Création de solutions de génération augmentée de récupération AWS pour les soins de santé

AWS Conseils prescriptifs

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Conseils prescriptifs: Création de solutions de génération augmentée de récupération AWS pour les soins de santé

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Les marques et la présentation commerciale d’Amazon ne peuvent être utilisées en relation avec un produit ou un service qui n’est pas d’Amazon, d’une manière susceptible de créer une confusion parmi les clients, ou d’une manière qui dénigre ou discrédite Amazon. Toutes les autres marques commerciales qui ne sont pas la propriété d’Amazon appartiennent à leurs propriétaires respectifs, qui peuvent ou non être affiliés ou connectés à Amazon, ou sponsorisés par Amazon.

Table of Contents

Introduction 1
Soins aux patients et productivité 2
Gestion des talents 2
Opportunités et défis 4
Opportunités d'applications d'IA générative dans le secteur de la santé 4
Analyse d'image avancée 5
Défis liés à l'industrialisation des solutions 5
Cas d'utilisation : création d'une application de renseignement médical 6
Présentation de la solution 6
Étape 1 : Découverte des données 8
Étape 2 : Création d'un graphe des connaissances médicales 9
Étape 3 : Création d'agents de récupération du contexte 15
Agents Amazon Bedrock 15
Agents LangChain 17
Étape 4 : Création d'une base de connaissances 18
Utilisation du OpenSearch service 19
Création d'une architecture RAG 20
Étape 5 : Génération de réponses 23
Alignement sur le AWS Well-Architected Framework 25
Cas d'utilisation : prévision des taux de réadmission 26
Présentation de la solution 26
Étape 1 : Prédire les résultats pour les patients 29
Étape 2 : Prédire le comportement du patient 31
Étape 3 : Prédire la réadmission du patient 33
Étape 4 : Calcul du score de propension 36
Alignement sur le AWS Well-Architected Framework 39
Cas d'utilisation : gestion des talents 40
Présentation de la solution 41
Étape 1 : Création d'un profil de compétences 43
Étape 2 : Déterminer la pertinence du rôle par rapport aux compétences 44
Étape 3 : recommander une formation 45
Alignement sur le AWS Well-Architected Framework 46
Développement de solutions 48
Amazon Q Developer 48

Multi-retriever Design RAG 49
ReAct agents 51
Évaluation des solutions 53
Évaluation de l'extraction d'informations 53
Évaluation de plusieurs récupérateurs 54
Utiliser un LLM 54
Ressources 56
AWS documentation 56
AWS articles de blog 56
Autres ressources 56
Collaborateurs 57
Conception 57
Révision 57
Rédaction technique 57
Historique du document 58
..... lix

Création de solutions de génération augmentée de récupération AWS pour les soins de santé

Amazon Web ServicesAccenture, et Cadiem (contributeurs)

Mars 2025 (historique du document)

Avant les grands modèles de langage (LLMs) et l'IA générative, le développement d'applications automatisées et de haute précision dans le secteur de la santé était un défi. Les méthodes traditionnelles reposaient largement sur la saisie et l'analyse manuelles des données. La complexité de l'analyse de l'imagerie médicale et des dossiers des patients a nécessité une intervention humaine importante, ce qui a souvent entraîné des flux de travail fragmentés et inefficaces. Les progrès des technologies d'intelligence artificielle vous aident à créer des applications hyperpersonnalisées à grande échelle. Les applications de santé peuvent désormais s'intégrer aux bases de connaissances médicales, interpréter les images diagnostiques avec une précision accrue et prévoir les résultats des patients à l'aide de modèles prédictifs.

Ce guide explique comment LLMs révolutionnent les soins de santé grâce à des applications de génération augmentée de récupération avec lesquelles vous pouvez créer. Services AWSLa génération augmentée de récupération (RAG) est une technologie d'IA générative dans laquelle un LLM fait référence à une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation avant de générer une réponse. Les applications RAG fondent les résultats du modèle sur des connaissances du monde réel, ce qui réduit les hallucinations et augmente la pertinence des réponses. Dans le secteur de la santé, le RAG peut être utilisé pour fournir des informations up-to-date médicales précises, garantissant ainsi que les prestataires de soins de santé ont accès aux dernières recherches et directives cliniques. En transformant les données en informations exploitables et en automatisant des processus complexes, ces technologies contribuent à améliorer les soins aux patients, à rationaliser les opérations et à améliorer la productivité des professionnels de santé.

Dans Amazon Bedrock, vous pouvez les affiner LLMs et les intégrer à des agents intelligents pour créer des solutions de santé avancées. Soulignant la synergie entre Amazon OpenSearch Service et Amazon Neptune, le guide montre comment ces services améliorent les solutions RAG grâce à une pertinence de recherche améliorée et à une extraction avancée de données multi-sources. Vous pouvez orchestrer des solutions Amazon Bedrock complètes qui utilisent les agents Amazon Bedrock et coordonnent de manière fluide LangChainles interactions entre différents référentiels de données.

Cette intégration démontre le pouvoir de combiner des services spécialisés pour créer des systèmes pilotés par l'IA plus efficaces et efficients.

Soins aux patients et productivité

Ce guide présente deux cas d'utilisation concrets pour les soins aux patients et la productivité : l'augmentation des données sur les patients et la prévision des risques de réadmission. Il fournit des plans stratégiques pour la mise en œuvre de ces solutions à grande échelle, offrant aux établissements de santé une voie claire vers l'industrialisation des processus pilotés par l'IA. Grâce à ces informations, les établissements de santé peuvent utiliser des technologies d'intelligence artificielle avancées pour créer des flux de travail plus efficaces et plus intelligents.

Gestion des talents

Ce guide décrit également des stratégies visant à requalifier les professionnels de santé et à leur donner les moyens d'intégrer facilement l'IA générative dans leur routine quotidienne. Cela peut améliorer à la fois la productivité et la qualité des soins aux patients. En dotant leur personnel des compétences nécessaires pour utiliser efficacement les outils avancés d'intelligence artificielle, les établissements de santé peuvent optimiser leur retour sur investissement et stimuler l'innovation dans les soins aux patients.

Cette solution de gestion des talents basée sur l'IA inclut les fonctionnalités clés suivantes :

- Analyseur intelligent de CV de talents : en utilisant les fonctionnalités avancées LLMs disponibles dans Amazon Bedrock, cet outil extrait et analyse efficacement les compétences et attributs essentiels des CV en matière de talents. Cet outil permet de rationaliser le processus de recrutement.
- Base de connaissances sur les talents : alimentée par Amazon Neptune, cette base de données dynamique fournit des informations en temps réel sur les niveaux de personnel, la répartition des compétences et les tendances du secteur. Cela vous aide à prendre des décisions basées sur les données concernant la gestion des effectifs.
- Moteur de recommandation d'apprentissage — Cet outil piloté par l'IA identifie les lacunes en matière de compétences au sein de l'organisation et recommande des programmes de formation personnalisés pour le personnel médical. Cet outil favorise le développement professionnel continu et aide votre personnel à s'adapter à l'évolution des technologies de santé.

Ensemble, ces fonctionnalités basées sur l'IA aident à optimiser les performances du personnel, révolutionnant ainsi la gestion des talents grâce à une intelligence et une efficacité accrues.

Opportunités et défis

Amazon Bedrock peut améliorer la productivité, l'évolutivité, la rentabilité et fournir des informations basées sur les données. Amazon Bedrock permet aux établissements de santé de les utiliser LLMs efficacement dans différents cas d'utilisation, qu'il s'agisse de création de contenu, d'analyse de données ou de prise de décision automatisée. Ce guide propose des approches pour surmonter les défis courants liés à l'IA générative, tels que les problèmes de qualité des données, l'évolutivité de l'infrastructure, le maintien des performances des modèles et les exigences d'amélioration continue pendant la transition de la preuve de concept à la production.

Opportunités d'applications d'IA générative dans le secteur de la santé

Le secteur de la santé est sur le point de connaître un changement transformateur, stimulé par les opportunités offertes par les applications d'IA générative. L'IA générative a le potentiel d'améliorer les soins aux patients, de rationaliser les opérations et d'accélérer la recherche médicale. En utilisant des modèles d'IA avancés, les prestataires de soins de santé peuvent automatiser l'augmentation des dossiers médicaux. Les dossiers complets et les antécédents des up-to-date patients facilitent des diagnostics et des plans de traitement plus précis. L'analyse d'images pilotée par l'IA, telle que l'interprétation des sonogrammes et d'autres images médicales, peut fournir des informations rapides et précises, réduisant ainsi la charge de travail des professionnels de santé et minimisant le risque d'erreur humaine.

Au-delà du diagnostic et du traitement, l'IA générative peut jouer un rôle essentiel dans l'analyse prédictive. L'analyse prédictive aide les établissements de santé à anticiper les résultats pour les patients et à personnaliser les plans de soins en conséquence. Cette technologie peut également optimiser les processus administratifs, qu'il s'agisse de gérer les données des patients ou de rationaliser la communication entre les prestataires et les patients. En intégrant des solutions d'IA générative aux systèmes de santé existants, les établissements médicaux peuvent gagner en efficacité, réduire les coûts et, en fin de compte, fournir des soins de meilleure qualité. L'intégration de l'IA aux soins de santé n'est pas simplement une amélioration, mais une évolution fondamentale vers des soins plus intelligents, réactifs et centrés sur le patient.

Analyse d'image avancée

La combinaison d'Amazon Bedrock avec des magasins de données, tels qu'Amazon Neptune et OpenSearch Amazon Service, peut vous aider à surmonter les difficultés liées à l'analyse avancée d'images dans le secteur de la santé. Les solutions de récupération d'informations peuvent améliorer le processus de découverte des maladies et améliorer la précision de l'interprétation en évaluant les images diagnostiques et en interprétant les échographies. La solution peut intégrer les données d'évaluation visuelles et textuelles à l'examen manuel de l'évaluation des patients par les médecins.

Défis liés à l'industrialisation des solutions

Les principaux obstacles à surmonter lors de l'industrialisation des solutions d'IA dans le secteur de la santé sont la qualité et la disponibilité des données. Les données de santé existent souvent dans des formats fragmentés et incohérents. Il est essentiel de s'assurer que les modèles d'IA ont accès à des données propres, structurées et représentatives pour maintenir les performances dans des scénarios réels. L'évolutivité de l'infrastructure peut devenir un défi en raison des environnements de production. Ces environnements doivent gérer d'importants volumes de données sur les patients en temps réel tout en offrant des temps de réponse rapides et en respectant les réglementations relatives à la confidentialité des données, telles que la loi HIPAA (Health Insurance Portability and Accountability Act). De plus, compte tenu des nouvelles informations médicales et des données sur les patients qui évoluent au fil du temps, les modèles d'IA doivent être réentraînés et mis à jour pour rester pertinents et donner des recommandations précises. Enfin, l'intégration de ces solutions d'IA dans les systèmes de santé existants peut s'avérer complexe en raison de problèmes d'interopérabilité et de la nécessité de les aligner sur les flux de travail cliniques actuels. Cette intégration nécessite des modifications techniques et opérationnelles.

Cas d'utilisation : création d'une application d'intelligence médicale à partir de données augmentées sur les patients

L'IA générative peut contribuer à améliorer les soins aux patients et la productivité du personnel en améliorant les fonctions cliniques et administratives. AI-driven l'analyse d'images, telle que l'interprétation des échographies, accélère les processus de diagnostic et améliore la précision. Il peut fournir des informations essentielles qui soutiennent des interventions médicales en temps opportun.

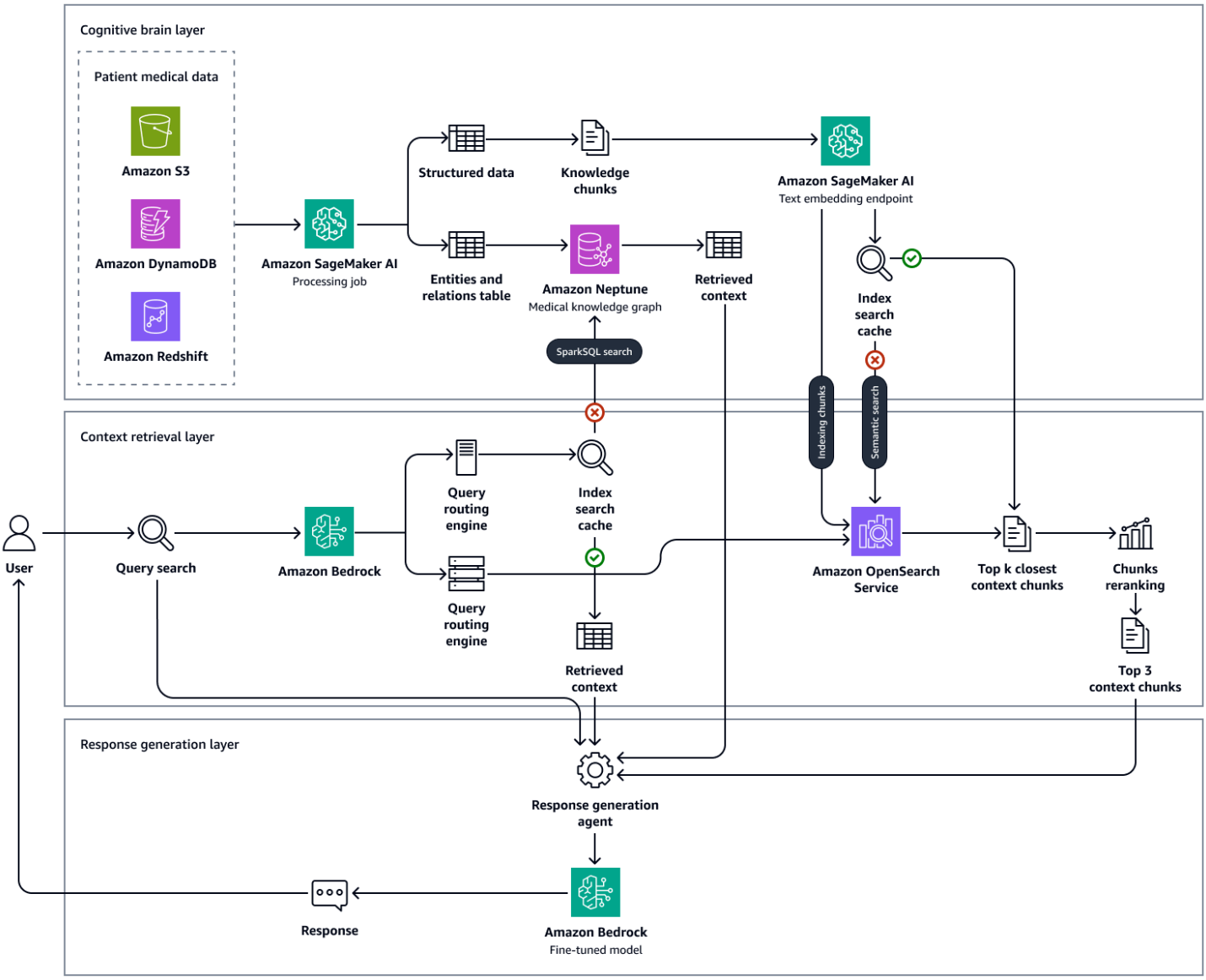
Lorsque vous combinez des modèles d'IA génératifs avec des graphes de connaissances, vous pouvez automatiser l'organisation chronologique des dossiers médicaux électroniques. Cela vous permet d'intégrer les données en temps réel issues des interactions médecin-patient, des symptômes, des diagnostics, des résultats de laboratoire et de l'analyse d'images. Cela fournit au médecin des données complètes sur le patient. Ces données aident le médecin à prendre des décisions médicales plus précises et plus rapides, améliorant à la fois les résultats pour les patients et la productivité des prestataires de soins de santé.

Présentation de la solution

L'IA peut renforcer les capacités des médecins et des cliniciens en synthétisant les données des patients et les connaissances médicales afin de fournir des informations précieuses. Cette solution RAG (Retrieval Augmented Generation) est un moteur d'intelligence médicale qui utilise un ensemble complet de données et de connaissances sur les patients issues de millions d'interactions cliniques. Il exploite le pouvoir de l'IA générative pour créer des informations fondées sur des preuves afin d'améliorer les soins aux patients. Il est conçu pour améliorer les flux de travail cliniques, réduire les erreurs et améliorer les résultats pour les patients.

La solution inclut une capacité de traitement d'image automatique alimentée par des LLM. Cette fonctionnalité réduit le temps que le personnel médical doit consacrer à la recherche manuelle d'images diagnostiques similaires et à l'analyse des résultats de diagnostic.

L'image suivante montre le flux de travail de bout en bout de cette solution. Il utilise Amazon Neptune, Amazon SageMaker AI, Amazon OpenSearch Service et un modèle de base dans Amazon Bedrock. Pour l'agent de récupération du contexte qui interagit avec le graphe des connaissances médicales dans Neptune, vous pouvez choisir entre un agent Amazon Bedrock et un agent. LangChain



Lors de nos expériences avec des exemples de questions médicales, nous avons observé que les réponses finales générées par notre approche utilisant le graphe de connaissances géré dans Neptune, une base de données OpenSearch vectorielle hébergeant la base de connaissances cliniques et les LLM d'Amazon Bedrock étaient fondées sur des faits et étaient bien plus précises en réduisant le nombre de faux positifs et en augmentant les vrais positifs. Cette solution peut générer des informations factuelles sur l'état de santé du patient et vise à améliorer les flux de travail cliniques, à réduire les erreurs et à améliorer les résultats pour les patients.

La création de cette solution comprend les étapes suivantes :

- Étape 1 : Découverte des données

- Étape 2 : Création d'un graphe des connaissances médicales
- Étape 3 : Création d'agents de récupération du contexte pour interroger le graphe des connaissances médicales
- Étape 4 : Création d'une base de connaissances contenant des données descriptives en temps réel
- Étape 5 : Utiliser les LLM pour répondre à des questions médicales

Étape 1 : Découverte des données

Il existe de nombreux ensembles de données médicales open source que vous pouvez utiliser pour soutenir le développement d'une AI-driven solution de santé. L'un de ces ensembles de données est le MIMIC-IV jeu de données, qui est un ensemble de données de dossiers de santé électroniques (DSE) accessible au public qui est largement utilisé dans le milieu de la recherche en santé. MIMIC-IV contient des informations cliniques détaillées, y compris des notes de sortie en texte libre extraites des dossiers des patients. Vous pouvez utiliser ces enregistrements pour expérimenter des techniques de sommation de texte et d'extraction d'entités. Ces techniques vous aident à extraire des informations médicales (telles que les symptômes du patient, les médicaments administrés et les traitements prescrits) à partir de texte non structuré.

Vous pouvez également utiliser un ensemble de données qui fournit des résumés de sortie de patients annotés et anonymisés, spécialement conçus à des fins de recherche. Un ensemble de données récapitulatives sur les sorties peut vous aider à expérimenter l'extraction d'entités, ce qui vous permet d'identifier les entités médicales clés (telles que les affections, les procédures et les médicaments) à partir du texte. Étape 2 : Création d'un graphe des connaissances médicalesdans ce guide, décrit comment vous pouvez utiliser les données structurées extraites des ensembles de données récapitulatifs MIMIC-IV et des ensembles de données récapitulatifs des sorties pour créer un graphique des connaissances médicales. Ce graphe de connaissances médicales sert de colonne vertébrale aux systèmes avancés d'interrogation et d'aide à la décision destinés aux professionnels de santé.

Outre les jeux de données textuels, vous pouvez utiliser des jeux de données d'images. Par exemple, le jeu de données sur les radiographies musculosquelettiques (MURA), qui est une base de données complète d'images radiographiques multivues des os. Utilisez ces ensembles de données d'images pour expérimenter l'évaluation diagnostique à l'aide de techniques de décodage d'images médicales. Ces techniques de décodage sont cruciales pour le diagnostic précoce de maladies telles que les maladies musculosquelettiques, les maladies cardiovasculaires et l'ostéoporose. En affinant

les modèles de base de la vision et du langage sur le jeu de données d'images médicales, vous pouvez détecter des anomalies dans les images diagnostiques. Cela permet au système de fournir aux cliniciens des informations diagnostiques précoces et précises. En utilisant des ensembles de données d'images et de textes, vous pouvez créer une application AI-driven médicale capable de traiter à la fois du texte et des données d'image pour améliorer les soins aux patients.

Étape 2 : Création d'un graphe des connaissances médicales

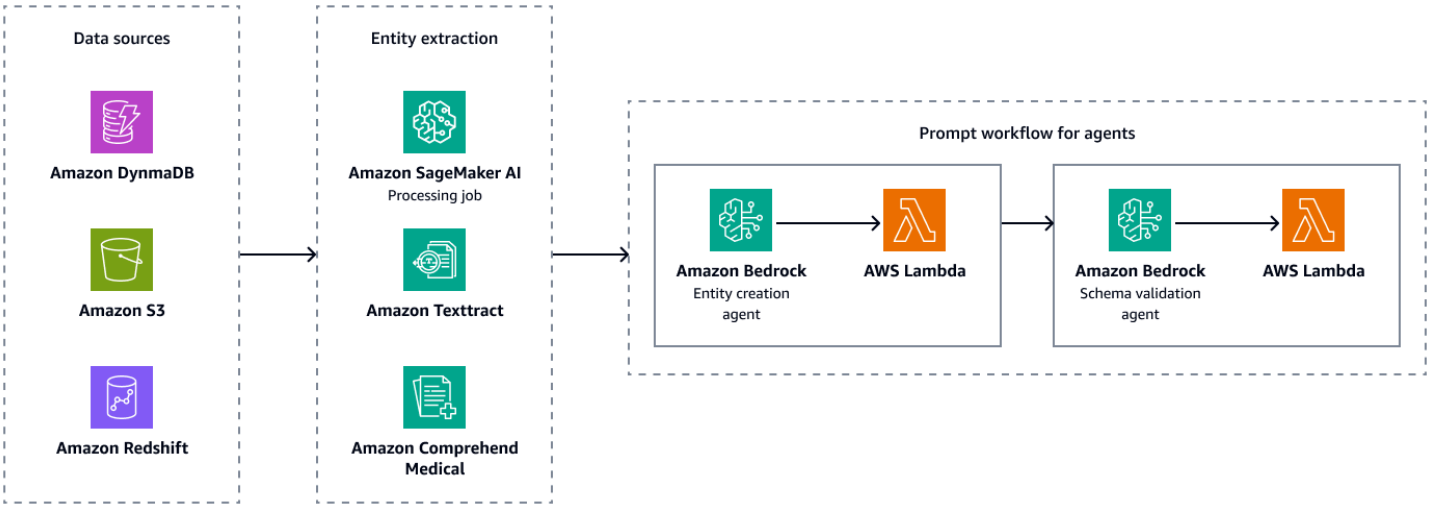
Pour tout établissement de santé qui souhaite créer un système d'aide à la décision basé sur une base de connaissances massive, l'un des principaux défis consiste à localiser et à extraire les entités médicales présentes dans les notes cliniques, les revues médicales, les résumés de sortie et les autres sources de données. Vous devez également capturer les relations temporelles, les sujets et les évaluations de certitude de ces dossiers médicaux afin d'utiliser efficacement les entités, les attributs et les relations extraits.

La première étape consiste à extraire des concepts médicaux d'un texte médical non structuré en utilisant quelques instructions pour un modèle de base, tel que Llama 3 dans Amazon Bedrock. Few-shot l'invite se produit lorsque vous fournissez à un LLM un petit nombre d'exemples illustrant la tâche et le résultat souhaité avant de lui demander d'effectuer une tâche similaire. À l'aide d'un extracteur d'entités LLM-based médicales, vous pouvez analyser le texte médical non structuré, puis générer une représentation des données structurées des entités de connaissances médicales. Vous pouvez également stocker les attributs du patient à des fins d'analyse et d'automatisation en aval. Le processus d'extraction des entités inclut les actions suivantes :

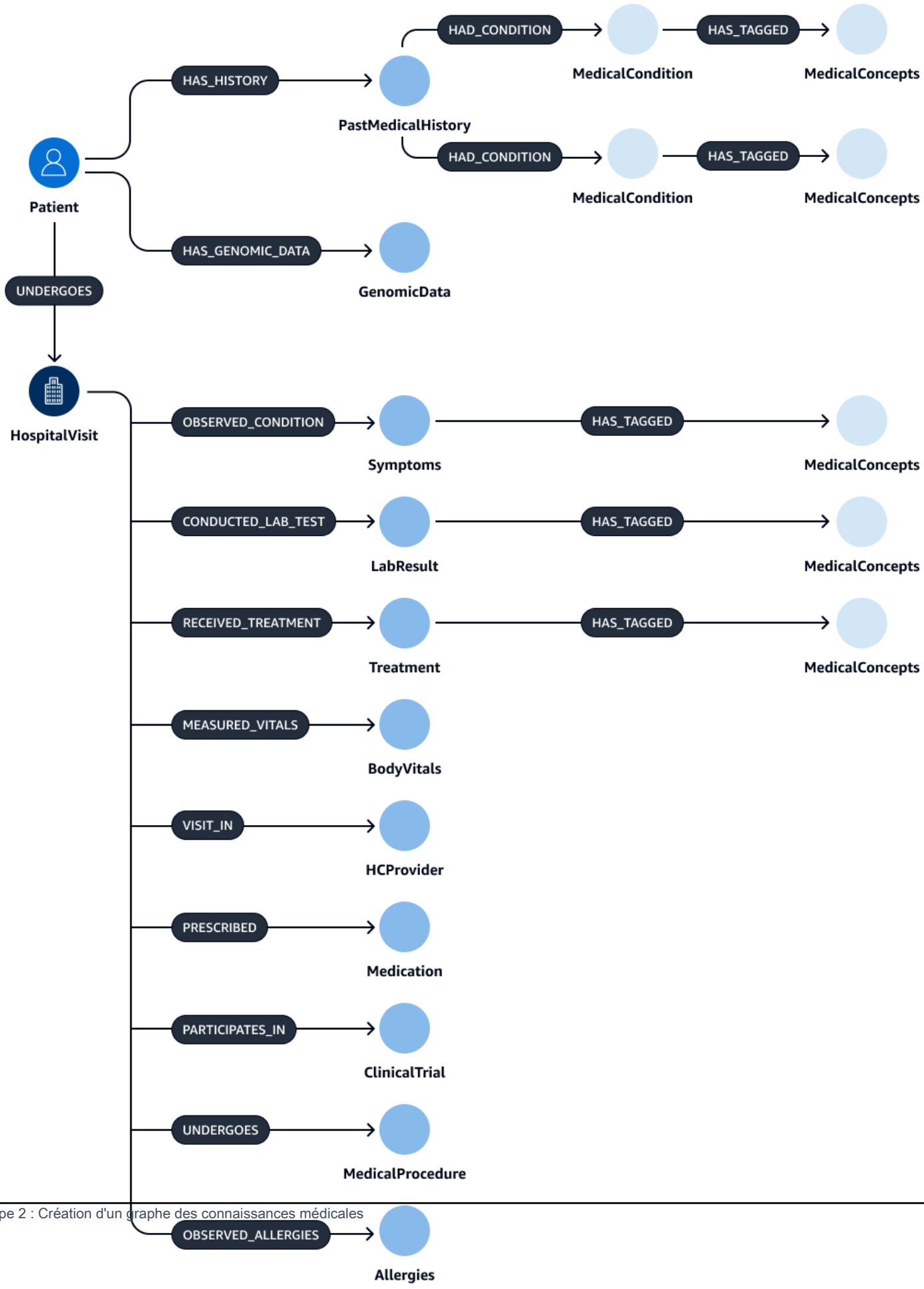
- Extrayez des informations sur des concepts médicaux, tels que les maladies, les médicaments, les dispositifs médicaux, la posologie, la fréquence des médicaments, la durée des médicaments, les symptômes, les procédures médicales et leurs attributs cliniquement pertinents.
- Capturez les caractéristiques fonctionnelles, telles que les relations temporelles entre les entités extraites, les sujets et les évaluations de certitude.
- Élargissez les vocabulaires médicaux standard, tels que les suivants :
  - Identifiants conceptuels (RxCUI) issus de la base de données RxNorm
  - Codes de la classification internationale des maladies, 10e révision, modification clinique (ICD-10-CM)
  - Termes tirés des rubriques médicales (MeSH)
  - Concepts issus de la nomenclature systématisée de la médecine, termes cliniques (SNOMED CT)

  - Codes issus du système de langage médical unifié (UMLS)
- Résumez les notes de sortie et tirez des informations médicales des transcriptions.

La figure suivante montre les étapes d'extraction d'entités et de validation du schéma pour créer des combinaisons appariées valides d'entités, d'attributs et de relations. Vous pouvez stocker des données non structurées, telles que des résumés de sortie ou des notes de patients, dans Amazon Simple Storage Service (Amazon S3). Vous pouvez stocker des données structurées, telles que les données de planification des ressources d'entreprise (ERP), les dossiers médicaux électroniques et les systèmes d'information de laboratoire, dans Amazon Redshift et Amazon DynamoDB. Vous pouvez créer un agent de création d'entités Amazon Bedrock. Cet agent peut intégrer des services, tels que les pipelines d'extraction de données Amazon SageMaker AI, Amazon Textract et Amazon Comprehend Medical, pour extraire des entités, des relations et des attributs à partir de sources de données structurées et non structurées. Enfin, vous utilisez un agent de validation de schéma Amazon Bedrock pour vous assurer que les entités et les relations extraites sont conformes au schéma de graphe prédéfini et préservent l'intégrité des connexions au bord du nœud et des propriétés associées.



Après extraction et validation des entités, des relations et des attributs, vous pouvez les lier pour créer un triplet sujet-objet-prédicat. Vous ingérez ces données dans une base de données de graphes Amazon Neptune, comme illustré dans la figure suivante. Les bases de données de graphes sont optimisées pour stocker et interroger les relations entre les éléments de données.


HAD_CONDITION
HAS_TAGGED
MedicalCondition
MedicalConcepts
HAS_HISTORY
PastMedicalHistory
HAD_CONDITION
HAS_TAGGED
MedicalCondition
MedicalConcepts
Patient
HAS_GENOMIC_DATA
GenomicData
UNDERGOES
HospitalVisit
OBSERVED_CONDITION
HAS_TAGGED
Symptoms
MedicalConcepts
CONDUCTED_LAB_TEST
HAS_TAGGED
LabResult
MedicalConcepts
RECEIVED_TREATMENT
HAS_TAGGED
Treatment
MedicalConcepts
MEASURED_VITALS
BodyVitals
VISIT_IN
HCProvider
PRESCRIBED
Medication
PARTICIPATES_IN
ClinicalTrial
UNDERGOES
MedicalProcedure
OBSERVED_ALLERGIES
Allergies

Vous pouvez créer un graphe de connaissances complet à partir de ces données. Un graphe de connaissances vous aide à organiser et à interroger toutes sortes d'informations connectées. Par exemple, vous pouvez créer un graphe de connaissances comportant les nœuds principaux suivants : `HospitalVisitPastMedicalHistory`,`Symptoms`,`Medication`,`MedicalProcedures`, et`Treatment`.

Les tableaux suivants répertorient les entités et leurs attributs que vous pouvez extraire des notes de décharge.

Entité	Attributes
`Patient`	`PatientID` , `Name`, `Age`, `Gender`, `Address`, `ContactInformation`
`HospitalVisit`	`VisitDate` , `Reason`, `Notes`
`HealthcareProvider`	`ProviderID` , `Name`, `Specialty` , `ContactInformation` , `Address`, `AffiliatedInstitution`
`Symptoms`	`Description` , `RiskFactors`
`Allergies`	`AllergyType` , `Duration`
`Medication`	`MedicationID` , `Name`, `Description` , `Dosage`, `SideEffects` , `Manufacturer`
`PastMedicalHistory`	`ContinuingMedicines`
`MedicalCondition`	`ConditionName` , `Severity`, `Treatment Received` , `DoctorinCharge` , `HospitalName` , `MedicinesFollowed`
`BodyVitals`	`HeartRate` , `BloodPressure` , `RespiratoryRate` , `BodyTemperature` , `BMI`
`LabResult`	`LabResultID` , `PatientID` , `TestName`, `Result`, `Date`

Entité	Attributes
ClinicalTrial	TrialID, Name, Description , Phase, Status, StartDate , EndDate
GenomicData	GenomicDataID , PatientID , SequenceData , VariantInformation
Treatment	TreatmentID , Name, Description , Type, SideEffects
MedicalProcedure	ProcedureID , Name, Description , Risks, Outcomes
MedicalConcepts	UMLSCodes , MedicalVocabularies

Le tableau suivant répertorie les relations que les entités peuvent avoir et leurs attributs correspondants. Par exemple, l'Patiententité peut se connecter à l'HospitalVisitentité associée à la [UNDERGOES] relation. L'attribut de cette relation estVisitDate.

Entité visée	Relation	Entité d'objet	Attributes
Patient	[UNDERGOES]	HospitalVisit	VisitDate
HospitalVisit	[VISIT_IN]	HealthcareProvider	ProviderName , Location, ProviderID , VisitDate
HospitalVisit	[OBSERVED_CONDITION]	Symptoms	Severity, CurrentStatus , VisitDate
HospitalVisit	[RECEIVED_TREATMENT]	Treatment	Duration, Dosage, VisitDate

Entité visée	Relation	Entité d'objet	Attributes
HospitalVisit	[PRESCRIBED]	Medication	Duration, Dosage, Adherence , VisitDate
Patient	[HAS_HISTORY]	PastMedicalHistory	Aucune
PastMedicalHistory	[HAD_CONDITION]	MedicalCondition	DiagnosisDate , CurrentStatus
HospitalVisit	[PARTICIPATES_IN]	ClinicalTrial	VisitDate , Status, Outcomes
Patient	[HAS_GENOMIC_DATA]	GenomicData	CollectionDate
HospitalVisit	[OBSERVED_ALLERGIES]	Allergies	VisitDate
HospitalVisit	[CONDUCTED_LAB_TEST]	LabResult	VisitDate , AnalysisDate , Interpretation
HospitalVisit	[UNDERGOES]	MedicalProcedure	VisitDate , Outcome
MedicalCondition	[HAS_TAGGED]	MedicalConcepts	Aucun
LabResult	[HAS_TAGGED]	MedicalConcepts	Aucun
Treatment	[HAS_TAGGED]	MedicalConcepts	Aucun
Symptoms	[HAS_TAGGED]	MedicalConcepts	Aucune

Étape 3 : Création d'agents de récupération du contexte pour interroger le graphe des connaissances médicales

Après avoir créé la base de données de graphes médicaux, l'étape suivante consiste à créer des agents pour l'interaction graphique. Les agents récupèrent le contexte correct et requis pour la requête saisie par un médecin ou un clinicien. Il existe plusieurs options pour configurer ces agents qui récupèrent le contexte à partir du Knowledge Graph :

- Agents Amazon Bedrock
- Agents LangChain

Agents Amazon Bedrock pour l'interaction graphique

Les agents Amazon Bedrock fonctionnent parfaitement avec les bases de données graphiques Amazon Neptune. Vous pouvez effectuer des interactions avancées via les groupes d'actions Amazon Bedrock. Le groupe d'actions lance le processus en appelant une AWS Lambda fonction qui exécute les requêtes Neptune OpenCypher.

Pour interroger un graphe de connaissances, vous pouvez utiliser deux approches distinctes : exécution directe de requêtes ou interrogation avec intégration de contexte. Ces approches peuvent être appliquées indépendamment ou combinées, en fonction de votre cas d'utilisation spécifique et de vos critères de classement. En combinant les deux approches, vous pouvez fournir un contexte plus complet au LLM, ce qui peut améliorer les résultats. Les deux approches d'exécution des requêtes sont les suivantes :

- Exécution directe de requêtes Cypher sans intégration — La fonction Lambda exécute des requêtes directement sur Neptune sans aucune recherche basée sur les intégrations. Voici un exemple de cette approche :

```
MATCH (p:Patient)-[u:UNDERGOES]->(h:HospitalVisit) WHERE h.Reason = 'Acute Diabetes'
 AND date(u.VisitDate) > date('2024-01-01')
RETURN p.PatientID, p.Name, p.Age, p.Gender, p.Address, p.ContactInformation
```

- Exécution directe de requêtes Cypher à l'aide de la recherche incorporée — La fonction Lambda utilise la recherche incorporée pour améliorer les résultats des requêtes. Cette approche améliore l'exécution des requêtes en incorporant des intégrations, qui sont des représentations vectorielles denses de données. Les intégrations sont particulièrement utiles lorsque la requête nécessite une

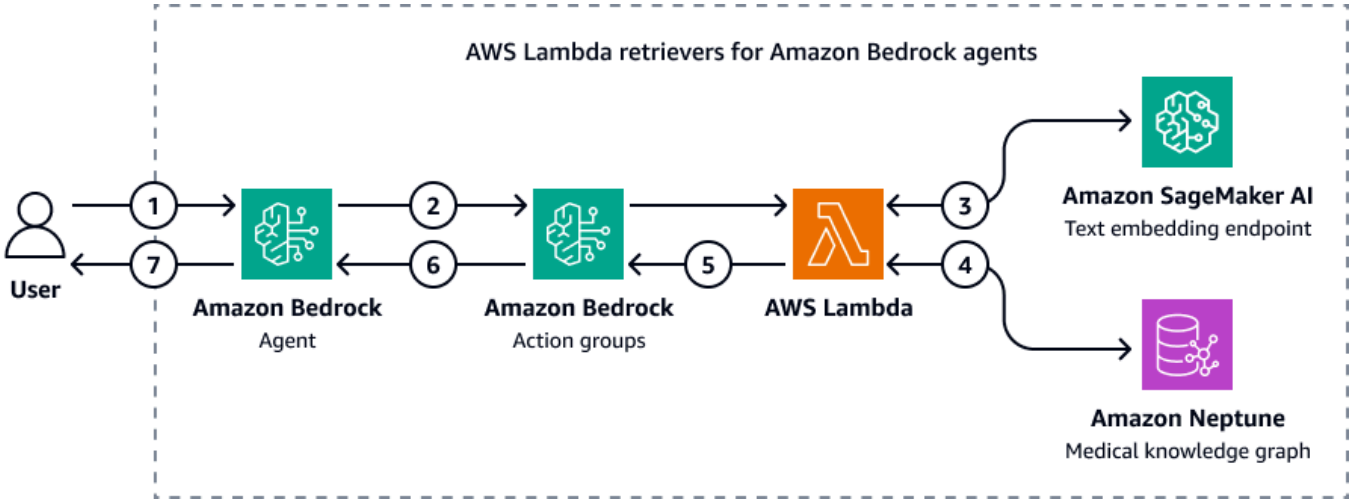
similitude sémantique ou une compréhension plus large au-delà des correspondances exactes. Vous pouvez utiliser des modèles pré-entraînés ou personnalisés pour générer des intégrations pour chaque condition médicale. Voici un exemple de cette approche :

```
CALL { WITH "Acute Diabetes" AS query_term RETURN search_embedding(query_term) AS
 similar_reasons }

MATCH (p:Patient)-[u:UNDERGOES]->(h:HospitalVisit) WHERE h.Reason IN similar reasons
 AND date(u.VisitDate) > date('2024-01-01')
RETURN p.PatientID, p.Name, p.Age, p.Gender, p.Address, p.ContactInformation
```

Dans cet exemple, la `search_embedding("Acute Diabetes")` fonction récupère des affections dont la sémantique est proche du « diabète aigu ». Cela permet à la requête de trouver également des patients atteints de maladies telles que le prédiabète ou le syndrome métabolique.

L'image suivante montre comment les agents Amazon Bedrock interagissent avec Amazon Neptune afin d'exécuter une requête Cypher sur un graphe de connaissances médicales.



Le schéma suivant illustre le flux de travail suivant :

1. L'utilisateur soumet une question à l'agent Amazon Bedrock.
2. L'agent Amazon Bedrock transmet la question et les variables du filtre d'entrée aux groupes d'action Amazon Bedrock. Ces groupes d'actions contiennent une AWS Lambda fonction qui interagit avec le point de terminaison d'intégration de texte Amazon SageMaker AI et le graphe de connaissances médicales Amazon Neptune.

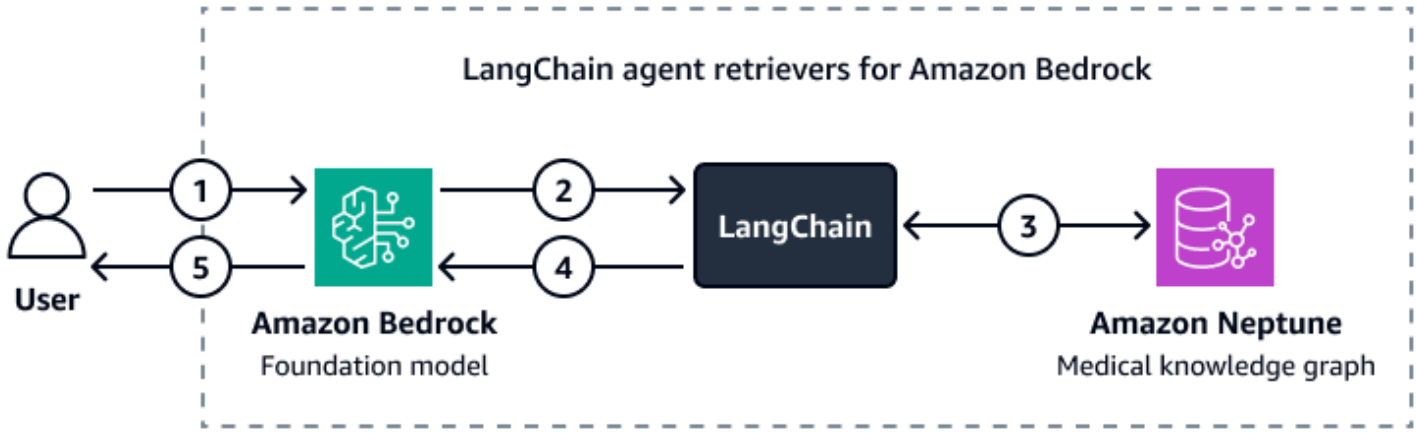
3. La fonction Lambda s'intègre au point de terminaison d'intégration de texte SageMaker AI pour effectuer une recherche sémantique dans la requête OpenCypher. Il convertit la requête en langage naturel en une requête OpenCypher en utilisant des agents sous-jacentsLangChain.
4. La fonction Lambda interroge le graphe de connaissances médicales de Neptune pour trouver le jeu de données correct et reçoit le résultat du graphe de connaissances médicales de Neptune.
5. La fonction Lambda renvoie les résultats de Neptune aux groupes d'actions Amazon Bedrock.
6. Les groupes d'action Amazon Bedrock envoient le contexte récupéré à l'agent Amazon Bedrock.
7. L'agent Amazon Bedrock génère la réponse en utilisant la requête utilisateur d'origine et le contexte extrait du graphe de connaissances.

LangChain agents pour l'interaction graphique

Vous pouvez intégrer Neptune pour permettre LangChain des requêtes et des extractions basées sur des graphes. Cette approche peut améliorer les AI-driven flux de travail en utilisant les fonctionnalités de base de données de graphes de Neptune. Le Custom LangChain Retriever joue le rôle d'intermédiaire. Le modèle de base d'Amazon Bedrock peut interagir avec Neptune en utilisant à la fois des requêtes Cypher directes et des algorithmes de graphes plus complexes.

Vous pouvez utiliser le récupérateur personnalisé pour affiner la façon dont l'LangChainagent interagit avec les algorithmes du graphe Neptune. Par exemple, vous pouvez utiliser des instructions ponctuelles, qui vous permettent d'adapter les réponses du modèle de base en fonction de modèles ou d'exemples spécifiques. Vous pouvez également appliquer des LLM-identified filtres pour affiner le contexte et améliorer la précision des réponses. Cela peut améliorer l'efficacité et la précision de l'ensemble du processus de récupération lors de l'interaction avec des données graphiques complexes.

L'image suivante montre comment un LangChain agent personnalisé orchestre l'interaction entre un modèle de fondation Amazon Bedrock et un graphe de connaissances médicales Amazon Neptune.



Le schéma suivant illustre le flux de travail suivant :

1. Un utilisateur soumet une question à Amazon Bedrock et à l'LangChainagent.
2. Le modèle Amazon Bedrock Foundation utilise le schéma Neptune, fourni par LangChain l'agent, pour générer une requête pour la question de l'utilisateur.
3. L'LangChainagent exécute la requête sur le graphe de connaissances médicales d'Amazon Neptune.
4. L'LangChainagent envoie le contexte récupéré au modèle de fondation Amazon Bedrock.
5. Le modèle Amazon Bedrock Foundation utilise le contexte extrait pour générer une réponse à la question de l'utilisateur.

Étape 4 : Création d'une base de connaissances contenant des données descriptives en temps réel

Ensuite, vous créez une base de connaissances contenant des notes descriptives en temps réel sur les interactions médecin-patient, des évaluations d'images diagnostiques et des rapports d'analyse de laboratoire. Cette base de connaissances est une base de données vectorielle. En utilisant une base de données vectorielle, qui peut stocker des connaissances médicales descriptives sous une forme indexée et vectorisée, les prestataires de soins de santé peuvent consulter et accéder efficacement aux informations pertinentes à partir d'un vaste référentiel. Ces représentations vectorisées vous aident à récupérer des données sémantiquement similaires. Les prestataires de soins peuvent parcourir rapidement les notes cliniques, les images médicales et les résultats de laboratoire. Cela accélère la prise de décision éclairée en offrant un accès instantané à des

informations contextuelles pertinentes, améliorant ainsi la précision et la rapidité des diagnostics et des plans de traitement.

Utilisation d'une base de connaissances médicales de OpenSearch service

Amazon OpenSearch Service peut gérer de gros volumes de données médicales de grande dimension. Il s'agit d'un service géré qui facilite les recherches de haute performance et les analyses en temps réel. Elle convient parfaitement comme base de données vectorielle pour les applications RAG. OpenSearch Le service agit comme un outil principal pour gérer de grandes quantités de données non structurées ou semi-structurées, telles que les dossiers médicaux, les articles de recherche et les notes cliniques. Ses fonctionnalités avancées de recherche sémantique vous aident à récupérer des informations contextuellement pertinentes. Cela le rend particulièrement utile dans des applications telles que les systèmes d'aide à la décision clinique, les outils de résolution des requêtes des patients et les systèmes de gestion des connaissances dans le domaine de la santé. Par exemple, un clinicien peut rapidement trouver des données pertinentes sur les patients ou des études de recherche correspondant à des symptômes ou à des protocoles de traitement spécifiques. Cela aide les cliniciens à prendre des décisions éclairées par les informations les plus récentes et les plus pertinentes.

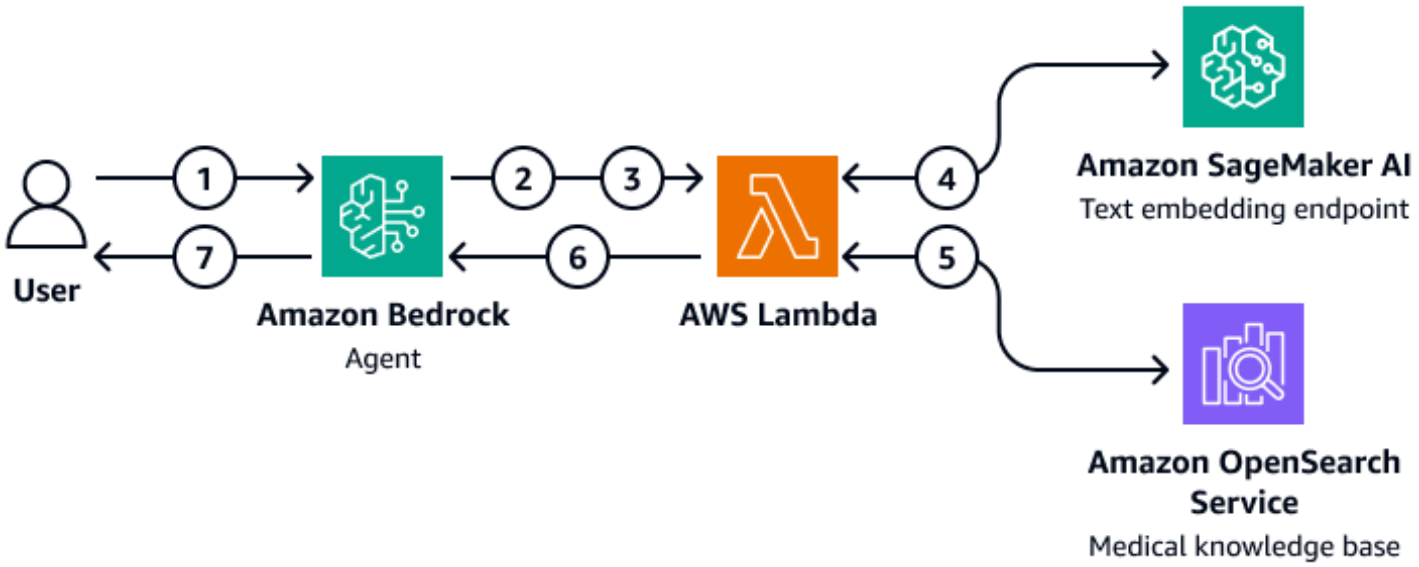
OpenSearch Le service peut évoluer et gérer l'indexation et l'interrogation des données en temps réel. Il est donc idéal pour les environnements de santé dynamiques où l'accès rapide à des informations précises est essentiel. En outre, il dispose de fonctionnalités de recherche multimodales qui sont optimales pour les recherches nécessitant plusieurs entrées, telles que des images médicales et des notes de médecin. Lors de la mise en œuvre du OpenSearch service pour les applications de santé, il est essentiel de définir des champs et des mappages précis afin d'optimiser l'indexation et la récupération des données. Les champs représentent les différents éléments de données, tels que les dossiers des patients, les antécédents médicaux et les codes de diagnostic. Les mappages définissent la manière dont ces champs sont stockés (sous forme incorporée ou sous forme originale) et interrogés. Pour les applications de santé, il est essentiel d'établir des mappages adaptés à différents types de données, notamment les données structurées (telles que les résultats de tests numériques), les données semi-structurées (telles que les notes des patients) et les données non structurées (telles que les images médicales)

Dans OpenSearch Service, vous pouvez effectuer des requêtes de recherche neurale en texte intégral à l'aide d'instructions organisées pour effectuer des recherches dans les dossiers médicaux, les notes cliniques ou les documents de recherche afin de trouver rapidement des informations pertinentes sur des symptômes, des traitements ou des antécédents de patients spécifiques. Les requêtes de recherche neuronale gèrent automatiquement l'intégration de l'invite d'entrée et des

images à l'aide de modèles de réseaux neuronaux intégrés. Cela l'aide à comprendre et à saisir les relations sémantiques les plus profondes des données multimodales, offrant ainsi des résultats de recherche plus précis et plus sensibles au contexte par rapport à d'autres algorithmes de requête de recherche, tels que la recherche K-Nearest Neighbor (K-nn).

Création d'une architecture RAG

Vous pouvez déployer une solution RAG personnalisée qui utilise les agents Amazon Bedrock pour interroger une base de connaissances médicales dans OpenSearch Service. Pour ce faire, vous créez une AWS Lambda fonction qui peut interagir avec le OpenSearch service et l'interroger. La fonction Lambda intègre la question saisie par l'utilisateur en accédant à un point de terminaison d'intégration de texte basé sur SageMaker l'IA. L'agent Amazon Bedrock transmet des paramètres de requête supplémentaires en tant qu'entrées à la fonction Lambda. La fonction interroge la base de connaissances médicales dans OpenSearch Service, qui renvoie le contenu médical pertinent. Après avoir configuré la fonction Lambda, ajoutez-la en tant que groupe d'action dans l'agent Amazon Bedrock. L'agent Amazon Bedrock prend les données saisies par l'utilisateur, identifie les variables nécessaires, transmet les variables et la question à la fonction Lambda, puis lance la fonction. La fonction renvoie un contexte qui aide le modèle de base à fournir une réponse plus précise à la question de l'utilisateur.

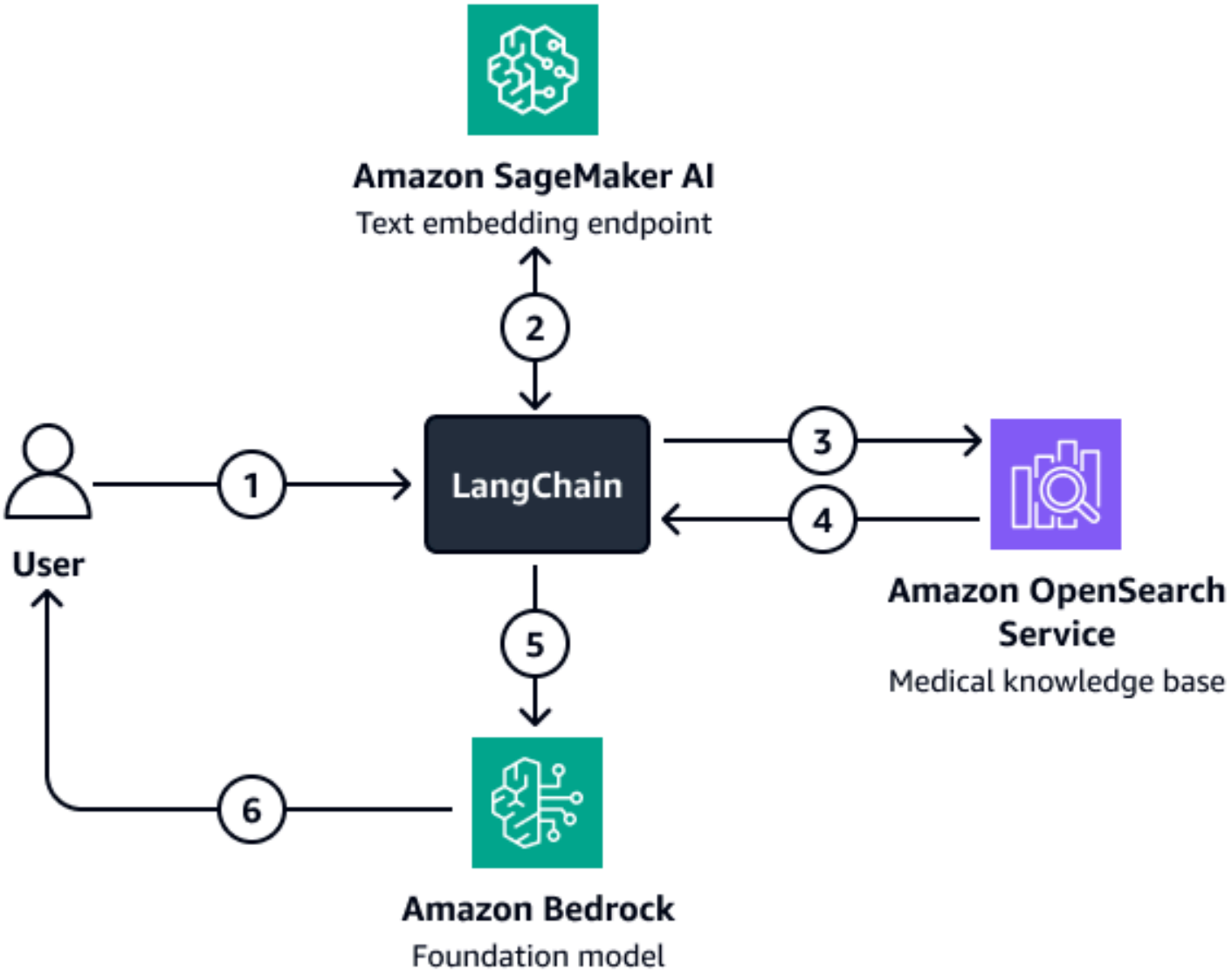


Le schéma suivant illustre le flux de travail suivant :

1. Un utilisateur soumet une question à l'agent Amazon Bedrock.
2. L'agent Amazon Bedrock sélectionne le groupe d'action à lancer.

3. L'agent Amazon Bedrock lance une AWS Lambda fonction et lui transmet des paramètres.
4. La fonction Lambda lance le modèle d'intégration de texte Amazon SageMaker AI pour intégrer la question de l'utilisateur.
5. La fonction Lambda transmet le texte intégré ainsi que les paramètres et filtres supplémentaires à Amazon OpenSearch Service. Amazon OpenSearch Service interroge la base de connaissances médicales et renvoie les résultats à la fonction Lambda.
6. La fonction Lambda transmet les résultats à l'agent Amazon Bedrock.
7. Le modèle de base de l'agent Amazon Bedrock génère une réponse basée sur les résultats et renvoie la réponse à l'utilisateur.

Pour les situations impliquant un filtrage plus complexe, vous pouvez utiliser un LangChain récupérateur personnalisé. Créez ce récupérateur en configurant un client de recherche vectorielle OpenSearch Service qui est chargé directement dansLangChain. Cette architecture vous permet de transmettre davantage de variables afin de créer les paramètres du filtre. Une fois le récupérateur configuré, utilisez le modèle et le récupérateur Amazon Bedrock pour configurer une chaîne de questions-réponses. Cette chaîne orchestre l'interaction entre le modèle et le récupérateur en transmettant les entrées utilisateur et les filtres potentiels au récupérateur. Le récupérateur renvoie un contexte pertinent qui aide le modèle de base à répondre à la question de l'utilisateur.



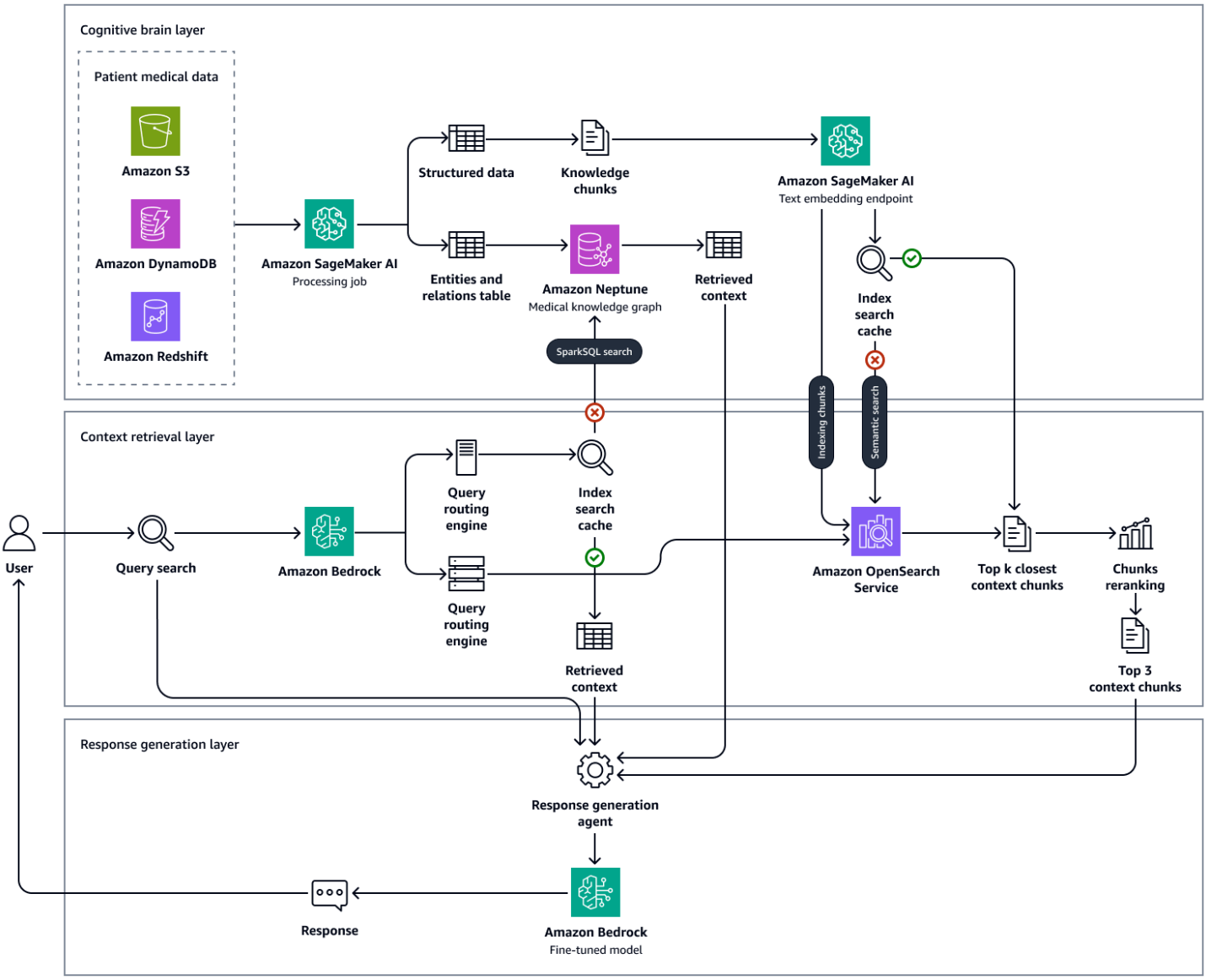
Le schéma suivant illustre le flux de travail suivant :

1. Un utilisateur soumet une question à l'agent LangChain de récupération.
2. L'agent de LangChain récupération envoie la question au point de terminaison d'intégration de texte Amazon SageMaker AI pour l'intégrer.
3. L'agent LangChain de récupération transmet le texte intégré à Amazon OpenSearch Service.
4. Amazon OpenSearch Service renvoie les documents récupérés à l'agent de LangChain récupération.
5. L'agent de LangChain récupération transmet la question de l'utilisateur et le contexte extrait au modèle Amazon Bedrock Foundation.
6. Le modèle de base génère une réponse et l'envoie à l'utilisateur.

Étape 5 : Utiliser les LLM pour répondre à des questions médicales

Les étapes précédentes vous aident à créer une application de renseignement médical capable de récupérer les dossiers médicaux d'un patient et de résumer les médicaments pertinents et les diagnostics potentiels. À présent, vous créez la couche de génération. Cette couche utilise les capacités génératives d'un LLM dans Amazon Bedrock, comme Llama 3, pour augmenter le rendement de l'application.

Lorsqu'un clinicien saisit une requête, la couche de récupération du contexte de l'application exécute le processus de récupération à partir du graphe de connaissances et renvoie les principaux enregistrements relatifs à l'historique, à la démographie, aux symptômes, au diagnostic et aux résultats du patient. À partir de la base de données vectorielle, il extrait également des notes descriptives en temps réel sur les interactions médecin-patient, des informations sur l'évaluation des images diagnostiques, des résumés de rapports d'analyse de laboratoire et des informations issues d'un vaste corpus de recherches médicales et de livres universitaires. Les meilleurs résultats obtenus, la requête du clinicien et les instructions (qui sont conçues pour sélectionner les réponses en fonction de la nature de la requête) sont ensuite transmis au modèle de base d'Amazon Bedrock. Il s'agit de la couche de génération de réponses. Le LLM utilise le contexte récupéré pour générer une réponse à la requête du clinicien. La figure suivante montre le flux de travail de bout en bout des étapes de cette solution.



Vous pouvez utiliser un modèle de base préformé dans Amazon Bedrock, tel que Llama 3, pour une gamme de cas d'utilisation que l'application de renseignement médical doit gérer. Le LLM le plus efficace pour une tâche donnée varie en fonction du cas d'utilisation. Par exemple, un modèle préformé peut être suffisant pour résumer les conversations patient-médecin, effectuer des recherches dans les médicaments et les antécédents des patients, et extraire des informations à partir d'ensembles de données médicales internes et de corpus de connaissances scientifiques. Cependant, un LLM affiné peut être nécessaire pour d'autres cas d'utilisation complexes, tels que les évaluations de laboratoire en temps réel, les recommandations de procédures médicales et les prédictions des résultats pour les patients. Vous pouvez affiner un LLM en l'entraînant sur des ensembles de données du domaine médical. Des exigences spécifiques ou complexes dans le

domaine des soins de santé et des sciences de la vie orientent le développement de ces modèles affinés.

Pour plus d'informations sur la mise au point d'un LLM ou sur le choix d'un LLM existant qui a été formé sur les données du domaine médical, voir Utilisation de grands modèles linguistiques pour les cas d'utilisation dans les domaines de la santé et des sciences de la vie.

Alignement sur le AWS Well-Architected Framework

La solution s'aligne sur les six piliers du AWS Well-Architected cadre comme suit :

- Excellence opérationnelle — L'architecture est découplée pour une surveillance et des mises à jour efficaces. Les agents Amazon Bedrock vous AWS Lambda aident à déployer et à annuler rapidement des outils.
- Sécurité — Cette solution est conçue pour se conformer aux réglementations sanitaires, telles que la loi HIPAA. Vous pouvez également implémenter le chiffrement, un contrôle d'accès précis et des garde-corps Amazon Bedrock pour protéger les données des patients.
- Fiabilité : les services AWS gérés, tels qu'Amazon OpenSearch Service et Amazon Bedrock, fournissent l'infrastructure nécessaire à une interaction continue avec les modèles.
- Efficacité des performances — La solution RAG récupère rapidement les données pertinentes en utilisant une recherche sémantique optimisée et des requêtes Cypher, tandis qu'un agent-routeur identifie les itinéraires optimaux pour les requêtes des utilisateurs.
- Optimisation des coûts — Le modèle de paiement par jeton d'Amazon Bedrock et de l'architecture RAG réduit les coûts d'inférence et de pré-formation.
- Durabilité — L'utilisation d'une infrastructure sans serveur et d'un calcul basé sur le paiement par jeton minimise l'utilisation des ressources et améliore la durabilité.

Cas d'utilisation : prévision des résultats pour les patients et des taux de réadmission

AI-powered l'analyse prédictive offre d'autres avantages en prévoyant les résultats pour les patients et en permettant des plans de traitement personnalisés. Cela peut améliorer la satisfaction des patients et les résultats de santé. En intégrant ces capacités d'IA à Amazon Bedrock et à d'autres technologies, les prestataires de soins de santé peuvent réaliser des gains de productivité significatifs, réduire les coûts et améliorer la qualité globale des soins aux patients.

Vous pouvez stocker des données médicales, telles que les antécédents des patients, les notes cliniques, les médicaments et les traitements, dans un graphe de connaissances. En combinant la compréhension contextuelle approfondie des LLM avec les données temporelles structurées d'un graphe de connaissances médicales, les prestataires de soins de santé peuvent obtenir des informations supplémentaires sur les modèles individuels des patients. Grâce à l'analyse prédictive, vous pouvez identifier à un stade précoce les cas potentiels de non-observance ou de complications liées au traitement et générer des scores personnalisés de propension à la réadmission.

Cette solution vous permet de prévoir la probabilité d'une réadmission. Ces prévisions peuvent améliorer les résultats pour les patients et réduire les coûts des soins de santé. Cette solution peut également aider les cliniciens et les administrateurs des hôpitaux à concentrer leur attention sur les patients présentant un risque plus élevé de réadmission. Cela les aide également à lancer des interventions proactives auprès de ces patients par le biais d'alertes, de libre-service et d'actions basées sur les données.

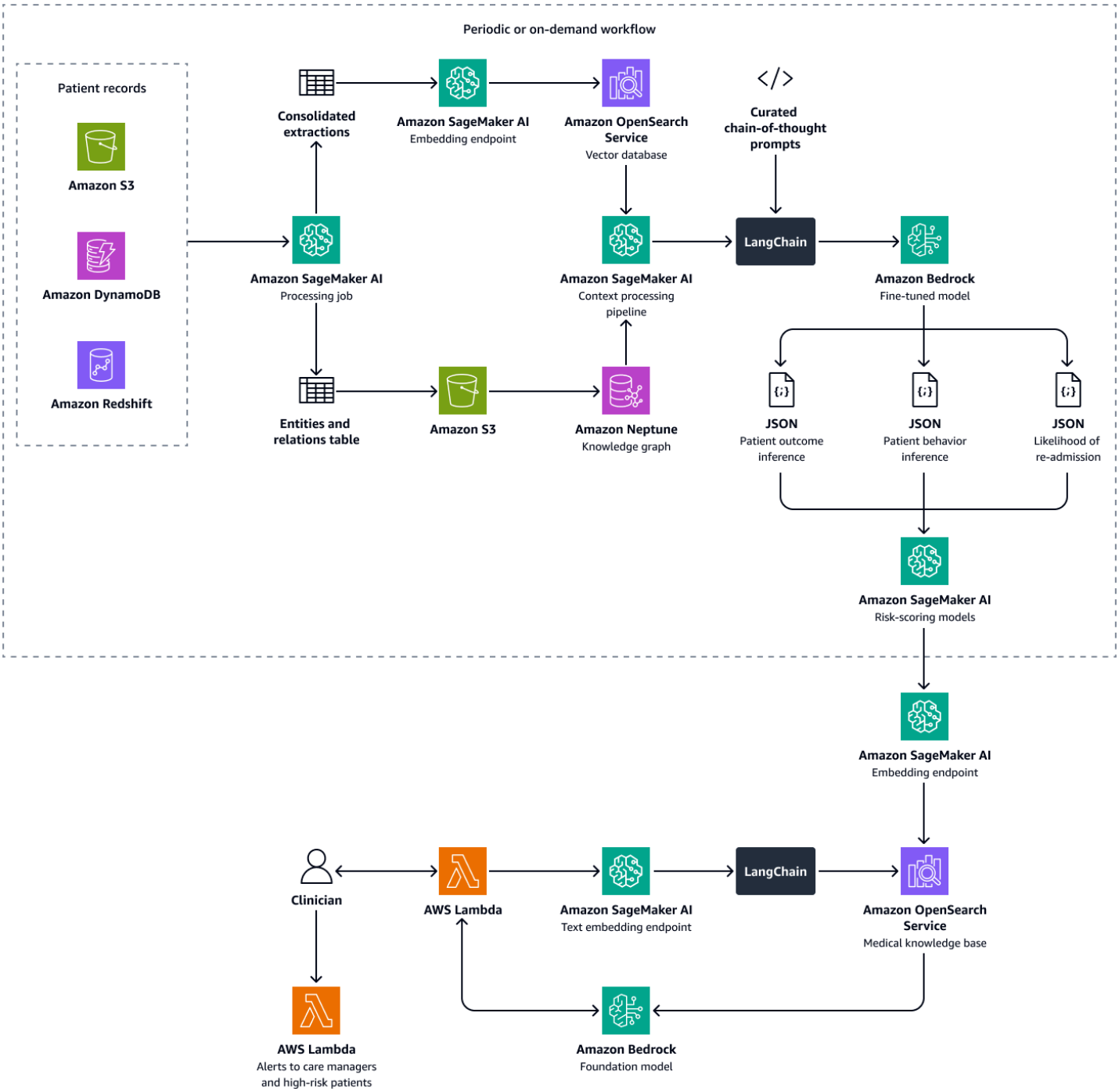
Présentation de la solution

Cette solution utilise un framework de génération augmentée de récupération multiple (RAG) pour analyser les données des patients. Il prédit la probabilité de réadmission à l'hôpital pour chaque patient et vous aide à calculer un score de propension à la réadmission au niveau de l'hôpital. Cette solution intègre les fonctionnalités suivantes :

- Graphe de connaissances — Stocke des données structurées et chronologiques sur les patients, telles que les consultations à l'hôpital, les réadmissions précédentes, les symptômes, les résultats de laboratoire, les traitements prescrits et l'historique d'observance du traitement

- Base de données vectorielle — Stocke des données cliniques non structurées, telles que les résumés des sorties, les notes du médecin et les enregistrements des rendez-vous manqués ou des effets secondaires signalés des médicaments
- Fine-tuned LLM — Consomme à la fois les données structurées du graphe de connaissances et les données non structurées de la base de données vectorielle afin de générer des inférences sur le comportement d'un patient, l'adhésion au traitement et la probabilité de réadmission

Les modèles de notation des risques quantifient les inférences du LLM sous forme de scores numériques. Vous pouvez agréger les scores pour obtenir un score de propension à la réadmission au niveau de l'hôpital. Ce score définit l'exposition au risque de chaque patient, et vous pouvez le calculer périodiquement ou selon les besoins. Toutes les inférences et les scores de risque sont indexés et stockés dans Amazon OpenSearch Service afin que les responsables de soins et les cliniciens puissent les récupérer. En intégrant un agent d'intelligence artificielle conversationnel à cette base de données vectorielle, les cliniciens et les responsables de soins peuvent extraire facilement des informations au niveau du patient, de l'ensemble de l'établissement ou de la spécialité médicale. Vous pouvez également configurer des alertes automatisées en fonction des scores de risque, ce qui encourage les interventions proactives.



La création de cette solution comprend les étapes suivantes :

- Étape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales
- Étape 2 : Prédire le comportement du patient à l'égard des médicaments ou traitements prescrits
- Étape 3 : Prédire la probabilité de réadmission du patient

- Étape 4 : Calcul du score de propension à la réadmission à l'hôpital

Étape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales

Dans Amazon Neptune, vous pouvez utiliser un graphe de connaissances pour stocker des informations temporelles sur les visites des patients et les résultats obtenus au fil du temps. Le moyen le plus efficace de créer et de stocker un graphe de connaissances consiste à utiliser un modèle de graphe et une base de données de graphes. Les bases de données de graphes sont spécialement conçues pour stocker et parcourir les relations. Les bases de données graphiques facilitent la modélisation et la gestion de données hautement connectées et disposent de schémas flexibles.

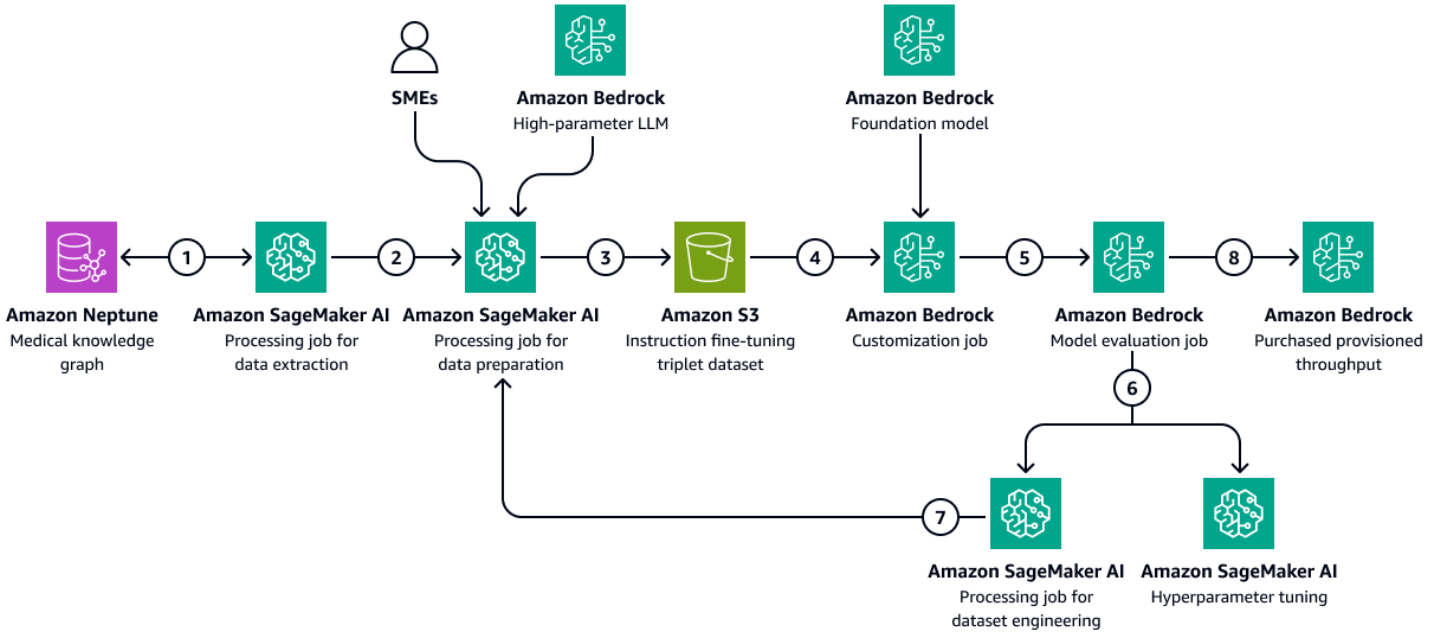
Le graphe de connaissances vous permet d'effectuer une analyse de séries chronologiques. Les éléments clés de la base de données de graphes utilisés pour la prédiction temporelle des résultats pour les patients sont les suivants :

- Données historiques — Diagnostics antérieurs, traitement continu, médicaments déjà utilisés et résultats de laboratoire pour le patient
- Visites des patients (chronologique) — Dates des visites, symptômes, allergies observées, notes cliniques, diagnostics, procédures, traitements, médicaments prescrits et résultats de laboratoire
- Symptômes et paramètres cliniques — Informations cliniques et basées sur les symptômes, notamment la gravité, les schémas de progression et la réponse du patient au médicament

Vous pouvez utiliser les informations du graphique des connaissances médicales pour peaufiner un LLM dans Amazon Bedrock, tel que Llama 3. Vous affinez le LLM à l'aide de données séquentielles du patient concernant la réponse du patient à un ensemble de médicaments ou de traitements au fil du temps. Utilisez un ensemble de données étiqueté qui classe un ensemble de médicaments ou de traitements et de données d'interaction patient-clinique dans des catégories prédéfinies indiquant l'état de santé d'un patient. La détérioration de l'état de santé, l'amélioration ou la stabilité des progrès sont des exemples de ces catégories. Lorsque le clinicien saisit un nouveau contexte concernant le patient et ses symptômes, le LLM affiné peut utiliser les modèles de l'ensemble de données de formation afin de prédire l'issue potentielle du patient.

L'image suivante montre les étapes séquentielles nécessaires pour peaufiner un LLM dans Amazon Bedrock à l'aide d'un ensemble de données de formation spécifique au secteur de la santé. Ces

données peuvent inclure l'état de santé des patients et les réponses aux traitements au fil du temps. Cet ensemble de données de formation aiderait le modèle à faire des prédictions généralisées sur les résultats pour les patients.



Le schéma suivant illustre le flux de travail suivant :

1. La tâche d'extraction de données Amazon SageMaker AI interroge le graphe de connaissances pour récupérer des données chronologiques sur les réponses des différents patients à un ensemble de médicaments ou de traitements au fil du temps.
2. Le travail de préparation des données basé sur l' SageMaker IA intègre un LLM Amazon Bedrock et des contributions d'experts en la matière (PME). Le travail classe les données extraites du graphe de connaissances dans des catégories prédéfinies (telles que détérioration de l'état de santé, amélioration ou progrès stables) qui indiquent l'état de santé de chaque patient.
3. Le travail crée un ensemble de données affiné qui inclut les informations extraites du graphe de connaissances, des instructions de la chaîne de réflexion et de la catégorie de résultats pour les patients. Il télécharge cet ensemble de données de formation dans un compartiment Amazon S3.
4. Une tâche de personnalisation d'Amazon Bedrock utilise cet ensemble de données de formation pour peaufiner un LLM.
5. Le travail de personnalisation d'Amazon Bedrock intègre le modèle fondamental de choix d'Amazon Bedrock dans l'environnement de formation. Il démarre le travail de réglage fin et utilise le jeu de données d'entraînement et les hyperparamètres d'entraînement que vous configurez.

6. Une tâche d'évaluation Amazon Bedrock évalue le modèle affiné à l'aide d'un cadre d'évaluation de modèle prédéfini.
7. Si le modèle doit être amélioré, la tâche de formation s'exécute à nouveau avec davantage de données après un examen attentif de l'ensemble de données d'apprentissage. Si le modèle ne montre pas d'amélioration progressive des performances, envisagez également de modifier les hyperparamètres d'entraînement.
8. Une fois que l'évaluation du modèle répond aux normes définies par les parties prenantes de l'entreprise, vous hébergez le modèle affiné sur le débit provisionné par Amazon Bedrock.

Étape 2 : Prédire le comportement du patient à l'égard des médicaments ou traitements prescrits

Fine-tuned Les LLM peuvent traiter les notes cliniques, les résumés des sorties et d'autres documents spécifiques au patient à partir du graphe temporel des connaissances médicales. Ils peuvent évaluer si le patient est susceptible de suivre les médicaments ou traitements prescrits.

Cette étape utilise le graphe de connaissances créé dansÉtape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales. Le graphe de connaissances contient des données issues du profil du patient, y compris l'historique d'adhésion du patient en tant que nœud. Cela inclut également les cas de non-observance des médicaments ou des traitements, les effets secondaires des médicaments, le manque d'accès aux médicaments ou les obstacles liés au coût de ceux-ci, ou les schémas posologiques complexes comme caractéristiques de ces nœuds.

Fine-tuned Les LLM peuvent consulter les données relatives à l'exécution des anciennes ordonnances issues du graphe des connaissances médicales et les résumés descriptifs des notes cliniques provenant d'une base de données vectorielle Amazon OpenSearch Service. Ces notes cliniques peuvent mentionner des rendez-vous fréquemment manqués ou le non-respect des traitements. Le LLM peut utiliser ces notes pour prédire la probabilité d'une future non-conformité.

1. Préparez les données d'entrée comme suit :
   - Données structurées — Extrayez les données récentes des patients, telles que les trois dernières visites et les résultats de laboratoire, à partir du graphe des connaissances médicales.
   - Données non structurées : récupérez les notes cliniques récentes de la base de données vectorielle Amazon OpenSearch Service.

2. Créez une invite de saisie qui inclut les antécédents du patient et le contexte actuel. Voici un exemple d'invite :

```
You are a highly specialized AI model trained in healthcare predictive analytics.
 Your task is to analyze a patient's historical medical records, adherence patterns,
 and clinical context to predict the **likelihood of future non-adherence** to
 prescribed medications or treatments.

### **Patient Details**
- **Patient ID:** {patient_id}
- **Age:** {age}
- **Gender:** {gender}
- **Medical Conditions:** {medical_conditions}
- **Current Medications:** {current_medications}
- **Prescribed Treatments:** {prescribed_treatments}

### **Chronological Medical History**
- **Visit Dates & Symptoms:** {visit_dates_symptoms}
- **Diagnoses & Procedures:** {diagnoses_procedures}
- **Prescribed Medications & Treatments:** {medications_treatments}
- **Past Adherence Patterns:** {historical_adherence}
- **Instances of Non-Adherence:** {past_non_adherence}
- **Side Effects Experienced:** {side_effects}
- **Barriers to Adherence (e.g., Cost, Access, Dosing Complexity):** {barriers}

### **Patient-Specific Insights**
- **Clinical Notes & Discharge Summaries:** {clinical_notes}
- **Missed Appointments & Non-Compliance Patterns:** {missed_appointments}

### **Let's think Step-by-Step to predict the patient behaviour**
1. You should first analyze past adherence trends and patterns of non-adherence.
2. Identify potential barriers, such as financial constraints, medication side
 effects, or complex dosing regimens.
3. Thoroughly examine clinical notes and documented patient behaviors that may hint
 at non-adherence.
4. Correlate adherence history with prescribed treatments and patient conditions.
5. Finally predict the likelihood of non-adherence based on these contextual
 insights.

### **Output Format (JSON)**
Return the prediction in the following structured format:
```json
{
```
```

```
  "patient_id": "{patient_id}",
  "likelihood_of_non_adherence": "{low | moderate | high}",
  "reasoning": "{detailed_explanation_based_on_patient_history}"
}
```

3. Passez l'invite au LLM affiné. Le LLM traite l'invite et prédit le résultat. Voici un exemple de réponse du LLM :

```
{
  "patient_id": "P12345",
  "likelihood_of_non_adherence": "high",
  "reasoning": "The patient has a history of missed appointments, has reported side
 effects to previous medications. Additionally, clinical notes indicate difficulty
 following complex dosing schedules."
}
```

4. Analysez la réponse du modèle pour extraire la catégorie de résultats prédite. Par exemple, la catégorie de l'exemple de réponse de l'étape précédente peut être une forte probabilité de non-conformité.
5. (Facultatif) Utilisez des logits modèles ou des méthodes supplémentaires pour attribuer des scores de confiance. Les logits sont les probabilités non normalisées que l'objet appartienne à une certaine classe ou catégorie.

Étape 3 : Prédire la probabilité de réadmission du patient

Les réadmissions à l'hôpital constituent une préoccupation majeure en raison du coût élevé de l'administration des soins de santé et de leur impact sur le bien-être du patient. Le calcul des taux de réadmission dans les hôpitaux est un moyen de mesurer la qualité des soins aux patients et les performances d'un professionnel de santé.

Pour calculer le taux de réadmission, vous avez défini un indicateur, tel qu'un taux de réadmission sur 7 jours. Cet indicateur est le pourcentage de patients admis qui retournent à l'hôpital pour une visite imprévue dans les sept jours suivant leur sortie. Pour prévoir les chances de réadmission d'un patient, un LLM affiné peut utiliser les données temporelles du graphe de connaissances médicales que vous avez créé dans. Étape 1 : Prédire les résultats des patients à l'aide d'un graphe de connaissances médicales Ce graphique de connaissances conserve des enregistrements chronologiques des rencontres avec les patients, des procédures, des médicaments et des symptômes. Ces enregistrements de données contiennent les éléments suivants :

- Durée écoulée depuis la dernière sortie du patient
- La réponse du patient aux traitements et médicaments antérieurs
- L'évolution des symptômes ou des affections au fil du temps

Vous pouvez traiter ces événements chronologiques pour prédire la probabilité de réadmission d'un patient grâce à une invite du système organisée. L'invite transmet la logique de prédiction au LLM affiné.

1. Préparez les données d'entrée comme suit :
   - Historique d'observance — Extrayez les dates de prise des médicaments, les fréquences de renouvellement des médicaments, les détails du diagnostic et des médicaments, les antécédents médicaux chronologiques et d'autres informations du graphique des connaissances médicales.
   - Indicateurs comportementaux — Récupérez et incluez les notes cliniques concernant les rendez-vous manqués et les effets secondaires signalés par les patients.
2. Créez une invite de saisie qui inclut l'historique d'adhésion et les indicateurs comportementaux. Voici un exemple d'invite :

```
You are a highly specialized AI model trained in healthcare predictive analytics.
 Your task is to analyze a patient's historical medical records, clinical events, and
 adherence patterns to predict the **likelihood of hospital readmission** within the
 next few days.

### **Patient Details**
- **Patient ID:** {patient_id}
- **Age:** {age}
- **Gender:** {gender}
- **Primary Diagnoses:** {diagnoses}
- **Current Medications:** {current_medications}
- **Prescribed Treatments:** {prescribed_treatments}

### **Chronological Medical History**
- **Recent Hospital Encounters:** {encounters}
- **Time Since Last Discharge:** {time_since_last_discharge}
- **Previous Readmissions:** {past_readmissions}
- **Recent Lab Results & Vital Signs:** {recent_lab_results}
- **Procedures Performed:** {procedures_performed}
- **Prescribed Medications & Treatments:** {medications_treatments}
- **Past Adherence Patterns:** {historical_adherence}
```

```
- **Instances of Non-Adherence:** {past_non_adherence}

### **Patient-Specific Insights**
- **Clinical Notes & Discharge Summaries:** {clinical_notes}
- **Missed Appointments & Non-Compliance Patterns:** {missed_appointments}
- **Patient-Reported Side Effects & Complications:** {side_effects}

### **Reasoning Process – You have to analyze this use case step-by-step.**
1. First assess **time since last discharge** and whether recent hospital encounters
 suggest a pattern of frequent readmissions.
2. Second examine **recent lab results, vital signs, and procedures performed** to
 identify clinical deterioration.
3. Third analyze **adherence history**, checking if past non-adherence to medications
 or treatments correlates with readmissions.
4. Then identify **missed appointments, self-reported side effects, or symptoms
 worsening** from clinical notes.
5. Finally predict the **likelihood of readmission** based on these contextual
 insights.

### **Output Format (JSON)**
Return the prediction in the following structured format:
```json
{
  "patient_id": "{patient_id}",
  "likelihood_of_readmission": "{low | moderate | high}",
  "reasoning": "{detailed_explanation_based_on_patient_history}"
}
```

3. Passez l'invite au LLM affiné. Le LLM traite l'invite et prédit la probabilité et les raisons de la réadmission. Voici un exemple de réponse du LLM :

```
{
  "patient_id": "P67890",
  "likelihood_of_readmission": "high",
  "reasoning": "The patient was discharged only 5 days ago, has a history of more
 than two readmissions to hospitals where the patient received treatment. Recent
 lab results indicate abnormal kidney function and high liver enzymes. These factors
 suggest a medium risk of readmission."
}
```

4. Catégorisez la prédiction selon une échelle standardisée, telle que faible, moyenne ou élevée.
```

5. Passez en revue le raisonnement fourni par le LLM et identifiez les facteurs clés qui contribuent à la prédiction.
6. Associez les résultats qualitatifs aux scores quantitatifs. Par exemple, une probabilité très élevée peut être égale à 0,9.
7. Utilisez des ensembles de données de validation pour calibrer les résultats du modèle par rapport aux taux de réadmission réels.

Étape 4 : Calcul du score de propension à la réadmission à l'hôpital

Ensuite, vous calculez un score de propension à la réadmission à l'hôpital par patient. Ce score reflète l'impact net des trois analyses effectuées au cours des étapes précédentes : les résultats potentiels pour les patients, le comportement du patient à l'égard des médicaments et des traitements, et la probabilité de réadmission du patient. En agrégeant le score de propension à la réadmission du patient au niveau de la spécialité, puis au niveau de l'hôpital, vous pouvez obtenir des informations utiles aux cliniciens, aux responsables des soins et aux administrateurs. Le score de propension à la réadmission à l'hôpital vous permet d'évaluer le rendement global par établissement, par spécialité ou par affection. Vous pouvez ensuite utiliser ce score pour mettre en œuvre des interventions proactives.

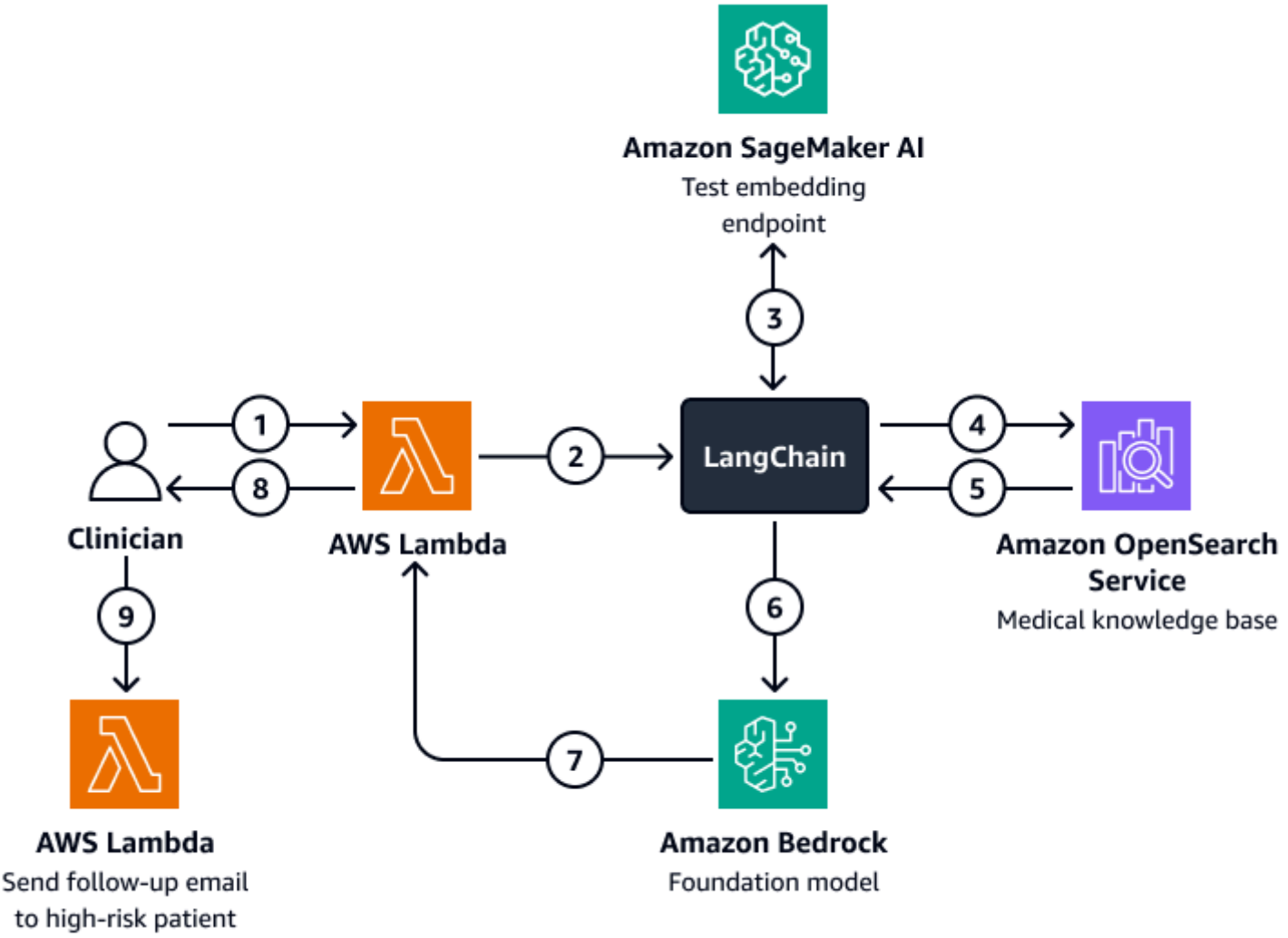
1. Attribuez des poids à chacun des différents facteurs (prédiction des résultats, probabilité d'adhésion, réadmission). Voici des exemples de poids :
   - Poids de prédiction des résultats : 0,4
   - Poids prévisionnel d'adhérence : 0,3
   - Re-admission Poids de probabilité : 0,3
2. Utilisez le calcul suivant pour calculer le score composite :

   `ReadadmissionPropensityScore` = (`OutcomeScore` × `OutcomeWeight`) + (`AdherenceScore` × `AdherenceWeight`) + (`ReadmissionLikelihoodScore` × `ReadmissionLikelihoodWeight`)
3. Assurez-vous que tous les scores individuels sont sur la même échelle, par exemple de 0 à 1.
4. Définissez les seuils d'action. Par exemple, les scores supérieurs à 0,7 déclenchent des alertes.

Sur la base des analyses ci-dessus et du score de propension à la réadmission d'un patient, les cliniciens ou les responsables de soins peuvent configurer des alertes pour surveiller leurs patients

individuels sur la base du score calculé. S'il est supérieur à un seuil prédéfini, ils sont avertis lorsque ce seuil est atteint. Cela permet aux responsables des soins d'être proactifs plutôt que réactifs lorsqu'ils élaborent des plans de soins pour la sortie de leurs patients. Enregistrez les résultats, le comportement et les scores de propension des patients sous forme indexée dans une base de données vectorielle Amazon OpenSearch Service afin que les responsables de soins puissent les récupérer facilement à l'aide d'un agent d'intelligence artificielle conversationnel.

Le schéma suivant montre le flux de travail d'un agent d'intelligence artificielle conversationnel qu'un clinicien ou un responsable de soins peut utiliser pour obtenir des informations sur les résultats des patients, le comportement attendu et la propension à la réadmission. Les utilisateurs peuvent récupérer des informations au niveau du patient, du département ou de l'hôpital. L'agent AI récupère ces informations, qui sont stockées sous forme indexée dans une base de données vectorielle Amazon OpenSearch Service. L'agent utilise la requête pour récupérer les données pertinentes et fournit des réponses personnalisées, y compris des suggestions d'actions pour les patients présentant un risque élevé de réadmission. En fonction du niveau de risque, l'agent peut également configurer des rappels pour les patients et les soignants.



Le schéma suivant illustre le flux de travail suivant :

1. Le clinicien pose une question à un agent d'intelligence artificielle conversationnel qui héberge une AWS Lambda fonction.
2. La fonction Lambda lance un agent. LangChain
3. L'LangChainagent envoie la question de l'utilisateur à un point de terminaison d'intégration de texte Amazon SageMaker AI. Le point de terminaison intègre la question.
4. L'LangChainagent transmet la question intégrée à une base de connaissances médicales sur Amazon OpenSearch Service.
5. Amazon OpenSearch Service renvoie à l'LangChainagent les informations spécifiques les plus pertinentes pour la requête de l'utilisateur.
6. Les LangChain agents envoient la requête et le contexte extrait de la base de connaissances à un modèle de base Amazon Bedrock.

7. Le modèle Amazon Bedrock Foundation génère une réponse et l'envoie à la fonction Lambda.
8. La fonction Lambda renvoie la réponse au clinicien.
9. Le clinicien lance une fonction Lambda qui envoie un e-mail de suivi à un patient présentant un risque élevé de réadmission.

Alignement sur le AWS Well-Architected Framework

L'architecture permettant de suivre le comportement des patients et de prévoir les taux de réadmission à l'hôpital intègre Services AWS des graphiques de connaissances médicales et des LLM afin d'améliorer les résultats des soins de santé tout en s'alignant sur les six piliers du cadre :AWS Well-Architected

- Excellence opérationnelle — La solution est un système découplé et automatisé qui utilise Amazon Bedrock et AWS Lambda fournit des alertes en temps réel.
- Sécurité — Cette solution est conçue pour se conformer aux réglementations sanitaires, telles que la loi HIPAA. Vous pouvez également implémenter le chiffrement, un contrôle d'accès précis et des garde-corps Amazon Bedrock pour protéger les données des patients.
- Fiabilité — L'architecture utilise un système sans serveur tolérant aux pannes. Services AWS
- Efficacité des performances — Amazon OpenSearch Service et les LLM affinés peuvent fournir des prévisions rapides et précises.
- Optimisation des coûts — Les technologies sans serveur et les modèles de paiement par inférence aident à minimiser les coûts. Bien qu'un LLM affiné puisse entraîner des frais supplémentaires, le modèle utilise une approche RAG qui réduit les données et le temps de calcul nécessaires au processus de réglage précis.
- Durabilité — L'architecture minimise la consommation de ressources grâce à l'utilisation d'une infrastructure sans serveur. Il soutient également des opérations de santé efficaces et évolutives.

Cas d'utilisation : gestion et renforcement des compétences de votre personnel de santé

La mise en œuvre de stratégies de transformation des talents et de renforcement des compétences aide les travailleurs à rester aptes à utiliser les nouvelles technologies et pratiques dans les services médicaux et de santé. Les initiatives proactives de renforcement des compétences permettent aux professionnels de santé de fournir des soins de haute qualité aux patients, d'optimiser l'efficacité opérationnelle et de respecter les normes réglementaires. De plus, la transformation des talents favorise une culture d'apprentissage continu. Cela est essentiel pour s'adapter à l'évolution du paysage des soins de santé et relever les nouveaux défis de santé publique. Les approches de formation traditionnelles, telles que la formation en classe et les modules d'apprentissage statiques, offrent un contenu uniforme à un large public. Ils ne disposent souvent pas de parcours d'apprentissage personnalisés, essentiels pour répondre aux besoins spécifiques et aux niveaux de compétence de chaque praticien. Cette stratégie universelle peut entraîner un désengagement et une rétention des connaissances sous-optimale.

Par conséquent, les établissements de santé doivent adopter des solutions innovantes, évolutives et axées sur la technologie qui peuvent déterminer l'écart entre la situation actuelle et la situation future de chacun de leurs employés. Ces solutions devraient recommander des parcours d'apprentissage hyperpersonnalisés et le bon ensemble de contenus d'apprentissage. Cela prépare efficacement le personnel au futur des soins de santé.

Dans le secteur de la santé, vous pouvez appliquer l'IA générative pour vous aider à comprendre et à améliorer les compétences de votre personnel. Grâce à la connexion de grands modèles linguistiques (LLM) et de récupérateurs avancés, les organisations peuvent comprendre les compétences dont elles disposent actuellement et identifier les compétences clés qui pourraient être nécessaires à l'avenir. Ces informations vous aident à combler le fossé en recrutant de nouveaux travailleurs et en améliorant les compétences de la main-d'œuvre actuelle. À l'aide d'Amazon Bedrock et des graphes de connaissances, les établissements de santé peuvent développer des applications spécifiques à un domaine qui facilitent l'apprentissage continu et le développement des compétences.

Les connaissances fournies par cette solution vous aident à gérer efficacement les talents, à optimiser les performances du personnel, à favoriser le succès de l'organisation, à identifier les compétences existantes et à élaborer une stratégie de gestion des talents. Cette solution peut vous aider à effectuer ces tâches en quelques semaines au lieu de plusieurs mois.

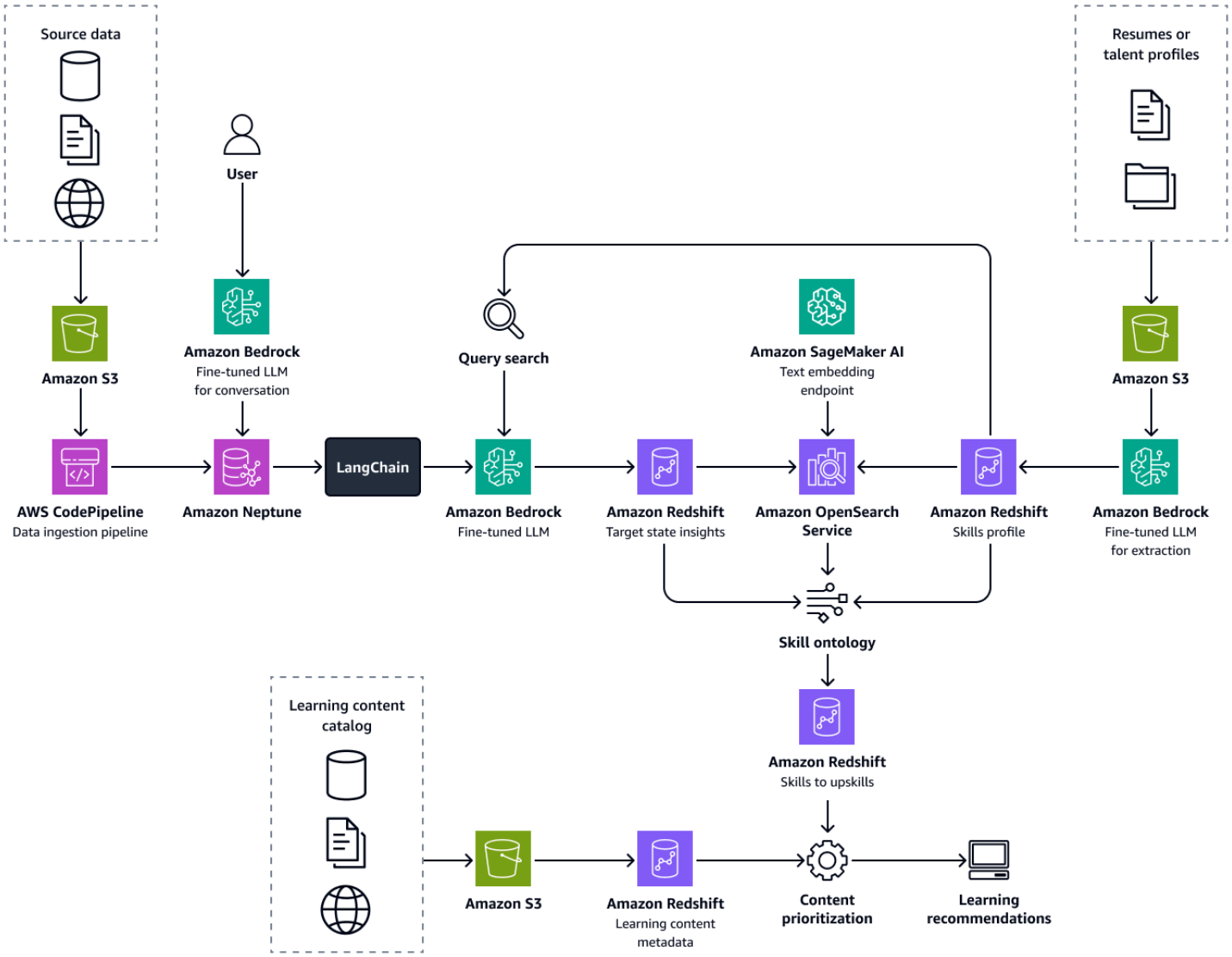
Présentation de la solution

Cette solution est un cadre de transformation des talents du secteur de la santé qui comprend les éléments suivants :

- Analyseur de CV intelligent — Ce composant peut lire le CV d'un candidat et extraire avec précision les informations relatives au candidat, y compris ses compétences. Solution intelligente d'extraction d'informations conçue à l'aide du modèle Llama 2 affiné d'Amazon Bedrock sur un ensemble de données de formation exclusif couvrant les CV et les profils de talents de plus de 19 secteurs d'activité. Ce LLM-based processus permet d'économiser des centaines d'heures en automatisant le processus de révision manuelle des CV et en associant les meilleurs candidats aux postes à pourvoir.
- Graphe de connaissances : graphe de connaissances basé sur Amazon Neptune, un référentiel unifié d'informations sur les talents, y compris la taxonomie des rôles et des compétences de l'organisation et du secteur, qui capture la sémantique des talents du secteur de la santé à l'aide de définitions des compétences, des rôles et de leurs propriétés, des relations et des contraintes logiques.
- Ontologie des compétences — La découverte de la proximité des compétences entre les compétences des candidats et les compétences de l'état actuel ou futur idéales (récupérées à l'aide d'un graphe de connaissances) est réalisée grâce à des algorithmes d'ontologie qui mesurent la similitude sémantique entre les compétences des candidats et les compétences de l'état cible.
- Parcours et contenu d'apprentissage — Ce composant est un moteur de recommandation pédagogique capable de recommander le bon contenu d'apprentissage à partir d'un catalogue de supports d'apprentissage de n'importe quel fournisseur en fonction des lacunes de compétences identifiées. Identifier les parcours de perfectionnement les plus optimaux pour chaque candidat en analysant les lacunes en matière de compétences et en recommandant des contenus d'apprentissage prioritaires, afin de permettre un développement professionnel continu et fluide pour chaque candidat pendant la transition vers un nouveau poste.

Cette solution automatisée basée sur le cloud est alimentée par des services d'apprentissage automatique, des LLM, des graphes de connaissances et la génération augmentée de récupération (RAG). Il peut évoluer pour traiter des dizaines ou des milliers de CV en un minimum de temps, créer des profils de candidats instantanés, identifier les lacunes dans leur état actuel ou futur potentiel, puis recommander efficacement le contenu de formation approprié pour combler ces lacunes.

L'image suivante montre le flux de bout en bout du framework. La solution repose sur des LLM affinés dans Amazon Bedrock. Ces LLM extraient des données de la base de connaissances sur les talents du secteur de la santé d'Amazon Neptune. Data-driven les algorithmes recommandent des parcours d'apprentissage optimaux pour chaque candidat.



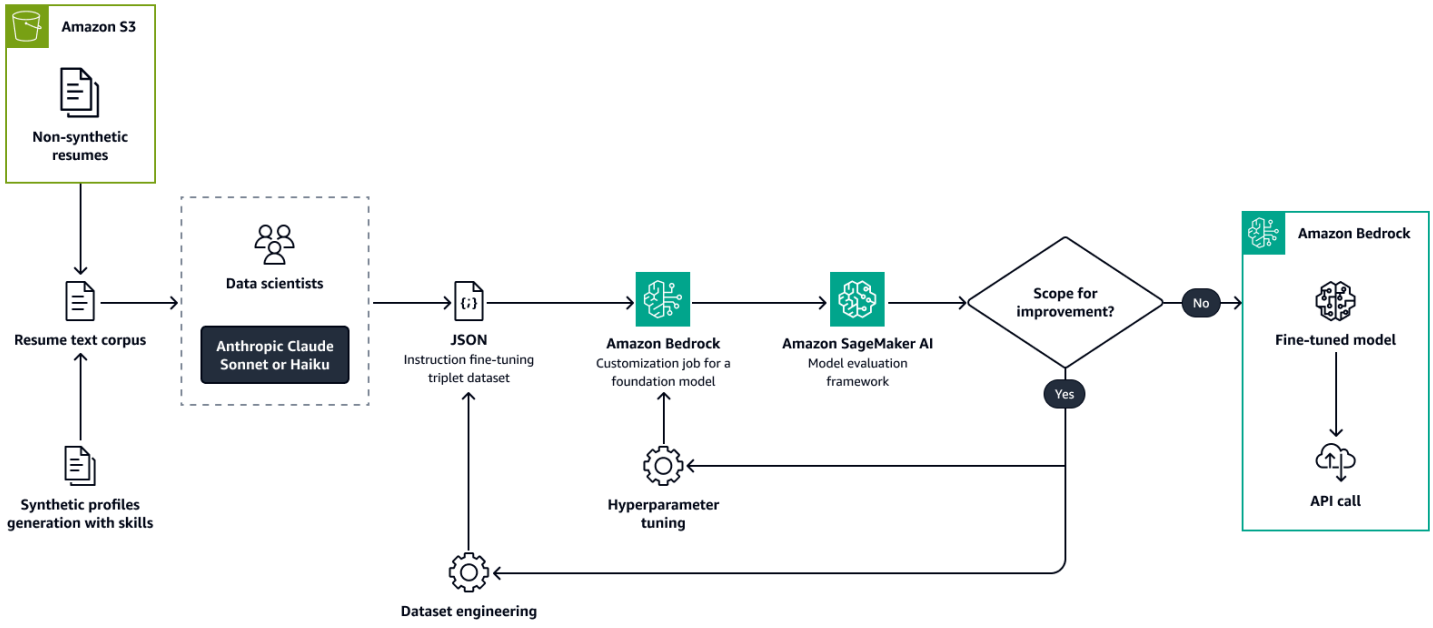
La création de cette solution comprend les étapes suivantes :

- Étape 1 : Extraction des informations sur les talents et établissement d'un profil de compétences
- Étape 2 : Déterminer la pertinence du rôle par rapport aux compétences à partir d'un graphe de connaissances
- Étape 3 : Identifier les lacunes en matière de compétences et recommander une formation

Étape 1 : Extraction des informations sur les talents et établissement d'un profil de compétences

Tout d'abord, vous peaufinez un grand modèle de langage, tel que Llama 2, dans Amazon Bedrock à l'aide d'un ensemble de données personnalisé. Cela adapte le LLM au cas d'utilisation. Au cours de la formation, vous extrayez de manière précise et cohérente les principaux attributs des talents des CV des candidats ou de profils de talents similaires. Ces attributs de talent incluent les compétences, le titre du poste actuel, les titres d'expérience avec dates, la formation et les certifications. Pour plus d'informations, consultez la section Personnaliser votre modèle afin d'améliorer ses performances en fonction de votre cas d'utilisation dans la documentation Amazon Bedrock.

L'image suivante montre le processus permettant d'affiner un modèle d'analyse de CV à l'aide d'Amazon Bedrock. Les CV réels et synthétiques sont transmis à un LLM afin d'en extraire des informations clés. Un groupe de data scientists valide les informations extraites par rapport au texte brut d'origine. Les informations extraites sont ensuite concaténées à l'aide d'une chaîne de pensée et du texte original pour obtenir un ensemble de données d'apprentissage à affiner. Ce jeu de données est ensuite transmis à une tâche de personnalisation Amazon Bedrock, qui affine le modèle. Une tâche par lots Amazon SageMaker AI exécute un cadre d'évaluation de modèle qui évalue le modèle affiné. Si le modèle doit être amélioré, la tâche est réexécutée avec davantage de données ou différents hyperparamètres. Une fois que l'évaluation répond aux normes, vous hébergez le modèle personnalisé via le débit provisionné d'Amazon Bedrock.



Étape 2 : Déterminer la pertinence du rôle par rapport aux compétences à partir d'un graphe de connaissances

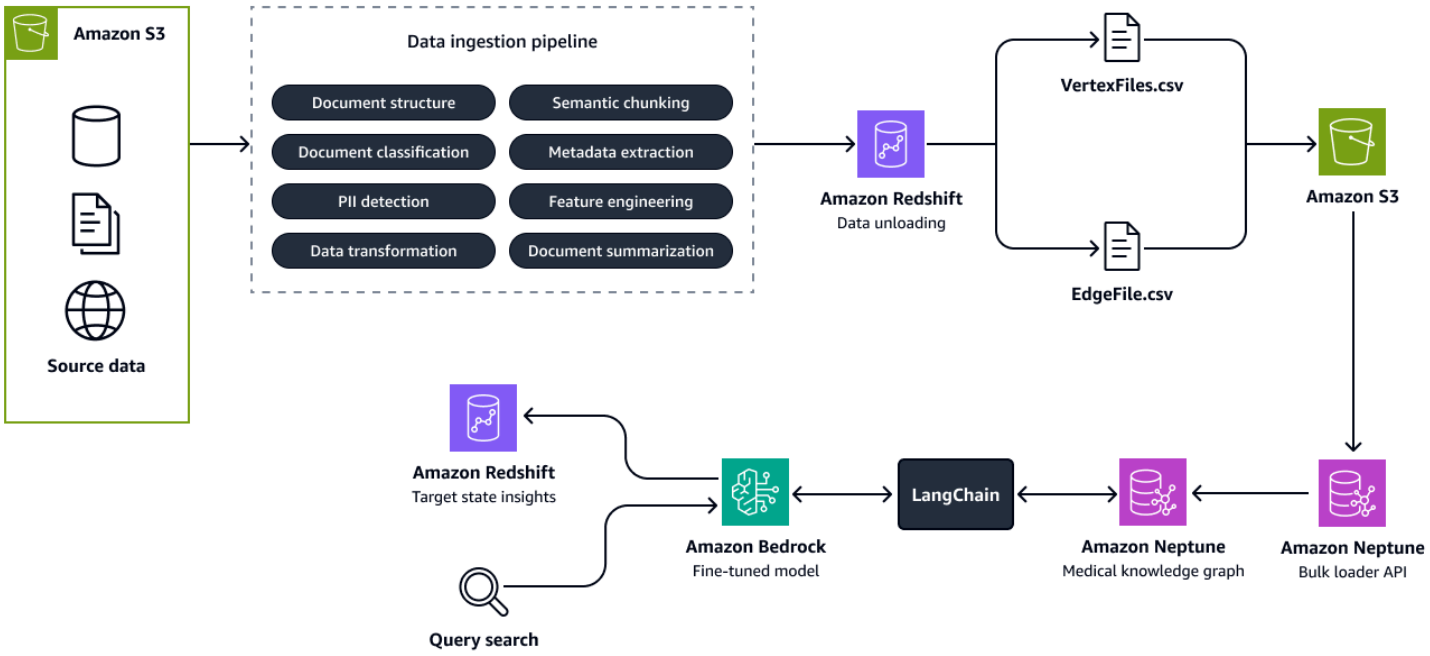
Ensuite, vous créez un graphe de connaissances qui résume les compétences et la taxonomie des rôles de votre organisation et des autres organisations du secteur de la santé. Cette base de connaissances enrichie provient de données agrégées sur les talents et les organisations dans Amazon Redshift. Vous pouvez recueillir des données sur les talents auprès de divers fournisseurs de données sur le marché du travail et de sources de données structurées et non structurées spécifiques à l'organisation, telles que les systèmes de planification des ressources d'entreprise (ERP), un système d'information sur les ressources humaines (HRIS), les CV des employés, les descriptions de poste et les documents relatifs à l'architecture des talents.

Créez le graphe de connaissances sur Amazon Neptune. Les nœuds représentent les compétences et les rôles, tandis que les arêtes représentent les relations entre eux. Enrichissez ce graphique avec des métadonnées pour inclure des détails tels que le nom de l'organisation, le secteur d'activité, la famille d'emplois, le type de compétence, le type de rôle et les tags sectoriels.

Ensuite, vous développez une application Graph Retrieval Augmented Generation (Graph RAG). Graph RAG est une approche RAG qui extrait les données d'une base de données de graphes. Les composants de l'application Graph RAG sont les suivants :

- Intégration avec un LLM dans Amazon Bedrock — L'application utilise un LLM dans Amazon Bedrock pour comprendre le langage naturel et générer des requêtes. Les utilisateurs peuvent interagir avec le système en utilisant le langage naturel. Cela le rend accessible aux parties prenantes non techniques.
- Orchestration et récupération d'informations — Utilisez LlamaIndexnos LangChainorchestrateurs pour faciliter l'intégration entre le LLM et le graphe de connaissances Neptune. Ils gèrent le processus de conversion des requêtes en langage naturel en requêtes OpenCypher. Ils exécutent ensuite les requêtes sur le graphe de connaissances. Utilisez une ingénierie rapide pour informer le LLM des meilleures pratiques pour créer des requêtes OpenCypher. Cela permet d'optimiser les requêtes pour récupérer le sous-graphe pertinent, qui contient toutes les entités et relations pertinentes concernant les rôles et les compétences demandés.
- Génération d'informations — Le LLM d'Amazon Bedrock traite les données graphiques récupérées. Il génère des informations détaillées sur l'état actuel et projette les états futurs pour le rôle demandé et les compétences associées.

L'image suivante montre les étapes à suivre pour créer un graphe de connaissances à partir des données sources. Vous transmettez les données sources structurées et non structurées au pipeline d'ingestion de données. Le pipeline extrait et transforme les informations en une formation de chargement groupé CSV compatible avec Amazon Neptune. L'API Bulk Loader télécharge les fichiers CSV stockés dans un compartiment Amazon S3 vers le graphe de connaissances Neptune. Pour les requêtes des utilisateurs relatives à l'état futur des talents, aux rôles pertinents ou aux compétences, le LLM affiné d'Amazon Bedrock interagit avec le graphe de connaissances via un orchestrateur. LangChain L'orchestrateur extrait le contexte pertinent à partir du graphe de connaissances et transmet les réponses au tableau d'informations d'Amazon Redshift. L'LangChainorchestrateur, comme GraphQaChain, convertit la requête en langage naturel de l'utilisateur en une requête OpenCypher afin d'interroger le graphe de connaissances. Le modèle affiné d'Amazon Bedrock génère une réponse basée sur le contexte extrait.



Étape 3 : Identifier les lacunes en matière de compétences et recommander une formation

Au cours de cette étape, vous calculez avec précision la proximité entre l'état actuel d'un professionnel de santé et les futurs rôles gouvernementaux potentiels. Pour ce faire, vous effectuez une analyse d'affinité des compétences en comparant les compétences de l'individu avec le rôle professionnel. Dans une base de données vectorielle Amazon OpenSearch Service, vous stockez les informations de taxonomie des compétences et les métadonnées des compétences, telles que

la description des compétences, le type de compétence et les groupes de compétences. Utilisez un modèle d'intégration Amazon Bedrock, tel que les modèles Amazon Titan Text Embeddings, pour intégrer la compétence clé identifiée dans des vecteurs. Grâce à une recherche vectorielle, vous pouvez récupérer les descriptions des compétences de l'état actuel et des compétences de l'état cible et effectuer une analyse ontologique. L'analyse fournit des scores de proximité entre les paires de compétences de l'État actuel et de l'État cible. Pour chaque paire, vous utilisez les scores ontologiques calculés pour identifier les écarts dans les affinités de compétences. Ensuite, vous recommandez la voie optimale pour le renforcement des compétences, que le candidat peut envisager lors des transitions de rôle.

Pour chaque rôle, la recommandation du contenu d'apprentissage approprié à des fins de perfectionnement ou de requalification implique une approche systématique qui commence par la création d'un catalogue complet de contenus d'apprentissage. Ce catalogue, que vous stockez dans une base de données Amazon Redshift, regroupe le contenu de différents fournisseurs et inclut des métadonnées, telles que la durée du contenu, le niveau de difficulté et le mode d'apprentissage. L'étape suivante consiste à extraire les compétences clés proposées par chaque élément de contenu, puis à les associer aux compétences individuelles requises pour le rôle cible. Vous réalisez cette cartographie en analysant la couverture fournie par le contenu grâce à une analyse de proximité des compétences. Cette analyse évalue dans quelle mesure les compétences enseignées par le contenu correspondent aux compétences souhaitées pour le poste. Les métadonnées jouent un rôle essentiel dans la sélection du contenu le plus approprié pour chaque compétence, en veillant à ce que les apprenants reçoivent des recommandations personnalisées adaptées à leurs besoins d'apprentissage. Utilisez les LLM dans Amazon Bedrock pour extraire des compétences des métadonnées du contenu, réaliser l'ingénierie des fonctionnalités et valider les recommandations de contenu. Cela améliore la précision et la pertinence du processus de mise à niveau ou de requalification.

Alignement sur le AWS Well-Architected Framework

La solution s'aligne sur les six piliers du AWS Well-Architected cadre :

- Excellence opérationnelle — Un pipeline modulaire et automatisé améliore l'excellence opérationnelle. Les composants clés du pipeline sont découplés et automatisés, ce qui permet d'accélérer les mises à jour des modèles et de faciliter la surveillance. En outre, les pipelines de formation automatisés permettent de publier plus rapidement des modèles affinés.
- Sécurité — Cette solution traite les informations sensibles et personnellement identifiables (PII), telles que les données figurant dans les CV et les profils de talents. Dans Gestion des identités et

des accès AWS (IAM), mettez en œuvre des politiques de contrôle d'accès précises et assurez-vous que seul le personnel autorisé a accès à ces données.

- Fiabilité — La solution utilise des Services AWS outils tels que Neptune, Amazon Bedrock et OpenSearch Service, qui offrent une tolérance aux pannes, une haute disponibilité et un accès ininterrompu aux informations, même en cas de forte demande.
- Efficacité des performances — Les Fine-tuned LLM des bases de données vectorielles Amazon Bedrock et OpenSearch Service sont conçus pour traiter rapidement et avec précision de grands ensembles de données afin de fournir des recommandations d'apprentissage personnalisées en temps opportun.
- Optimisation des coûts — Cette solution utilise une approche RAG, qui réduit le besoin de pré-entraînement continu des modèles. Au lieu de peaufiner l'ensemble du modèle à plusieurs reprises, le système ne peaufine que des processus spécifiques, tels que l'extraction d'informations à partir de CV et la structuration des résultats. Cela se traduit par des économies de coûts importantes. En minimisant la fréquence et l'ampleur de la formation des modèles gourmands en ressources et en utilisant des services cloud payants à l'utilisation, les établissements de santé peuvent optimiser leurs coûts opérationnels tout en maintenant des performances élevées.
- Durabilité — Cette solution utilise des services évolutifs et natifs du cloud qui allouent les ressources informatiques de manière dynamique. Cela permet de réduire la consommation d'énergie et l'impact environnemental tout en soutenant les initiatives de transformation des talents à grande échelle et gourmandes en données.

Développement et orchestration de solutions d'IA générative pour le secteur de la santé

Pour créer les solutions présentées dans ce guide, vous devez créer une architecture RAG qui utilise des LLM affinés pour fournir aux prestataires de soins de santé des données augmentées sur les patients, des informations cliniques et diagnostiques et des résultats prédits pour les patients. Cela nécessite l'intégration de plusieurs Services AWS outils pour créer un flux de travail cohérent et efficace. Cette section aborde les points suivants :

- Amazon Q Developer— Utilisez Amazon Q Developer pour résoudre les questions d'ingénierie et les erreurs de code pendant le processus de développement.
- Multi-retriever Design RAG— Concevez et implémentez des solutions RAG qui utilisent plusieurs récupérateurs pour trouver le contexte médical approprié à la question de l'utilisateur.
- ReAct agents— Implémentez des agents qui combinent raisonnement et action dynamique.

Amazon Q Developer

Lors de la création d'une solution d'IA générative, il peut être difficile de créer des agents d'IA et de connecter des services clés. Cependant, Amazon Q Developer aide les scientifiques des données et les ingénieurs en IA en leur donnant accès à un assistant d'IA générative avancé. Amazon Q peut répondre rapidement et précisément aux questions des utilisateurs et aux erreurs de code, ce qui peut vous aider à optimiser le processus de développement du LLM. Amazon Q offre des avantages considérables aux développeurs qui créent des applications utilisant les modèles de base d'Amazon Bedrock. Il peut rationaliser les flux de travail et améliorer la qualité du code. Il automatise la génération de scripts Python et de configurations d'infrastructure sous forme de code (IaC), réduisant ainsi considérablement le temps et les efforts de développement. Grâce à des fonctionnalités de refactoring avancées, Amazon Q peut améliorer les performances du code, identifier les failles de sécurité et s'assurer que les développeurs respectent les meilleures pratiques. En outre, il facilite l'apprentissage et l'adoption par les débutants en fournissant des suggestions et des explications contextuelles, ce qui rend les tâches de codage complexes plus accessibles et plus efficaces.

Multi-retriever Design RAG

Dans une application d'IA générative, un pipeline RAG multi-récupérateurs peut récupérer efficacement des informations à partir de plusieurs sources de données pour aider les prestataires de soins de santé et les cliniciens à répondre aux questions médicales. Ce pipeline utilise différents types de récupérateurs pour extraire les données pertinentes de différentes bases de connaissances. Chaque récupérateur est spécialisé dans la récupération d'un type particulier d'informations, telles que les antécédents des patients, les informations diagnostiques, les notes cliniques ou le contenu de recherches médicales et de textes universitaires.

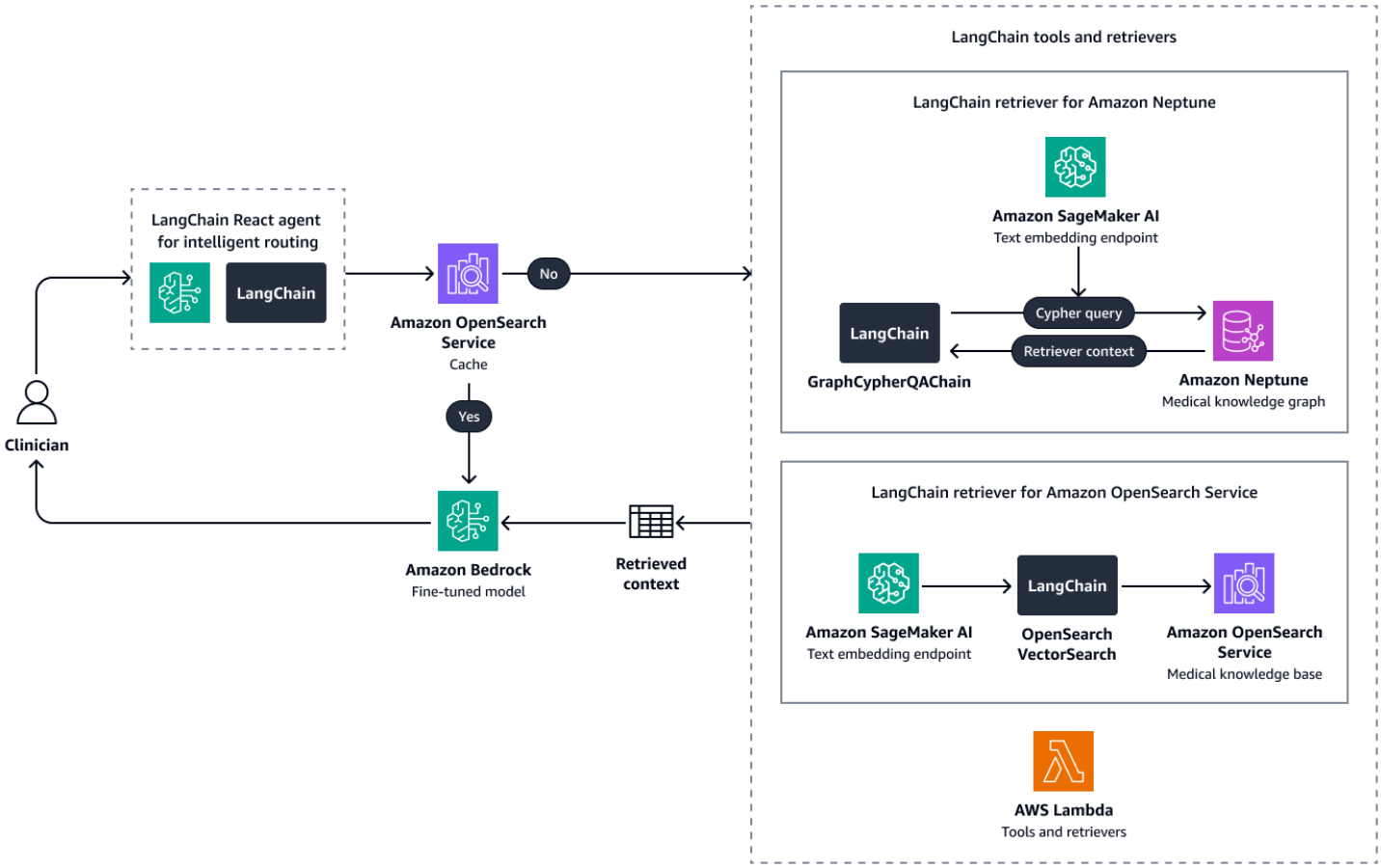
Utilisez la nature des données et les exigences spécifiques de l'application pour déterminer quelle est la base de connaissances backend adaptée à votre cas d'utilisation. Une base de données vectorielle Amazon OpenSearch Service convient parfaitement à de grands volumes de données de santé non structurées ou semi-structurées, notamment des résumés d'évaluation diagnostique par image, des résumés de sortie, des rapports cliniques, des recherches médicales et des textes universitaires. D'autre part, un service de base de données graphique, tel qu'Amazon Neptune, peut être idéal pour les cas d'utilisation dans le secteur de la santé qui nécessitent une exploration approfondie des relations temporelles entre les entités, telles que le patient, les antécédents du patient, le prestataire de soins de santé, les médicaments, les symptômes et les traitements.

Un élément essentiel de ce pipeline est la prédiction de l'intention des requêtes des utilisateurs. Cela garantit que le système achemine la requête vers la bonne chaîne de récupération. Par exemple, si un clinicien pose des questions sur les antécédents thérapeutiques d'un patient, ses symptômes, son interaction avec l'hôpital, la probabilité d'une réadmission à l'hôpital ou les résultats potentiels du patient, le module de prédiction de l'intention des requêtes identifie cette intention. Il dirige la demande vers la chaîne de recherche qui peut récupérer les dossiers des patients ou les données chronologiques des traitements à partir du graphe des connaissances médicales. Par ailleurs, si la question porte sur la découverte d'une maladie, des évaluations diagnostiques spécifiques ou des détails de procédures cliniques spécifiques tirés de manuels universitaires, la requête est acheminée vers la chaîne de recherche qui peut récupérer ces informations dans la OpenSearch base de données des vecteurs de service. Vous pouvez utiliser la fonctionnalité d'appel d'outils LangChain pour associer un outil personnalisé au programme Amazon Bedrock LLM qui permet de classer une question d'utilisateur selon des intentions prédéfinies.

Ce système RAG multi-retriever inclut des LangChain agents conçus pour gérer l'accès à la base de connaissances spécifique. Vous pouvez l'utiliser LangChain pour orchestrer l'interaction entre l'Amazon Bedrock LLM, les différents récupérateurs et les outils. LangChaininclut une classe

d'appel d'outils qui vous aide à créer des outils personnalisés, tels qu'un classificateur d'intention, un récupérateur pour Neptune, un récupérateur OpenSearch pour Service ou tout autre outil pouvant être développé pour classer les intentions de l'utilisateur et accéder aux données d'une base de connaissances spécifique dans un format structuré. Vous transmettez ensuite ces outils à la classe pour créer un agent Reasoning and Acting (ReAct). L' ReAct agent traite la question de l'utilisateur, planifie les étapes séquentielles pour y répondre, puis exécute de manière itérative les outils disponibles et traite les réponses des outils pour enfin répondre à la requête de l'utilisateur.

L'image suivante montre le fonctionnement d'un système RAG à plusieurs récupérateurs conçu pour une récupération efficace des connaissances et une résolution intelligente des requêtes. Un LangChain ReAct agent analyse l'intention de l'utilisateur, formule un plan d'exécution structuré et sélectionne les outils de récupération les plus pertinents. Le système interroge un cache de questions précédentes et recherche des requêtes similaires en fonction d'attributs clés, tels que l'identifiant du patient, son état de santé et la date de la visite. Si une question très similaire est trouvée, la réponse correspondante est récupérée directement. Dans le cas contraire, l'agent exécute le récupérateur approprié. Pour récupérer des informations centrées sur le patient, telles que l'historique du traitement, les symptômes, les interactions avec l'hôpital ou la probabilité de réadmission, le système utilise un extracteur de graphes. Pour les évaluations diagnostiques, les procédures cliniques et les résultats médicaux structurés, l'agent utilise un récupérateur de base de données vectorielles. Dans les scénarios qui nécessitent une combinaison de connaissances contextuelles provenant des deux magasins de données, pour générer une réponse complète, le système utilise une stratégie de récupération hybride qui intègre les résultats du graphe de connaissances et de la base de données vectorielle.



ReAct agents

Les agents Reasoning and Acting (ReAct) sont conçus pour les applications RAG à multiples facettes. Ces agents fournissent une puissante combinaison de raisonnement et d'action dynamique, en particulier pour les applications complexes qui impliquent des flux de travail logiques de récupération d'informations étape par étape. Pour plus d'informations, voir ReAct: Synergie du raisonnement et de l'action dans les modèles linguistiques.

Dans les contextes médicaux et de soins de santé, les demandes d'un clinicien ou d'un médecin comportent souvent de multiples facettes. Par exemple, un clinicien peut demander « Quels traitements ont été administrés à des patients similaires souffrant à la fois d'hypertension et de diabète de type 2 ? » Après avoir identifié l'intention de l'utilisateur, qui est de récupérer les traitements pour l'hypertension et le diabète de type 2, l'agent d'intelligence artificielle doit diviser cette requête en sous-tâches, puis choisir la stratégie de récupération la plus efficace. Dans ce cas, l'agent d'IA doit identifier les nœuds les plus pertinents (tels que l'âge, le sexe, les affections, les traitements et les médicaments du patient), puis interroger le graphique pour ces entités,

leurs attributs et leurs relations. ReAct les agents sont très utiles car ils combinent la capacité de raisonnement (inférence logique) d'un LLM avec une action (interrogation ou interaction avec des ressources externes ou des bases de connaissances).

Pour répondre à la question de l'utilisateur « Quels traitements ont été administrés à des patients similaires souffrant à la fois d'hypertension et de diabète de type 2 ? » , l'exemple suivant illustre le fonctionnement d'un ReAct agent :

1. Raisonnement de l' ReAct agent — L'agent en déduit que la question implique de récupérer des informations sur des affections (diabète et hypertension). Il tient compte de l'âge du patient, des traitements, des médicaments et de la période à analyser.
2. Action de l'agent — L'agent utilise OpenCypher pour interroger le graphe de connaissances sur les traitements spécifiques au diabète de type 2 et à l'hypertension. Il permet également de récupérer les médicaments administrés, les dates des visites à l'hôpital, les effets secondaires des médicaments, les résultats connus des patients et les données de référence croisées pour des patients similaires (tels que des patients du même sexe et du même âge).
3. Observation de l'agent — À partir du graphe de connaissances, l'agent extrait les données tabulaires les plus récentes des six derniers mois sur les traitements administrés à des patients atteints à la fois d'hypertension et de diabète de type 2.
4. Raisonnement de l'agent — Pour classer les résultats à partir des dossiers récupérés, l'agent identifie des attributs importants, tels que la récence, les effets secondaires des médicaments ou les résultats connus pour les patients.
5. Action de l'agent : l'agent reclasse les enregistrements en fonction des attributs identifiés et de la logique prédéfinie transmise par le biais de l'invite du système.
6. Génération de réponses — Le LLM d'Amazon Bedrock génère une réponse basée sur le contexte préparé par l' ReAct agent.

Évaluation des solutions d'IA générative pour le secteur de la santé

L'évaluation des solutions d'IA pour le secteur de la santé que vous créez est essentielle pour garantir leur efficacité, leur fiabilité et leur évolutivité dans des environnements médicaux réels. Utilisez une approche systématique pour évaluer les performances de chaque composant de la solution. Vous trouverez ci-dessous un résumé des méthodologies et des mesures que vous pouvez utiliser pour évaluer votre solution.

Rubriques

- Évaluation de l'extraction d'informations
- Évaluation des solutions RAG avec plusieurs récupérateurs
- Évaluation d'une solution à l'aide d'un LLM

Évaluation de l'extraction d'informations

Évaluez les performances des solutions d'extraction d'informations, telles que l'analyseur de CV intelligent et l'extracteur d'entités personnalisé. Vous pouvez mesurer l'alignement des réponses de ces solutions à l'aide d'un jeu de données de test. Si vous ne disposez pas d'un ensemble de données couvrant les profils de talents polyvalents du secteur de la santé et les dossiers médicaux des patients, vous pouvez créer un ensemble de données de tests personnalisé en utilisant la capacité de raisonnement d'un LLM. Par exemple, vous pouvez utiliser un modèle à grands paramètres, tel que des Anthropic Claude modèles, pour générer un ensemble de données de test.

Voici trois indicateurs clés que vous pouvez utiliser pour évaluer les modèles d'extraction d'informations :

- Exactitude et exhaustivité — Ces mesures évaluent la mesure dans laquelle le résultat a capturé les informations correctes et complètes présentes dans les données de vérité sur le terrain. Cela implique de vérifier à la fois l'exactitude des informations extraites et la présence de tous les détails pertinents dans les informations extraites.
- Similarité et pertinence — Ces mesures évaluent les similitudes sémantiques, structurelles et contextuelles entre les données de sortie et les données de base (la similitude) et la mesure dans laquelle la sortie s'aligne sur le contenu, le contexte et l'intention des données de base (la pertinence) et les prend en compte.

- Taux de rappel ou de capture ajusté — Ces taux déterminent de manière empirique le nombre de valeurs actuelles des données de vérité sur le terrain qui ont été correctement identifiées par le modèle. Le taux doit inclure une pénalisation pour toutes les fausses valeurs extraites par le modèle.
- Score de précision : le score de précision vous aide à déterminer le nombre de faux positifs présents dans les prédictions, par rapport aux vrais positifs. Par exemple, vous pouvez utiliser des mesures de précision pour mesurer l'exactitude de la compétence extraite.

Évaluation des solutions RAG avec plusieurs récupérateurs

Pour évaluer dans quelle mesure le système récupère les informations pertinentes et dans quelle mesure il utilise ces informations pour générer des réponses précises et adaptées au contexte, vous pouvez utiliser les métriques suivantes :

- Pertinence de la réponse — Mesurez la pertinence de la réponse générée, qui utilise le contexte extrait, par rapport à la requête d'origine.
- Précision du contexte — Sur le total des résultats récupérés, évaluez la proportion de documents ou d'extraits extraits pertinents pour la requête. Une précision contextuelle plus élevée indique que le mécanisme de récupération est efficace pour sélectionner les informations pertinentes.
- Fidélité — Évalue avec quelle précision la réponse générée reflète les informations contenues dans le contexte extrait. En d'autres termes, mesurez si la réponse reste fidèle à l'information source.

Évaluation d'une solution à l'aide d'un LLM

Vous pouvez utiliser une technique appelée LLM- as-a-judge pour évaluer les réponses textuelles de votre solution d'IA générative. Cela implique d'utiliser LLMs pour évaluer et évaluer les performances des résultats du modèle. Cette technique utilise les fonctionnalités d'Amazon Bedrock pour évaluer divers attributs, tels que la qualité des réponses, la cohérence, l'adhésion, la précision et l'exhaustivité par rapport aux préférences humaines ou aux données de base. Vous utilisez des techniques chain-of-thought (CoT) et quelques techniques d'incitation pour une évaluation complète. L'invite demande au LLM d'évaluer la réponse générée à l'aide d'une grille de notation, et les quelques échantillons de l'invite illustrent le processus d'évaluation réel. L'invite comprend également des directives à suivre par l'évaluateur LLM. Par exemple, vous pouvez envisager d'utiliser une ou plusieurs des techniques d'évaluation suivantes qui utilisent un LLM pour évaluer les réponses générées :

- Comparaison par paires — Donnez à l'évaluateur LLM une question médicale et plusieurs réponses générées par différentes versions itératives des systèmes RAG que vous avez créés. Demandez à l'évaluateur LLM de déterminer la meilleure réponse en fonction de la qualité de la réponse, de la cohérence et du respect de la question initiale.
- Notation à réponse unique : cette technique convient parfaitement aux cas d'utilisation dans lesquels vous devez évaluer l'exactitude de la catégorisation, comme la classification des résultats des patients, la catégorisation du comportement des patients, la probabilité de réadmission des patients et la catégorisation des risques. Utilisez l'évaluateur LLM pour analyser la catégorisation ou la classification individuelle de manière isolée, et évaluez le raisonnement qu'il a fourni par rapport aux données de base.
- Notation guidée par référence — Fournissez à l'évaluateur du LLM une série de questions médicales nécessitant des réponses descriptives. Créez des exemples de réponses à ces questions, tels que des réponses de référence ou des réponses idéales. Demandez à l'évaluateur LLM de comparer la réponse générée par le LLM aux réponses de référence ou aux réponses idéales, et demandez-lui d'évaluer la réponse générée en termes d'exactitude, d'exhaustivité, de similitude, de pertinence ou d'autres attributs. Cette technique vous permet d'évaluer si les réponses générées correspondent à une réponse standard ou exemplaire bien définie.

Ressources

AWS documentation

- Documentation Amazon Bedrock
- Documentation Amazon Neptune
- Documentation Amazon OpenSearch Service
- Appliquer le framework AWS Well-Architected pour Amazon Neptune (directives prescriptives)AWS
- Bonnes pratiques opérationnelles pour Amazon OpenSearch Service (documentation OpenSearch du service)
- Utilisation d'Amazon Comprehend Medical LLMs et pour les soins de santé et les sciences de la vieAWS (directives prescriptives)

AWS articles de blog

- Créez des applications RAG et d'IA générative basées sur des agents avec le nouveau modèle Amazon Titan Text Premier, disponible sur Amazon Bedrock
- Complétez l'intelligence commerciale en créant un graphe de connaissances à partir d'un entrepôt de données avec Amazon Neptune
- Utilisation de graphes de connaissances pour créer des applications GraphRag avec Amazon Bedrock et Amazon Neptune

Autres ressources

- Intégrer la génération augmentée par extraction à de grands modèles linguistiques en néphrologie : faire progresser les applications pratiques (PubMed Central, National Library of Medicine)
- Présentation de LangChain (LangChain documentation)

Collaborateurs

Conception

- Nitu Nivedita, directeur général, responsable de l'intelligence artificielle, des données et de l'IA, Accenture
- Manoj Appully, fondateur et directeur technique, Cadiem
- Conor Folan, consultant en données et intelligence artificielle, Accenture
- Deepak Krishna AR, consultant en données et intelligence artificielle, Accenture
- Almore Cato, responsable des données et de l'IA, Accenture
- Soonam Kurian, architecte de solutions principal, AWS

Révision

- Sally Lin, directrice principale de la science des données, données et intelligence artificielle, Accenture
- Terry Huang, responsable de la science des données — Données et intelligence artificielle, Accenture
- William Lorenz, architecte des solutions pour les partenaires, AWS

Rédaction technique

- Lilly AbouHarb, rédactrice technique senior, AWS

Historique du document

Le tableau suivant décrit les modifications importantes apportées à ce guide. Pour être averti des mises à jour à venir, abonnez-vous à un fil RSS.

Modification	Description	Date
Publication initiale	—	14 mars 2025

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.