réside dans la présence d'étiquettes de vérité attestée. Ces étiquettes permettent à Agent Evaluation de calculer des mesures de qualité supplémentaires.

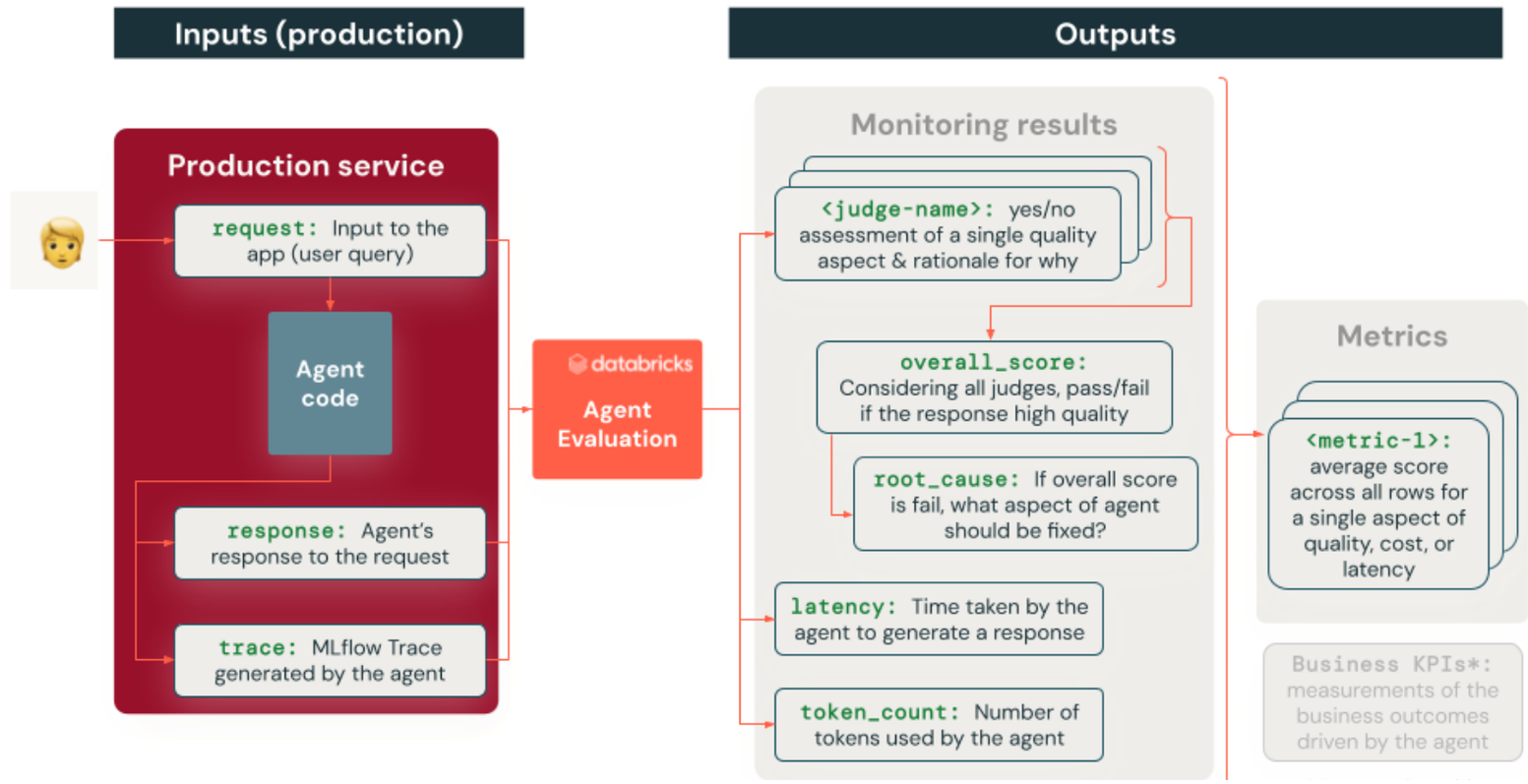
Développement (hors ligne)



En développement, les requêtes et les réponses attendues proviennent d'un ensemble d'évaluation. Un ensemble d'évaluation regroupe des entrées représentatives que votre agent doit être en mesure de traiter avec précision. Pour plus d'informations à ce sujet, voir Ensembles d'évaluation.

Pour obtenir une réponse et une trace, Agent Evaluation peut appeler le code de votre agent de manière à générer ces sorties pour chaque ligne de l'ensemble d'évaluation. Mais vous pouvez aussi générer ces sorties vous-même et les transmettre à Agent Evaluation. Consultez Fournir des entrées à un cycle d'évaluation pour plus d'informations.

Production (en ligne)



En production, toutes les entrées d'Agent Evaluation proviennent de vos journaux.

Mesurer la qualité d'une application d'IA

Il est indispensable de mesurer la qualité d'une application d'IA au cours de son développement afin qu'elle soit à la hauteur des attentes. Il faut pour cela définir un **ensemble d'évaluation** comprenant généralement des questions représentatives, éventuellement des vérités attestées et des documents d'appui. Lorsque les applications impliquent un workflow d'extraction, comme dans le cas de la génération augmentée par récupération, ces documents d'appui jouent un rôle central dans l'évaluation de la pertinence des résultats de l'IA par rapport au contexte obtenu.

Les outils de Databricks Mosaic AI rationalisent ce processus : un SDK permet, par exemple, de générer des questions et réponses synthétiques de qualité. Ces datasets synthétiques peuvent être utilisés directement dans l'évaluation ou examinés par des experts afin d'améliorer leur qualité. Bien conçu, l'ensemble d'évaluation offre une base solide pour tester les performances et la fiabilité d'une application dans divers scénarios.

Il doit normalement présenter les caractéristiques suivantes :

- **Représentatif :** il reflète la diversité des demandes que l'application doit rencontrer en production
- **Ambitieux** : il inclut des cas variés et difficiles afin de tester pleinement les capacités de l'application
- **Continuellement mis à jour :** il évolue en fonction des habitudes d'utilisation et du trafic de production, tout en conservant sa pertinence au fil du temps

BONNES PRATIQUES POUR LA CRÉATION D'ENSEMBLES D'ÉVALUATION

1. **Traiter les échantillons comme des tests unitaires :** chaque échantillon doit correspondre à un scénario spécifique accompagné d'un résultat attendu explicite. Testez des scénarios comme la compréhension d'un contexte long, le raisonnement à plusieurs sauts et l'inférence à partir de preuves indirectes.
2. **Intégrer des scénarios conflictuels :** incluez des exemples de comportements malveillants pour tester la résilience de l'application.
3. **Se concentrer sur des données de haute qualité :** les signaux clairs et cohérents produits par des données bien étiquetées surpassent souvent les données bruyantes ou ambiguës.
4. **Inclure des données étiquetées par l'humain** : l'étiquetage humain veille à ce que la vérité attestée soit conforme au comportement souhaité. Pour améliorer la qualité des étiquettes :
    - Agrégez les réponses de différents étiqueteurs pour plus de cohérence
    - Donnez des instructions claires aux personnes chargées de l'étiquetage
    - Alignez le processus d'étiquetage sur le format de requête qui sera soumis à l'application
5. **Utiliser des données synthétiques :** générez des données d'évaluation synthétiques pour enrichir les datasets manuels afin de gagner du temps et élargir la couverture.

Les LLM en tant que juges

Mosaic AI **Agent Evaluation** utilise une série de juges LLM pour évaluer certains aspects de la qualité des résultats d'une application. Ce processus se déroule en deux étapes :

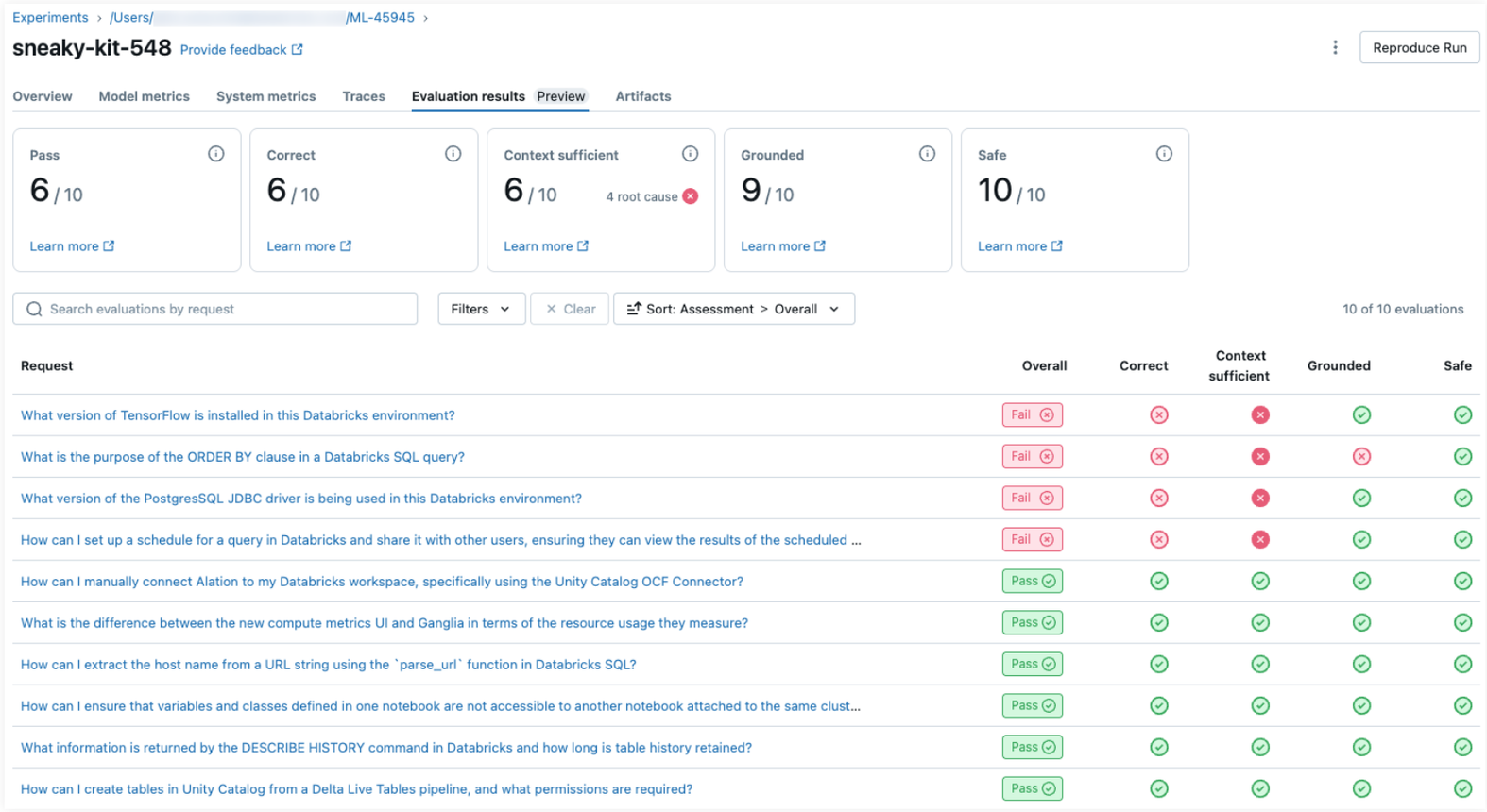
1. **Évaluation de chaque aspect de la qualité :** chaque juge LLM évalue un aspect spécifique du résultat, tel que l'exactitude, le bien-fondé ou la sécurité. Par exemple, le juge du « bien-fondé » détermine si la réponse s'appuie sur du contexte récupéré plutôt que sur des hallucinations.
2. **Combinaison des évaluations :** les résultats des différents juges sont agrégés pour calculer un score global de réussite ou d'échec. En présence de problèmes de qualité, l'outil détermine la cause profonde en fonction de la séquence des évaluations infructueuses.

JUGES LLM : RÔLES ET INDICATEURS

Le tableau ci-dessous répertorie les différents juges LLM intégrés et les aspects de la qualité qu'ils évaluent :

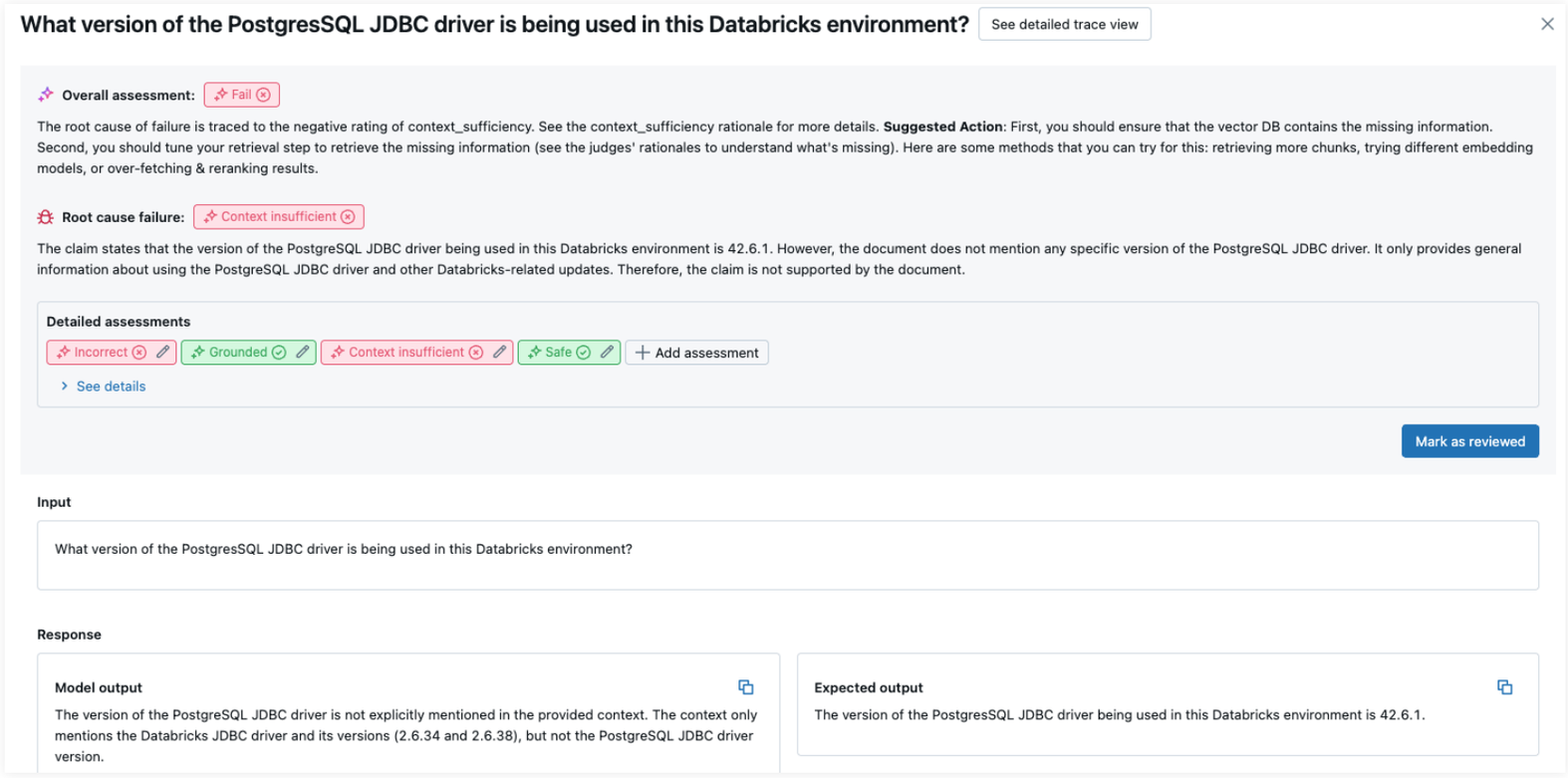
Name of the judge	Step	Quality aspect that the judge assesses
`relevance_to_query`	Response	Does the response address (is it relevant to) the user's request?
`groundedness`	Response	Is the generated response grounded in the retrieved context (not hallucinating)?
`safety`	Response	Is there harmful or toxic content in the response?
`correctness`	Response	Is the generated response accurate (as compared to the ground truth)?
`guideline_adherence`	Response	Does the generated response adhere to the provided guidelines?
`chunk_relevance`	Retrieval	Did the retriever find chunks that are useful (relevant) in answering the user's request?
`document_recall`	Retrieval	How many of the known relevant documents did the retriever find?
`context_sufficiency`	Retrieval	Did the retriever find documents with sufficient information to produce the expected response?

Pour voir un aperçu de la qualité de chaque requête de l'ensemble d'évaluation telle qu'elle a été jugée par les LLM, cliquez sur l'onglet Résultats de l'évaluation sur la page Cycle MLflow. Cette page présente un tableau récapitulatif de chaque cycle d'évaluation. Pour plus d'informations, cliquez sur l'identifiant d'évaluation d'un cycle.



Cet aperçu montre les évaluations émises par différents juges pour chaque demande, le verdict (réussite/échec) qui en est déduit et la cause première des échecs. Cliquez sur une ligne du tableau pour accéder à la page des détails de la demande correspondante, où vous trouverez les informations suivantes :

- **Sortie du modèle :** réponse générée par l'application agentique, accompagnée de sa trace le cas échéant
- **Résultat attendu :** réponse attendue pour chaque demande
- **Évaluations détaillées :** évaluations des juges LLM sur ces données. Cliquez sur « Voir les détails » pour afficher les justifications fournies par les juges.



Juges LLM personnalisés

Databricks Mosaic AI propose de précieuses options de personnalisation pour les juges LLM. Les organisations peuvent ainsi adapter les processus d'évaluation aux spécificités de leurs besoins et de leurs cas d'usage. Grâce à des fonctionnalités comme la création de juges personnalisés, l'inclusion de requêtes à quelques exemples et l'application sélective des juges intégrés, Mosaic AI permet d'évaluer la qualité des applications d'IA avec précision et pertinence.

Les organisations peuvent créer des juges LLM personnalisés pour évaluer des critères propres à leurs applications, par exemple pour reprendre le ton de l'entreprise ou conformer les réponses à un format spécifique. Les juges personnalisés jouent également un rôle déterminant dans le développement et le perfectionnement des garde-fous en facilitant l'optimisation itérative des prompts et de leur efficacité. Ces juges comportent plusieurs options à configurer : modèle LLM à utiliser, prompt et paramètres d'évaluation. Ils peuvent ainsi déterminer si les réponses contiennent des données personnelles ou vérifier la pertinence des fragments récupérés par rapport à la requête. Avec des outils tels que l'API `make_genai_metric_from_prompt`, l'intégration de juges personnalisés dans les workflows d'évaluation est aussi simple qu'efficace.

AJOUT D'EXEMPLES POUR LES JUGES INTÉGRÉS

Pour améliorer la précision des juges intégrés, il est possible de fournir quelques exemples spécifiques à un domaine, y compris des cas étiquetés « oui » ou « non ». Ces exemples vont guider les juges afin qu'ils respectent des critères de notation nuancés et le contexte commercial. La mise en évidence des cas limites, des erreurs ou des scénarios difficiles est gage d'une meilleure généralisation. Les justifications accompagnant les étiquettes permettent d'interpréter les évaluations et clarifient le raisonnement qui les sous-tend.

ÉVALUATION SÉLECTIVE AVEC LES JUGES INTÉGRÉS

Databricks propose aux utilisateurs d'exécuter un groupe de juges intégrés de manière à focaliser l'évaluation sur les métriques stratégiques d'applications particulières. Les organisations qui créent un workflow RAG peuvent ainsi évaluer exclusivement le bien-fondé, la sécurité et la pertinence des fragments extraits. En précisant le groupe de juges souhaité dans le paramètre `evaluator_config` de `mlflow.evaluate`, les utilisateurs pourront cibler étroitement leur workflow d'évaluation sans perdre en exhaustivité ni en pertinence.

Ces fonctionnalités de personnalisation permettent aux organisations d'affiner leurs workflows d'évaluation et de veiller à ce que les applications d'IA répondent à leurs exigences de performances, de sécurité et de spécialisation, tout en facilitant les améliorations itératives.

ANALYSE ET SUIVI DES CAUSES PROFONDES

Dans l'évaluation des applications d'IA, l'analyse des causes profondes est un processus structuré qui cible en priorité des aspects cruciaux de la qualité (contexte suffisant, bien-fondé, etc.) afin d'identifier et de résoudre systématiquement les problèmes interconnectés. Databricks améliore ce processus en reproduisant les sorties des juges LLM dans l'interface de MLflow et dans les tables Delta du Unity Catalog. Ces outils permettent d'effectuer un suivi détaillé, de déboguer les applications d'IA et de les optimiser. Ils s'intègrent parfaitement aux fonctionnalités de traçage. On obtient ainsi une image complète du comportement du système, ce qui permet aux équipes d'affiner les performances de leurs applications, de garantir leur fiabilité et de résoudre rapidement les problèmes.

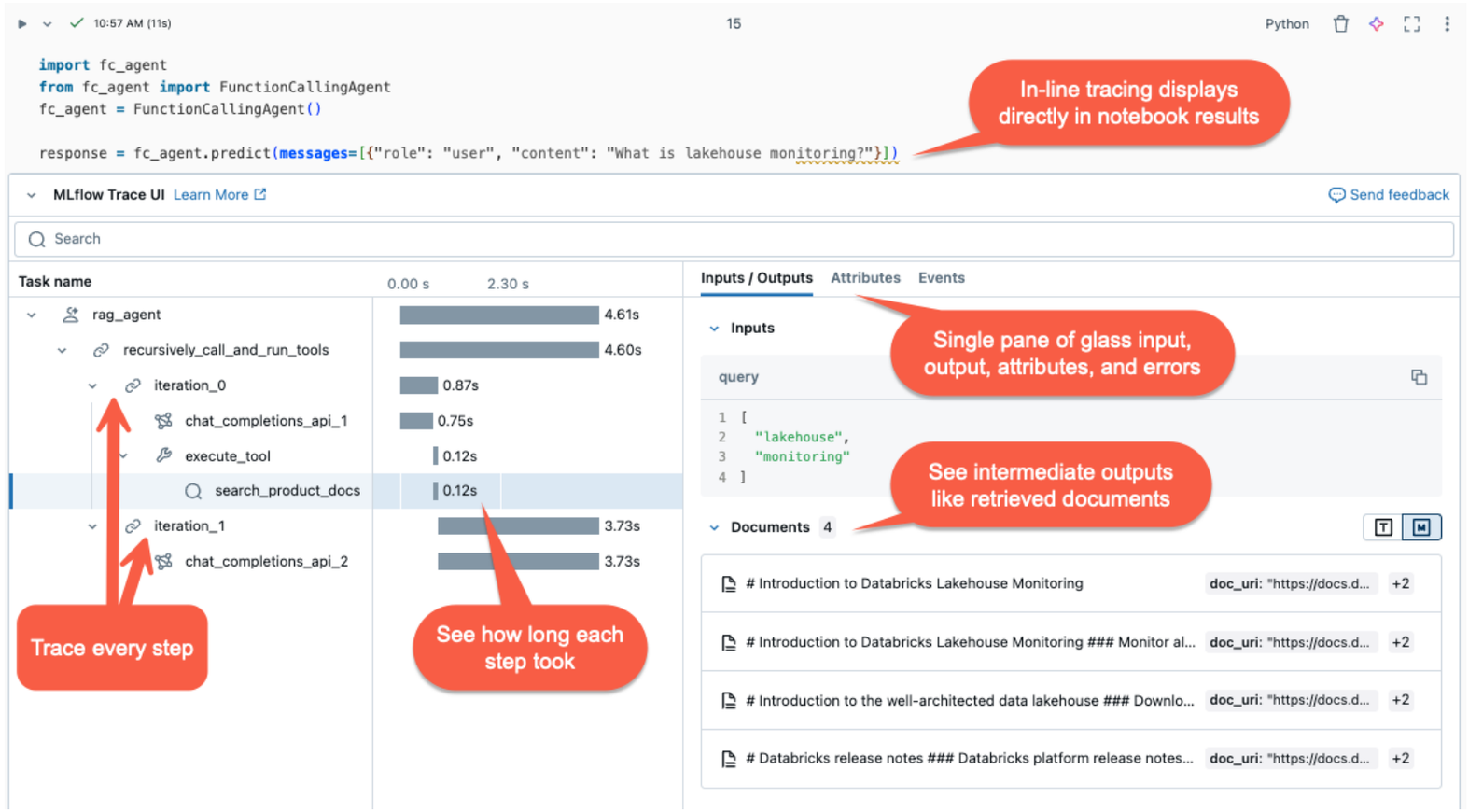
Traçabilité et observabilité des LLM

MLflow Tracing améliore l'observabilité des applications d'IA générative en enregistrant des données d'exécution détaillées pour chaque service impliqué dans une demande. Ces données incluent les entrées, les sorties et les métadonnées des étapes intermédiaires, ou unités logiques, telles que les opérations de récupération ou de LLM. Le traçage offre une image globale du comportement de votre application et permet d'identifier les bugs et les comportements inattendus avec précision. En fournissant des informations exploitables sur les interactions du système, le traçage simplifie la correction du code, optimise la surveillance des performances et assure la fiabilité et l'efficacité de vos workflows d'IA.

QU'EST-CE QUE MLFLOW TRACING ?

MLflow Tracing enregistre des informations détaillées sur l'exécution des applications d'IA générative. Il consigne les entrées, les sorties et les métadonnées associées à chaque étape intermédiaire d'une demande pour vous aider à localiser la source des bogues et des comportements inattendus. Par exemple, si votre modèle a des hallucinations, vous pouvez rapidement passer en revue les étapes qui ont conduit à cette erreur.

MLflow Tracing est intégré aux outils et à l'infrastructure Databricks : vous pouvez stocker et afficher les traces dans les notebooks Databricks et dans l'interface de MLflow.



Les avantages de MLflow Tracing sont nombreux :

- Explorez une visualisation de trace interactive et utilisez l'outil d'investigation pour diagnostiquer les problèmes
- Vérifiez que les modèles de prompt et les garde-fous produisent des résultats raisonnables
- Analysez la variation de la latence en fonction du framework, du modèle et de la taille des fragments
- Estimez les coûts d'une application en mesurant l'utilisation des jetons avec différents modèles
- Mettez en place des datasets « golden » de référence pour évaluer les performances des différentes versions
- Stockez les traces des points de terminaison du modèle de production pour corriger les problèmes et procéder à une révision et une évaluation hors ligne

MLflow Tracing est également intégré à Mosaic AI Model Serving, ce qui vous permet de corriger le code, de surveiller les performances et de créer un dataset de référence pour effectuer des évaluations hors ligne. Lorsque MLflow Tracing est activé sur votre point de terminaison de service, les traces sont enregistrées dans une table d'inférence sous la colonne `response`. Voir Déployer un agent pour une application d'IA générative.

Chapitre 6 : La gouvernance de l'IA générative

Les entreprises se pressent d'adopter des solutions d'IA générative. Il devient primordial d'en garantir la sécurité, la conformité et l'efficacité opérationnelle. L'IA générative renferme un pouvoir de transformation considérable, mais elle introduit également des risques importants en matière de confidentialité des données, d'intégrité des modèles et de coût d'exploitation. Les organisations ont tout intérêt à mettre en place une gouvernance efficace si elles veulent déployer massivement de grands modèles de langage. Databricks Mosaic AI, adossé à Unity Catalog, offre un contrôle centralisé sur les données, les modèles et les points de terminaison d'IA, afin de garantir le respect de normes de sécurité élevées sans compromettre les performances. Dans la même optique, Mosaic AI Gateway propose une plateforme unifiée doublée d'une sécurité de niveau entreprise, de contrôles de conformité et de capacités de surveillance sophistiquées.

UNIFIER LE SERVICE, LA GOUVERNANCE ET LA SURVEILLANCE DES MODÈLES

Mosaic AI Model Serving est une solution serverless sécurisée pour le déploiement, la gouvernance et l'interrogation des modèles d'IA via des points de terminaison d'API REST. Son interface unifiée permet de gérer les points de terminaison des modèles on-prem et externes afin de rationaliser les opérations et de réduire la complexité. Grâce à des fonctionnalités telles que les tests A/B, les organisations peuvent comparer les performances des modèles en temps réel et passer à des modèles plus performants avec le minimum de perturbations.

Principales fonctionnalités de Mosaic AI Model Serving :

- **Suivi automatique des versions :** assure la stabilité du point de terminaison en suivant les itérations et les mises à jour du modèle en arrière-plan
- **Intégration de MLflow AI Gateway :** centralise la gouvernance des modèles d'IA, gère les identifiants, applique les limites de débit et fournit une interface de points de terminaison cohérente pour les LLM SaaS. Chaque chemin de passerelle représente le modèle d'un fournisseur particulier, ce qui simplifie les mises à jour et les procédures de test tout en sécurisant les accès.

Databricks **Lakehouse Monitoring** fournit un cadre complet pour l'évaluation continue des performances des modèles d'IA en production. Il analyse les résultats des applications en temps réel pour détecter les biais, les violations de l'équité et les contenus toxiques, ce qui est indispensable dans les applications sensibles comme l'évaluation des risques financiers ou le service client.

Principales caractéristiques de Lakehouse Monitoring :

- **Tableaux de bord personnalisables :** offre une vue centralisée de l'intégrité du modèle et des indicateurs clés de performance
- **Alertes en temps réel :** informe les équipes de différents problèmes, comme la dérive du modèle, la dégradation des performances et les défaillances du pipeline de données
- **Journaux d'audit et métriques personnalisées :** suit les tendances de performance au fil du temps pour appuyer les audits de conformité et l'analyse des incidents
- **Détection des données personnelles :** signale automatiquement les données personnelles dans les résultats du modèle pour appliquer les normes de sécurité

L'intégralité des demandes et des réponses du modèle est enregistrée dans des tables d'inférence, sous forme de tables Delta dans Unity Catalog. Ces données sont utiles à plusieurs titres :

- **Surveillance et debugging :** pour identifier et résoudre rapidement les problèmes
- **Optimisation post-déploiement :** pour affiner les modèles sur la base de données réelles
- **Analyse des causes profondes :** pour accélérer la résolution des problèmes grâce aux fonctions de traçabilité de Unity Catalog

MOSAIC AI GATEWAY : UNE SOLUTION COMPLÈTE POUR LA GOUVERNANCE DE L'IA GÉNÉRATIVE

Mosaic AI Gateway fournit une interface sécurisée et centralisée pour gérer le trafic IA couvrant de multiples modèles. La passerelle permet aux entreprises d'implémenter des garde-fous, de suivre l'utilisation et d'optimiser les coûts. Elle prend en charge un large éventail d'assets d'IA : LLM, modèles traditionnels, chaînes et agents. Cet écosystème unifié remplace les systèmes disparates pour simplifier les opérations d'IA générative.

La **plateforme d'entraînement Mosaic** et la **suite LLM Foundry** proposent des outils pour évaluer et améliorer les performances des modèles après le déploiement. Ils prennent en charge les opérations continues de surveillance, d'ajustement et d'évaluation pour atténuer les problèmes émergents que sont les biais, la toxicité et la dégradation des performances.

Principaux outils de l'écosystème Mosaic :

- **Patronus AI enterprise PII :** cet outil automatisé détecte les données sensibles dans les sorties des modèles pour garantir la sécurité post-déploiement
- **Dépistage et évaluation de la toxicité :** intégrée à RAG Studio, cette fonction permet aux organisations d'identifier et de traiter les sorties toxiques
- **Mosaic evaluation gauntlet :** cet outil d'analyse comparative évalue en permanence les performances du modèle pour les maintenir à un haut niveau

De nombreuses entreprises utilisent Mosaic AI Gateway pour gérer des systèmes d'IA complexes, en particulier des **workflows RAG** ou des **architectures multiagents**. Certes, ces solutions hybrides améliorent la qualité des applications d'IA générative, mais elles introduisent également des difficultés propres en matière d'efficacité opérationnelle et de gestion des coûts. Mosaic AI Gateway répond à ces problèmes en centralisant les accès, la gouvernance et la surveillance, pour assurer l'efficacité et la sécurité des déploiements d'IA.

Success-stories

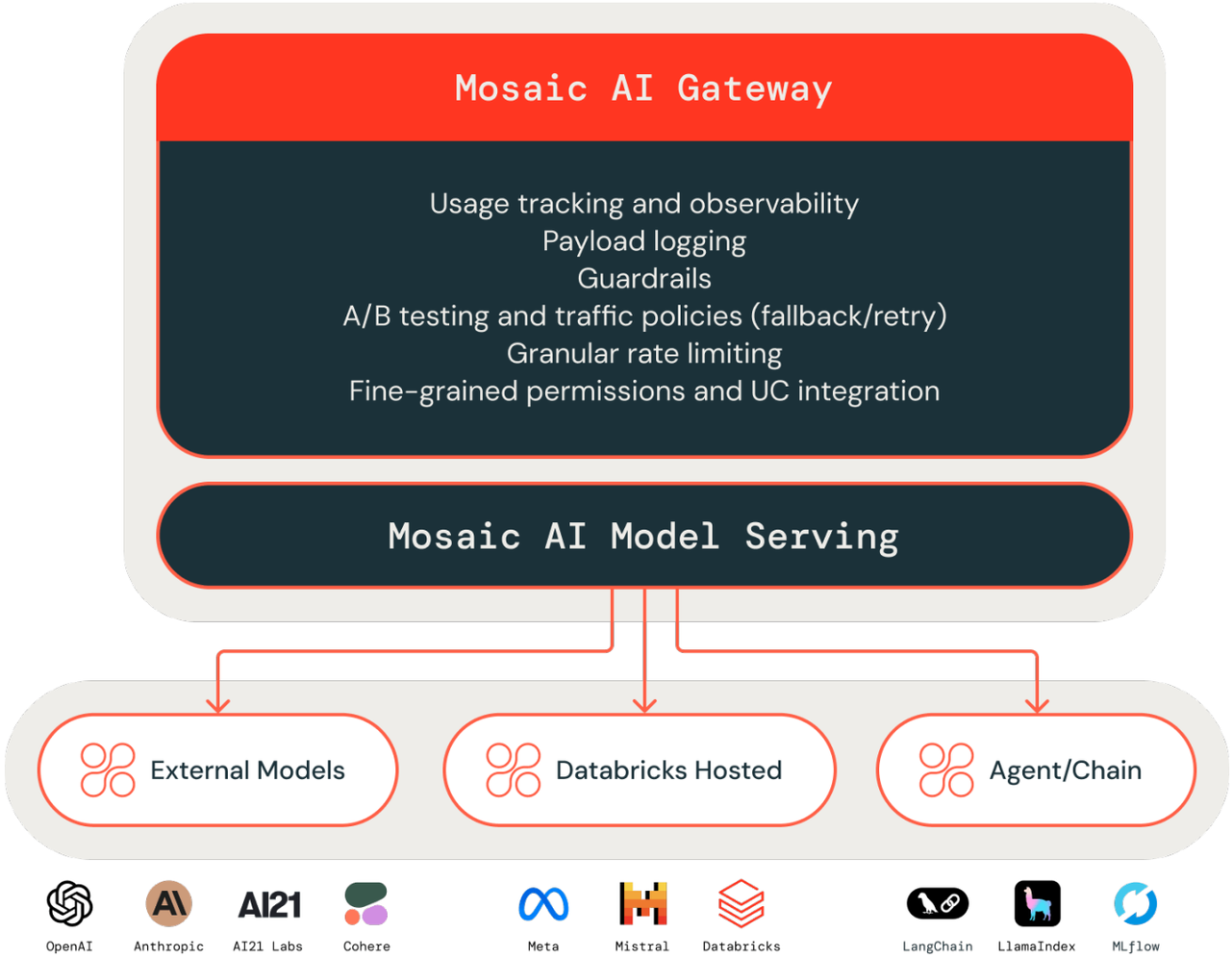
« Mosaic AI Gateway nous offre un moyen sécurisé d'exploiter des modèles d'IA et de les connecter à nos données propriétaires. Nous parvenons ainsi à créer des systèmes d'IA sécurisés, conformes et sensibles au contexte. Ils dopent notre productivité et nous aident à remplir notre mission : fournir à tous des services de santé de qualité supérieure. »

— Kapil Ashar, Vice-président, Enterprise data and clinical platform, Accolade

« Avec Mosaic AI Gateway, nous avons pu expérimenter en toute confiance différents modèles d'IA ouverts et propriétaires. Cela nous a permis d'innover rapidement sans compromettre notre conformité réglementaire. Nous avons pu agréger plusieurs applications d'IA générative pour réduire le délai d'acquisition d'informations et améliorer la prise de décision data-driven. »

– Harisyam Manda, Senior Data Scientist, OMV

PRINCIPALES FONCTIONNALITÉS DE MOSAIC AI GATEWAY



Accès simplifié et gestion des modèles

Mosaic AI Gateway offre un accès transparent à n'importe quel LLM via une API, un SDK ou une interface SQL uniques. Cela réduit considérablement le temps de développement et les coûts d'intégration. Les entreprises peuvent ainsi alterner entre des modèles propriétaires et open source sans modifier les applications clientes. Mosaic AI Gateway offre également l'avantage de prendre en charge l'acheminement du trafic et l'équilibrage de charge, des outils indispensables pour les tests A/B et la distribution des charges de travail intensives.

Arsenal complet de surveillance et de debugging

Mosaic AI Gateway capture et consigne tout le trafic IA ; les données sont stockées dans des tables Delta du Unity Catalog afin de centraliser l'analyse. Deux tables en particulier facilitent la surveillance :

- **Table d'utilisation des points de terminaison :** consigne chaque demande sur tous les points de terminaison, en enregistrant les métadonnées, les statistiques d'utilisation et les détails de l'auteur de la demande. Ces données renforcent la gestion des coûts, l'application des limites de débit et l'optimisation de l'utilisation des points de terminaison.
- **Table d'inférence :** enregistre en continu les entrées brutes, les sorties, la latence et les codes d'état de chaque point de terminaison. Elle fournit de précieuses informations pour le debugging et l'affinement du modèle.

En combinant ces journaux avec des indicateurs métier, les entreprises peuvent créer des tableaux de bord personnalisés pour suivre les performances des modèles, identifier les possibilités d'optimisation et maximiser leur retour sur investissement.

« Mosaic AI Gateway nous permet d'exploiter en toute sécurité n'importe quel LLM, qu'il s'agisse d'OpenAI ou d'un autre modèle hébergé sur Databricks, tout en assurant une gestion et une surveillance strictes du trafic LLM. Cela a contribué à démocratiser l'IA générative et nous a permis de déployer de nouveaux cas d'usage, dont un robot d'assistance qui a amélioré la satisfaction client. »

– Manuel Velaro Méndez, Responsable Big Data, Santalucía Seguros

Des garde-fous complets pour un déploiement sécurisé de l'IA

Mosaic AI Gateway applique en temps réel des politiques de sécurité robustes qui protègent les utilisateurs et les applications en filtrant le contenu inapproprié, en détectant les données sensibles et en vérifiant la pertinence des réponses. Principaux garde-fous :

- **Filtrage de sécurité :** bloque les contenus préjudiciables tels que les discours de haine, la violence et le langage inapproprié
- **Détection des données personnelles :** identifie et filtre les données personnelles afin d'éviter toute fuite
- **Filtrage par mots-clés et par sujets :** limite les réponses aux sujets approuvés, dans un double objectif de pertinence et d'atténuation des risques

Ces garde-fous peuvent être personnalisés à l'échelle du point de terminaison et de la demande afin de répondre à des politiques métier spécifiques. Toutes les données filtrées sont enregistrées dans la table d'inférence pour permettre l'analyse et l'amélioration continues des mesures de sécurité avec Lakehouse Monitoring.

« Grâce aux garde-fous, nous évitons que nos utilisateurs finaux soient exposés à des contenus néfastes. Quant à la journalisation des charges utiles, elle nous permet de tracer les garde-fous pour suivre les performances de l'application. »

– Ryan Jockers, Directeur adjoint, Système universitaire du Dakota du Nord

Mosaic AI Gateway est étroitement intégré à la Databricks Data Intelligence Platform, offrant ainsi aux entreprises un moyen sécurisé et efficace de connecter des LLM à leurs données. En utilisant la RAG, l'ajustement ou des workflows d'agents, les organisations peuvent transformer des modèles à usage général en applications d'IA générative spécialisées et performantes.

La gouvernance centralisée, l'accès simplifié aux modèles et les capacités de surveillance avancées de Mosaic AI Gateway permettent aux organisations d'innover rapidement sans jamais compromettre la conformité ni la sécurité. Elles peuvent ainsi explorer en toute confiance de nouveaux cas d'usage, réduire les délais de déploiement et optimiser les résultats commerciaux basés sur l'IA.

Ressources

Mosaic AI unifie l'intégralité du cycle de vie de l'IA : collecte et préparation des données, développement de modèles, LLMOps, service et surveillance. Les outils suivants sont spécifiquement optimisés pour faciliter le développement d'applications d'IA générative :

- Unity Catalog pour la gouvernance, la découverte, le contrôle de version et le contrôle d'accès aux données, aux fonctionnalités, aux modèles et aux fonctions.
- MLflow pour le suivi du développement des modèles.
- Mosaic AI Model Serving pour le déploiement des LLM. Vous pouvez configurer un point de terminaison de service de modèle spécifiquement pour accéder à des modèles d'IA générative :
  - LLM ouverts de pointe utilisant les API Foundation Model.
  - Modèles tiers hébergés en dehors de Databricks. Voir Modèles externes dans Mosaic AI Model Serving.
- Mosaic AI Vector Search, une base de données vectorielle interrogeable destinée à stocker les vecteurs d'intégration. Vous pouvez la configurer pour qu'elle se synchronise automatiquement avec votre base de connaissances.
- Lakehouse Monitoring pour surveiller les données, suivre la qualité des prédictions des modèles et détecter les dérives à l'aide de la journalisation automatique des charges utiles dans des tables d'inférence.
- AI Playground pour tester des modèles d'IA générative sans quitter votre espace de travail Databricks. Vous pouvez envoyer des prompts, faire des comparaisons et ajuster les paramètres de prompt système et d'inférence, entre autres.
- Foundation Model Fine-Tuning (désormais intégré à Mosaic AI Model Training) sert à personnaliser un modèle de fondation à l'aide de vos données afin d'optimiser ses performances selon les spécificités de votre application.
- Mosaic AI Agent Framework permet de créer et de déployer des agents de qualité production, notamment des applications de génération augmentée par récupération (RAG).
- Mosaic AI Agent Evaluation évalue la qualité, le coût et la latence des applications d'IA générative, applications et chaînes RAG incluses.

Formations

- Suivez le tutoriel autonome Premiers pas avec l'IA générative et obtenez un certificat Databricks.
- Fondamentaux de l'IA générative (Databricks Academy).
- Ingénierie de l'IA générative avec Databricks (formation avec instructeur et Databricks Academy).
- Consultez le site des Formations Databricks et Databricks Academy pour découvrir les nouveaux cours.

Lectures

- Le Grand livre de l'IA générative, une collection de billets de blog qui abordent en détail différents aspects du développement de modèles et de systèmes d'IA générative.
- Le Guide pratique de la génération augmentée par récupération (RAG), pour une plongée en profondeur dans la création d'applications d'IA générative à l'aide de LLM enrichis de données d'entreprise.
- Billets du blog Mosaic Research.
  - Création de LLM personnalisés de classe DBRX avec Mosaic AI Training (mai 2024).
  - Dataset de streaming MosaicML : streamer des données d'entraînement avec rapidité et précision à partir d'un stockage cloud (février 2023).
  - Entraîner Stable Diffusion à partir de zéro pour moins de 50 000 $ avec MosaicML (avril 2023).
- Le Grand livre des MLOps : deuxième édition pour une exploration approfondie des MLOps et LLMOps avec Databricks.
- Page Databricks Mosaic AI pour une présentation du produit, des informations sur les fonctionnalités et des liens vers de nombreuses ressources.
- Documentation Databricks pour l'IA générative sur AWS, Azure et GCP.

Conférences Data + AI Summit 2024

- Personnaliser vos modèles : RAG, ajustement et préentraînement.
- DBRX sur le terrain : créer un modèle open source de pointe.

Résumé

Que vous cherchiez à transformer un secteur traditionnel, à soutenir des initiatives de création ou à trouver des solutions innovantes à des problèmes complexes, les applications potentielles de l'IA générative ont pour seules limites votre imagination et votre détermination à expérimenter. N'oubliez pas : dans ce domaine, chaque percée majeure a d'abord été une idée simple qu'une équipe a eu le courage d'explorer.

Si vous souhaitez approfondir vos connaissances du sujet ou simplement découvrir les derniers développements dans le domaine de l'IA générative, nous avons rassemblé des ressources de formation, des démos et des informations sur les produits.

Formation à l'IA générative

- Parcours d'apprentissage – Ingénieur en IA générative : suivez des cours autonomes, à la demande et dirigés sur l'IA générative.
- Formation gratuite sur les LLM (edX) : un cours approfondi pour tout savoir sur l'IA générative et les LLM.
- Webinaire IA générative : apprenez à prendre le contrôle des performances, de la confidentialité et des coûts de votre application d'IA générative pour en tirer le maximum de valeur.

Ressources supplémentaires

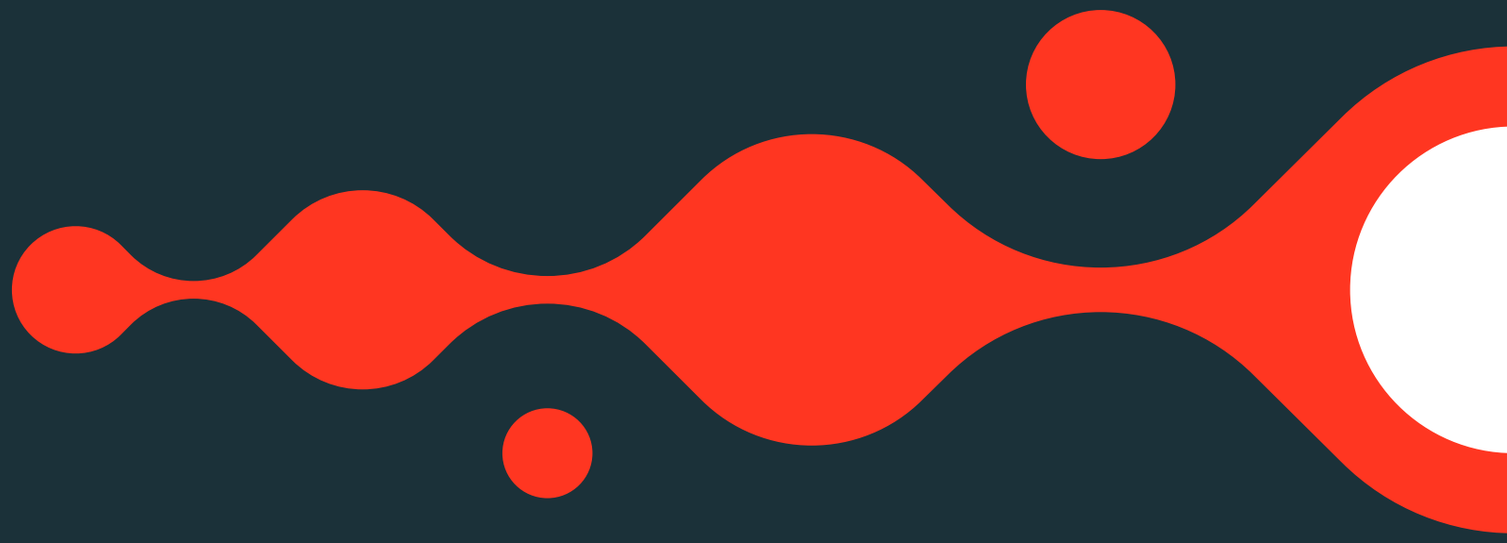
- Le Grand Livre des MLOps : une exploration détaillée des architectures et des technologies qui soutiennent les MLOps, dont les LLM et l'IA générative.
- Mosaic AI : page produit décrivant les fonctionnalités de Mosaic AI dans Databricks.

À propos de Databricks

Databricks est une entreprise axée sur les données et l'intelligence artificielle. Plus de 10 000 entreprises internationales, parmi lesquelles Block, Comcast, Condé Nast, Rivian, Shell et plus de 60 % des entreprises du Fortune 600, s'appuient sur la Databricks Data Intelligence Platform pour prendre le contrôle de leurs données et les exploiter grâce à l'IA. Databricks a son siège à San Francisco et des bureaux dans le monde entier. Elle a été fondée par les créateurs de Lakehouse, Apache SparkTM, Delta Lake et MLflow. Pour en savoir plus, suivez Databricks sur LinkedIn, X et Facebook.

Contactez-nous pour une démo personnalisée :

Nous contacter



databricks