ヘルスケアとライフサイエンスに Amazon Comprehend Medical と LLMs を使用する

AWS 規範ガイダンス

AWS 規範ガイダンス: ヘルスケアとライフサイエンスに Amazon Comprehend Medical と LLMs を使用する

Table of Contents

序章 1
概要: 1
対象者 2
目的 2
技術的なアプローチ 4
Amazon Comprehend Medical の使用 4
機能 5
ユースケース 6
Amazon Comprehend Medical と LLMs の組み合わせ 7
アーキテクチャ 7
ユースケース 9
ベストプラクティス 10
プロンプト - エンジニアリング 11
LLMs の使用 20
LLM のユースケース 20
カスタマイズ 21
LLM の選択 24
LLMs微調整 26
コストと ROI の見積もり 27
戦略の選択 28
データセットの構築 30
ファインチューニング 31
モニタリング 32
アプローチの選択 33
ビジネスの成熟度に関する考慮事項 34
LLMs の評価 36
データのトレーニングとテスト 36
メトリクス 37
よくある質問 39
Amazon Comprehend Medical と LLM のどちらを選択するか 39
Amazon Comprehend Medical の結果を LLM に提供するにはどうすればよいですか? 39
LLMs で Amazon Comprehend Medical を使用する際のベストプラクティスは何ですか? 39
医療ユースケースでは、事前トレーニング済みの医療 LLM を使用するか、一般的な LLM を微調整する必要がありますか? 40

医療 NLP タスクの LLMs のパフォーマンスを評価するにはどうすればよいですか? 40
高複雑度と低複雑度の LLM ソリューションのトレードオフは何ですか? 40
次の手順 41
AWS リソース 41
その他のリソース 42
寄稿者 43
オーサリング 43
レビューアー 43
テクニカルライター 43
ドキュメント履歴 44
用語集 45
..... 45
A 46
B 48
C 50
D 53
E 57
F 60
G 61
H 62
I 64
L 66
M 67
O 71
P 74
Q 77
R 77
S 80
T 84
U 85
V 86
W 86
Z 87
..... lxxxviii

ヘルスケアとライフサイエンスに Amazon Comprehend Medical と LLMs を使用する

アマゾン ウェブ サービス (寄???稿者)

2025 年 12 月 (ドキュメント履歴)

概要:

増え続ける医療データと効率的で正確な処理の必要性により、人工知能と機械学習 (AI/ML) テクノロジーによる自然言語処理 (NLP) の導入が促進されています。事前トレーニング済みの分類子モデルと大規模言語モデル (LLMs) は、臨床的な質問への回答、レポートの要約、インサイトの生成など、さまざまな医療 NLP タスクのための強力なツールとして浮上しています。ただし、医療とライフサイエンスのドメインは、医学用語、ドメイン固有の知識、規制要件が複雑であるため、固有の課題があります。このドメインで事前トレーニング済みの分類子または LLMs を効果的に使用するには、これらのモデルの長所をドメイン固有のリソースや手法と組み合わせる、適切に設計されたアプローチが必要です。

ヘルスケアとライフサイエンスにおける業界のプラクティスは、従来、ルールベースのシステム、手動コーディング、エキスパートによるレビュープロセスに依存してきました。これらのシステムやプロセスは時間がかかり、エラーが発生しやすくなります。Amazon Comprehend Medical や Amazon Bedrock の基盤モデルなどの AI と NLP テクノロジーの統合は、精度と一貫性を向上させながら医療データを処理するための効率的でスケーラブルなソリューションを提供します。

このガイドでは、ヘルスケア業界におけるインテリジェントな自動化のための Amazon Comprehend Medical と LLMs の使用について説明します。医療コーディング、患者情報抽出、レコード要約プロセスを合理化するためのベストプラクティス、課題、実践的なアプローチの概要を説明します。Amazon Comprehend Medical と LLMs の機能を使用することで、医療組織は新しいレベルの運用効率を引き出し、コストを削減し、患者ケアを向上させることができます。

このガイドでは、医療用語の理解、ドメイン固有の LLMs の使用、AI/ML システムの制限への対応など、ヘルスケアドメイン固有の考慮事項について詳しく説明します。ヘルスケア IT マネージャー、アーキテクト、技術リーダーが組織の準備状況を評価し、実装オプションを評価し、自動化を成功させるための適切な および AWS のサービス ツールを使用するための包括的な決定パスを提供します。

このガイドで概説されているガイドラインとベストプラクティスに従うことで、医療組織は AI/ML テクノロジーの能力を活用しながら、医療分野の複雑さに対処できます。このアプローチは、倫理お

よび規制ガイドラインへの準拠をサポートし、医療における AI システムの責任ある使用を促進します。これは、正確でプライベートなインサイトを生成するように設計されています。

対象者

このガイドは、医療データ分析と自動化のための AI を活用した自然言語処理ソリューションを実装したい技術関係者、アーキテクト、技術リーダー、意思決定者を対象としています。

目的

ヘルスケアおよびライフサイエンス組織は、Amazon Comprehend Medical と LLMs を使用することで、複数のビジネス目標を達成できます。これらの結果には通常、運用効率の向上、コストの削減、患者ケアの改善が含まれます。このセクションでは、主要なビジネス目標と、このガイドで説明されている戦略とベストプラクティスを実装することに伴う利点について説明します。

以下は、このガイドのガイドラインとベストプラクティスを実装することで組織が達成できる目標の一部です。

- 開発時間の短縮 – このガイドの最終的な目標は、コストによる開発時間の短縮、技術的負債の削減、POC による潜在的なプロジェクト障害の軽減です。Amazon Comprehend Medical などの主要な AI/ML サービスと、ヘルスケアタスクにおける LLM の使用の利点と制限を理解することで、企業は市場投入までの時間を短縮し、ビジネス目標の達成速度を向上させることができます。
- 医療コーディングタスクを自動化するための情報を抽出する – 患者の訪問後、コーディングスペシャリストとプロバイダーは、主観的、目的、評価、計画 (SOAP) ノートなどの医療テキストからインサイトを抽出できます。これにより、手動によるドキュメント作成作業が減り、プロバイダーが患者のニーズに集中できるようになります。Amazon Comprehend Medical のエンティティ認識機能を LLMs と組み合わせることで、組織は患者記録、臨床ノート、その他の医療データソースから関連する医療情報を抽出できます。これにより、ヒューマンエラーを最小限に抑え、一貫したプラクティスを促進することができます。
- 患者記録と臨床文書の要約 – 患者の履歴、治療計画、医療結果の自動要約により、医療提供者にとって貴重な時間を節約できます。LLMsは、包括的で構造化された臨床ドキュメントの生成に役立ちます。Amazon Comprehend Medical で追加のコンテキストを取得したり、医療ドメイン LLM を使用したり、医療データで LLM を微調整したりできます。これらのアプローチは、正確な概要を提供し、ドキュメントがコンプライアンス要件と標準に準拠していることを確認するのに役立ちます。
- 臨床上の意思決定と患者ケアのサポート – Amazon Comprehend Medical でオントロジーリンクを使用し、LLMs を使用することで、プロバイダーは医療上の質問に答えたり、患者ケアに対処する

ための推奨事項を求めることができます。これにより、医療専門家は、患者の成果を向上させ、医療ミスのリスクを減らすための情報に基づいた意思決定を行うことができます。

ヘルスケアとライフサイエンスのための生成 AI と NLP アプローチ

自然言語処理 (NLP) は、コンピュータが人間の言語を解釈、操作、理解できるようにする機械学習テクノロジーです。ヘルスケアおよびライフサイエンス組織は、患者記録から大量のデータを持っています。NLP ソフトウェアを使用して、このデータを自動的に処理できます。例えば、NLP と生成 AI を組み合わせて、医療コーディングを合理化し、患者情報を抽出し、レコードを要約できます。

実行する NLP タスクによっては、ユースケースに最適なアーキテクチャが異なる場合があります。このガイドでは、ヘルスケアおよびライフサイエンスアプリケーションの以下の生成 AI および NLP オプションについて説明します AWS。

- Amazon Comprehend Medical の使用 – Amazon Comprehend Medical を大規模言語モデル (LLM) と統合せずに個別に使用する方法について説明します。
- Amazon Comprehend Medical と大規模言語モデルの組み合わせ – 検索拡張生成 (RAG) アーキテクチャで Amazon Comprehend Medical を LLM と組み合わせる方法について説明します。
- ヘルスケアとライフサイエンスのユースケースでの大規模言語モデルの使用 – 微調整された LLM または RAG アーキテクチャを使用して、ヘルスケアおよびライフサイエンスアプリケーションに LLM を使用する方法について説明します。

Amazon Comprehend Medical の使用

Amazon Comprehend Medical は、医師のメモ、退床の概要、テスト結果、ケースノートなど、構造化されていない臨床テキストで有用な情報を検出して返 AWS のサービス す です。自然言語処理 (NLP) モデルを使用してエンティティを検出します。エンティティは、病状、薬剤、保護医療情報 (PHI) などの医療情報へのテキスト参照です。

> **Important**
>
> Amazon Comprehend Medical は専門家による医療の助言、診断、治療の代用品ではありません。Amazon Comprehend Medical は、検出されたエンティティの精度に対する信頼度を示す信頼スコアを提供します。ユースケースに適した信頼しきい値を特定し、高い精度を必要とする状況では高い信頼しきい値を使用してください。特定のユースケースでは、適切な訓練を受けたレビュー担当者によって人的に結果を見直し、検証する必要があります。例え

ば、Amazon Comprehend Medical は、訓練を受けた医療専門家による正確さと健全な医療判断の確認後、患者ケアのシナリオでのみ使用してください。

Amazon Comprehend Medical には、、 AWS Command Line Interface (AWS CLI) AWS マネジメントコンソール、または AWS SDKs からアクセスできます。 AWS SDKs は、Java、Python、Ruby、.NET、iOS、Android など、さまざまなプログラミング言語とプラットフォームで使用できます。SDKs を使用して、クライアントアプリケーションから Amazon Comprehend Medical にプログラムでアクセスできます。

このセクションでは、Amazon Comprehend Medical の主な機能について説明します。また、大規模言語モデル (LLM) と比較して、このサービスを使用する利点についても説明します。

Amazon Comprehend Medical の機能

Amazon Comprehend Medical はAPIs を提供しています。これらの APIs は、医療エンティティの認識とエンティティの関係の特定を使用して、医療テキストを取り込み、医療 NLP タスクの結果を提供できます。Amazon Simple Storage Service (Amazon S3) バケットに保存されている単一のファイルでも、複数のファイルのバッチ分析としても分析を実行できます。Amazon Comprehend Medical では、同期エンティティ検出用に次のテキスト分析 API オペレーションを提供しています。

- エンティティの検出 – 解剖学、病状、PHI カテゴリ、手順、時間式などの一般的な医療カテゴリを検出します。
- PHI の検出 – 年齢、日付、名前、類似の個人情報などの特定のエンティティを検出します。

Amazon Comprehend Medical には、臨床文書に対してバッチテキスト分析を実行するために使用できる複数の API オペレーションも含まれています。これらの API オペレーションの使用方法の詳細については、「テキスト分析バッチ APIs」を参照してください。

Amazon Comprehend Medical では、臨床テキスト内のエンティティを検出し、これらのエンティティを RxNorm、ICD-10-CM、および SNOMED CT のナレッジベースなどの標準化された医療オントロジー内の概念にリンクできます。分析は、1 つのファイルで実行することも、Amazon S3 バケットに保存されている大きなドキュメントや複数のファイルのバッチ分析として実行することもできます。Amazon Comprehend Medical には、次のオントロジーリンク API オペレーションが用意されています。

- InferICD10CM – InferICD10CM オペレーションは、潜在的な病状を検出し、2019 年版の国際疾病分類第 10 版リビジョン、臨床変更 (ICD-10-CM) のコードにリンクします。検出された可能性のあ

る病状ごとに、Amazon Comprehend Medical は一致する ICD-10-CM コードと説明をリストします。結果内のリストされた病状には、結果内の一致した概念に対するエンティティの精度に関して Amazon Comprehend Medical が持っている信頼度を示す信頼スコアが含まれます。

- InferRxNorm – InferRxNorm オペレーションは、患者レコードにエンティティとしてリストされている薬剤を識別します。このオペレーションは、エンティティを米国国立医学図書館の RxNorm データベースの概念識別子 (RxCUI) にリンクします。各 RxCUI は、さまざまな強度と処方に対して一意です。結果内のリストされた薬剤には、RxNorm ナレッジベースの概念に一致したエンティティの精度に関して Amazon Comprehend Medical が持っている信頼度を示す信頼スコアが含まれます。Amazon Comprehend Medical は、信頼スコアに基づいて、検出された各薬剤について一致する可能性のある上位の RxCUI を降順にリストします。
- InferSNOMEDCT – InferSNOMEDCT オペレーションは、考えられる医療概念をエンティティとして識別し、Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) の 2021 年 3 月バージョンのコードにリンクします。SNOMED CTは、病状や解剖学のほか、医学的検査、治療、処置など、医療概念の包括的な語彙を提供します。Amazon Comprehend Medical は、一致する概念 ID ごとに、上位 5 つの医療概念を返します。各概念には、信頼スコアと、特性や属性などのコンテキスト情報が含まれます。SNOMED CT の概念 ID を SNOMED CT 多重階層と併用すると、医療事務、レポート、または臨床分析用に患者の臨床データの構造化できます。

詳細については、Amazon Comprehend Medical ドキュメントの「テキスト分析 APIs」と「オントロジーリンク APIs」を参照してください。

Amazon Comprehend Medical のユースケース

スタンドアロンサービスとして、Amazon Comprehend Medical は組織のユースケースに対応する場合があります。Amazon Comprehend Medical は、次のようなタスクを実行できます。

- 患者レコードでの医療コーディングのヘルプ
- 保護された医療情報 (PHI) データを検出する
- 投与量、頻度、フォームなどの属性を含む薬剤の検証

Amazon Comprehend Medical の結果は、ほとんどの医療行為でダイジェスト可能です。ただし、次のような制限がある場合は、代替案を検討する必要があります。

- 異なるエンティティ定義 – 例えば、FREQUENCY薬剤エンティティの定義は異なる場合があります。頻度については、Amazon Comprehend Medical は必要に応じて予測しますが、組織は pro re nata (PRN) という用語を使用する場合があります。

- 結果の圧倒的な量 – たとえば、患者メモには、複数の ICD-10-CM コードにマッピングされる複数の症状とキーワードが頻繁に含まれます。ただし、いくつかのキーワードは診断には適用されません。この場合、プロバイダーは多数の ICD-10-CM エンティティとその信頼スコアを評価する必要があり、手動処理時間が必要になります。
- カスタムエンティティまたは NLP タスク – 例えば、プロバイダーは、必要に応じて痛むなどの PRN 証拠を抽出したい場合があります。これは Amazon Comprehend Medical では利用できないため、別の AI/ML モデルが必要です。NLP タスクが要約、質疑応答、感情分析などのエンティティ認識の範囲外である場合は、別の AI/ML ソリューションが必要です。

Amazon Comprehend Medical と大規模言語モデルの組み合わせ

NEJM AI による 2024 年の研究では、医療コーディングタスクにゼロショットプロンプトで LLM を使用すると、一般的にパフォーマンスが低下することがわかりました。LLM で Amazon Comprehend Medical を使用すると、これらのパフォーマンスの問題を軽減できます。Amazon Comprehend Medical の結果は、NLP タスクを実行している LLM に役立つコンテキストです。例えば、Amazon Comprehend Medical から大規模言語モデルにコンテキストを提供することは、以下に役立ちます。

- Amazon Comprehend Medical からの最初の結果を LLM のコンテキストとして使用して、エンティティ選択の精度を向上させる
- カスタムエンティティの認識、要約、質疑応答、その他のユースケースを実装する

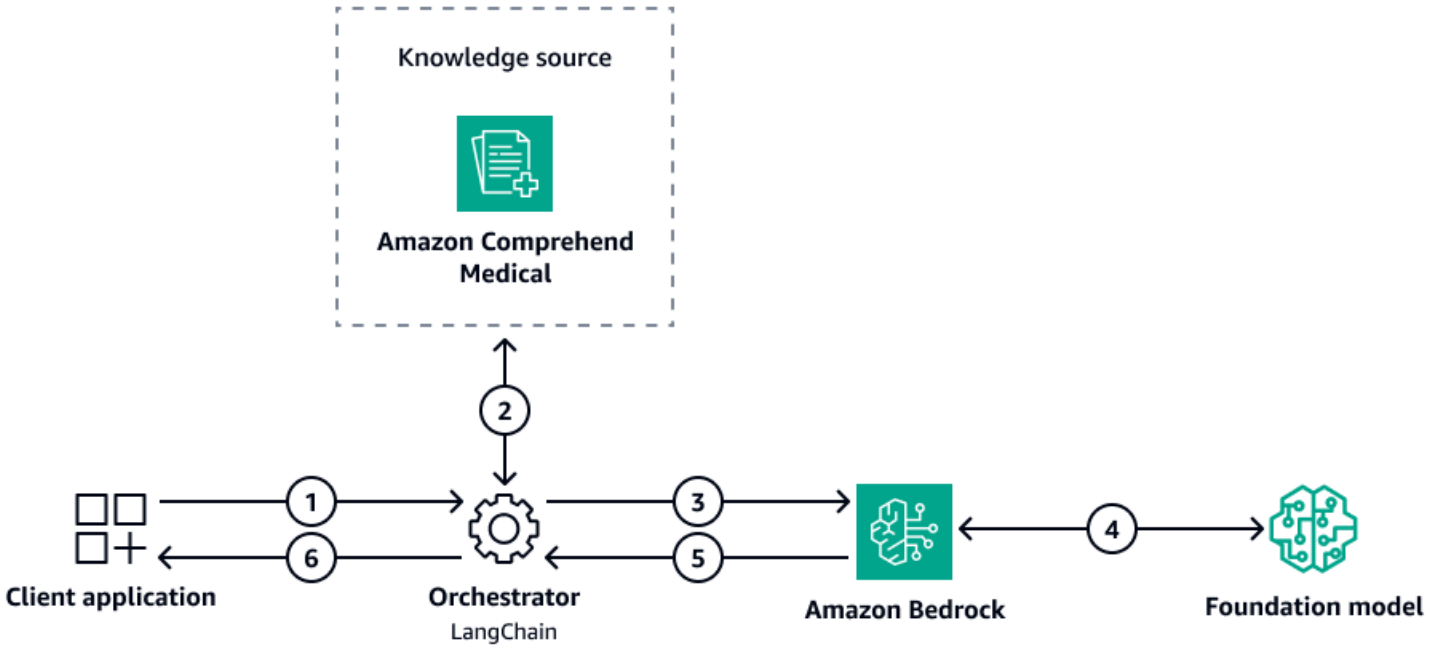
このセクションでは、検索拡張生成 (RAG) アプローチを使用して Amazon Comprehend Medical を LLM と組み合わせる方法について説明します。Retrieval Augmented Generation (RAG) は、LLM がレスポンスを生成する前にトレーニングデータソースの外部にある信頼できるデータソースを参照する生成 AI テクノロジーです。細については、「RAG (検索拡張生成) とは何ですか?」を参照してください。

このアプローチを説明するために、このセクションでは ICD-10-CM に関連する医療 (診断) コーディングの例を使用します。これには、イノベーションを加速するためのサンプルアーキテクチャとプロンプトエンジニアリングテンプレートが含まれています。また、RAG ワークフロー内で Amazon Comprehend Medical を使用するためのベストプラクティスも含まれています。

Amazon Comprehend Medical を使用した RAG ベースのアーキテクチャ

次の図は、患者ノートから ICD-10-CM 診断コードを識別するための RAG アプローチを示しています。Amazon Comprehend Medical をナレッジソースとして使用します。RAG アプローチでは、取

得方法は通常、該当する知識を含むベクトルデータベースから情報を取得します。ベクトルデータベースの代わりに、このアーキテクチャは取得タスクに Amazon Comprehend Medical を使用します。オーケストレーターは、患者メモ情報を Amazon Comprehend Medical に送信し、ICD-10-CM コード情報を取得します。オーケストレーターはこのコンテキストを Amazon Bedrock を介してダウンストリーム基盤モデル (LLM) に送信します。LLM は ICD-10-CM コード情報を使用してレスポンスを生成し、そのレスポンスはクライアントアプリケーションに返されます。



この図は、次の RAG ワークフローを示しています。

1. クライアントアプリケーションは、患者メモをクエリとしてオーケストレーターに送信します。これらの患者メモの例は、「患者は X 博士の 71 歳女性患者です。患者は昨日の夜に緊急治療室を訪問し、約 7 日から 8 日前には、永続的に持続しているおなかの痛みの履歴がありました。明らかな症状や症状はなく、黄疸の履歴もありません。The patient denies any significant recent weight loss.(患者は、最近の大きな体重減少はないと言っている。)」
2. オーケストレーターは Amazon Comprehend Medical を使用して、クエリ内の医療情報に関連する ICD-10-CM コードを取得します。InferICD10CM API を使用して、患者ノートから ICD-10-CM コードを抽出して推測します。
3. オーケストレーターは、プロンプトテンプレート、元のクエリ、Amazon Comprehend Medical から取得した ICD-10-CM コードを含むプロンプトを作成します。この拡張コンテキストを Amazon Bedrock に送信します。
4. Amazon Bedrock は入力を処理し、基盤モデルを使用して、クエリから ICD-10-CM コードとそれに対応する証拠を含むレスポンスを生成します。生成されたレスポンスには、識別された ICD-10-

CM コードと、各コードをサポートする患者ノートからの証拠が含まれます。レスポンスの例を次に示します。

```
<response>
<icd10>
<code>R10.9</code>
<evidence>history of abdominal pain</evidence>
</icd10>
<icd10>
<code>R10.30</code>
<evidence>history of abdominal pain</evidence>
</icd10>
</response>
```

5. Amazon Bedrock は、生成されたレスポンスをオーケストレーターに送信します。
6. オーケストレーターはレスポンスをクライアントアプリケーションに送信し、ユーザーはレスポンスを確認できます。

RAG ワークフローで Amazon Comprehend Medical を使用するユースケース

Amazon Comprehend Medical は、特定の NLP タスクを実行できます。詳細については、Amazon Comprehend Medical のユースケース」を参照してください。

Amazon Comprehend Medical を次のような高度なユースケースの RAG ワークフローに統合できます。

- 抽出された医療エンティティと患者記録からのコンテキスト情報を組み合わせて、詳細な臨床概要を生成する
- コード割り当てにオントロジーにリンクされた情報を持つ抽出されたエンティティを使用して、複雑なケースの医療コーディングを自動化する
- 抽出された医療エンティティを使用して、非構造化テキストから構造化された臨床メモの作成を自動化する
- 抽出された薬剤名と属性に基づいて薬剤の副作用を分析する
- 抽出された医療情報とup-to-date研究とガイドラインを組み合わせたインテリジェントな臨床サポートシステムを開発する

RAG ワークフローで Amazon Comprehend Medical を使用するためのベストプラクティス

Amazon Comprehend Medical の結果を LLM のプロンプトに統合するときは、ベストプラクティスに従うことが重要です。これにより、パフォーマンスと精度を向上させることができます。主な推奨事項は次のとおりです。

- Amazon Comprehend Medical 信頼スコアを理解する – Amazon Comprehend Medical は、検出された各エンティティとオントロジーリンクの信頼スコアを提供します。これらのスコアの意味を理解し、特定のユースケースに適したしきい値を設定することが重要です。信頼スコアは、信頼性の低いエンティティを除外し、ノイズを減らし、LLM の入力の品質を向上させるのに役立ちます。
- プロンプトエンジニアリングで信頼スコアを使用する – LLM のプロンプトを作成するときは、追加のコンテキストとして Amazon Comprehend Medical 信頼スコアを組み込むことを検討してください。これにより、LLM は信頼度に基づいてエンティティの優先順位付けや重み付けを行い、出力の品質を向上させることができます。
- Amazon Comprehend Medical の結果をグラウンドトゥルースデータで評価する – グラウンドトゥルースデータは true として知られている情報です。AI/ML アプリケーションが正確な結果を生成していることを検証するために使用できます。Amazon Comprehend Medical の結果を LLM ワークフローに統合する前に、データの代表的なサンプルでサービスのパフォーマンスを評価します。結果をグラウンドトゥルース注釈と比較し、潜在的な不一致や改善点を特定します。この評価は、ユースケースにおける Amazon Comprehend Medical の長所と制限を理解するのに役立ちます。
- 関連する情報を戦略的に選択する – Amazon Comprehend Medical は大量の情報を提供できますが、そのすべてがタスクに関連するとは限りません。ユースケースに最も関連性の高いエンティティ、属性、メタデータを慎重に選択します。LLM に無関係な情報が多すぎると、ノイズが発生し、パフォーマンスが低下する可能性があります。
- エンティティ定義の整列 – Amazon Comprehend Medical で使用されるエンティティと属性の定義が解釈と一致していることを確認します。不一致がある場合は、Amazon Comprehend Medical 出力と要件の間のギャップを埋めるために、LLM に追加のコンテキストまたは明確化を提供することを検討してください。Amazon Comprehend Medical エンティティが期待を満たさない場合は、プロンプトに追加の指示 (および可能な例) を含めることで、カスタムエンティティ検出を実装できます。
- ドメイン固有の知識を提供する – Amazon Comprehend Medical は貴重な医療情報を提供しますが、特定のドメインのすべてのニュアンスを把握できるとは限りません。Amazon Comprehend Medical の結果を、オントロジー、用語、エキスパートが厳選したデータセットなどの追加のドメイン固有のナレッジソースで補完することを検討してください。これにより、LLM のより包括的なコンテキストが提供されます。

- 倫理および規制ガイドラインに従う – 医療データを扱うときは、データのプライバシー、セキュリティ、医療における AI システムの責任ある使用に関連するものなど、倫理原則と規制ガイドラインに従うことが重要です。実装が関連する法律と業界のベストプラクティスに準拠していることを確認します。

これらのベストプラクティスに従うことで、AI/ML 実務者は Amazon Comprehend Medical と LLMs。医療 NLP タスクの場合、これらのベストプラクティスは潜在的なリスクを軽減し、パフォーマンスを向上させるのに役立ちます。

Amazon Comprehend Medical コンテキストのプロンプトエンジニアリング

プロンプトエンジニアリングは、生成 AI ソリューションをガイドして必要な出力を生成するプロンプトを設計および改良するプロセスです。AI がユーザーとより意味のあるやり取りをするための最も適切な形式、フレーズ、単語、記号を選択します。

実行する API オペレーションに応じて、Amazon Comprehend Medical は検出されたエンティティ、オントロジーコードと説明、および信頼スコアを返します。これらの結果は、ソリューションがターゲット LLM を呼び出すときにプロンプト内のコンテキストになります。プロンプトテンプレート内でコンテキストを表示するようにプロンプトを設計する必要があります。

> **Note**
>
> このセクションのプロンプト例は、Anthropic のガイダンスに従います。別の LLM プロバイダーを使用している場合は、そのプロバイダーからの推奨事項に従ってください。

一般的に、元の医療テキストと Amazon Comprehend Medical の結果の両方をプロンプトに挿入します。一般的なプロンプト構造は次のとおりです。

```
<medical_text>
medical text
</medical_text>

<comprehend_medical_text_results>
comprehend medical text results
</comprehend_medical_text_results>

<prompt_instructions>
prompt instructions
```

```
</prompt_instructions>
```

このセクションでは、Amazon Comprehend Medical の結果を以下の一般的な医療 NLP タスクのプロンプトコンテキストとして含める戦略について説明します。

- Amazon Comprehend Medical の結果をフィルタリングする
- Amazon Comprehend Medical を使用して医療 NLP タスクを拡張する
- Amazon Comprehend Medical でガードレールを適用する

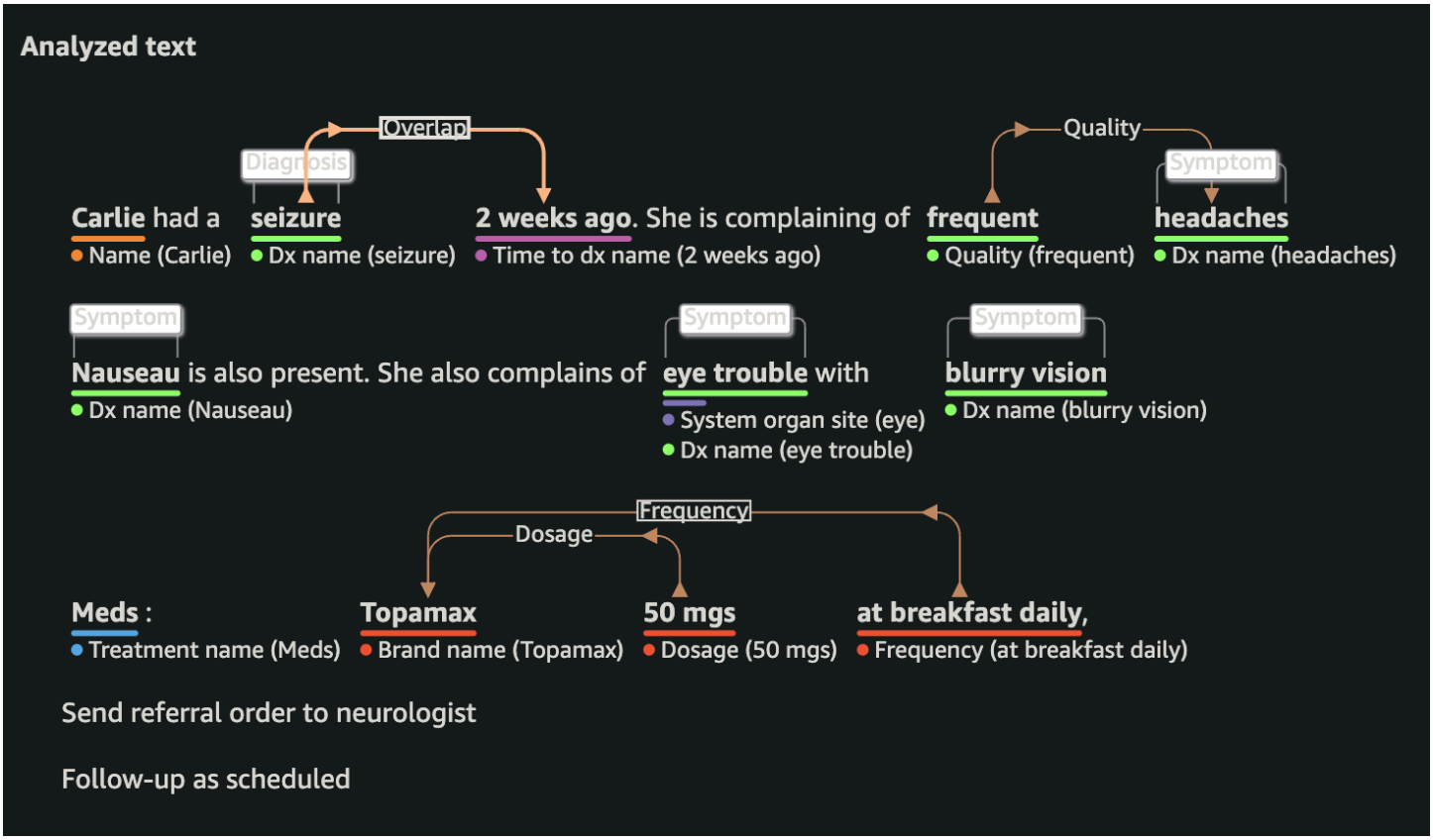
Amazon Comprehend Medical の結果をフィルタリングする

Amazon Comprehend Medical は通常、大量の情報を提供します。医療専門家が確認する必要がある結果の数を減らすことができます。この場合、LLM を使用してこれらの結果をフィルタリングできます。Amazon Comprehend Medical エンティティには、プロンプトの設計時にフィルタリングメカニズムとして使用できる信頼スコアが含まれています。

以下は、患者のメモの例です。

```
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
```

この患者メモでは、Amazon Comprehend Medical は次のエンティティを検出します。



エンティティは次の ICD-10-CM コードにリンクされ、けいれんや頭が痛くなります。

Category	ICD-10-CM コード	ICD-10-CM の説明	信頼スコア
けいれん	R56.9	指定されていないけいれん	0.8348
けいれん	G40.909	けいれん性ステータスのない、詳細不明、難読化不能のふるえ音	0.5424
けいれん	R56.00	単純な熱性けいれん	0.4937
けいれん	G40.09	その他のけいれん音	0.4397
けいれん	G40.409	その他の全般性遅発性および遅発性遅発性遅発性遅発性遅発性遅発性遅発性遅発性遅発性	0.4138

		遅発性遅発性遅発性遅発性遅発性遅発性遅発性遅発性遅発性遅発性遅	
ヘッダー	R51	ヘッダー	0.4067
ヘッダー	R51.9	ヘッダー、未指定	0.3844
ヘッダー	G44.52	新しい日次永続性痛 (NDPH)	0.3005
ヘッダー	G44	その他の頭部障害	0.2670
ヘッダー	G44.8	その他の指定された頭部症状	0.2542

ICD-10-CM コードをプロンプトに渡し、LLM の精度を高めることができます。ノイズを減らすために、Amazon Comprehend Medical の結果に含まれる信頼スコアを使用して ICD-10-CM コードをフィルタリングできます。以下は、信頼スコアが 0.4 を超える ICD-10-CM コードのみを含むプロンプトの例です。

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
</patient_note>

<comprehend_medical_results>
<icd-10>
  <entity>
    <text>seizure</text>
    <code>
      <description>Unspecified convulsions</description>
      <code_value>R56.9</code_value>
      <score>0.8347607851028442</score>
    </code>
    <code>
      <description>Epilepsy, unspecified, not intractable, without status epilepticus</
description>
```

```
      <code_value>G40.909</code_value>
      <score>0.542376697063446</score>
    </code>
    <code>
      <description>Other seizures</description>
      <code_value>G40.89</code_value>
      <score>0.43966275453567505</score>
    </code>
    <code>
      <description>Other generalized epilepsy and epileptic syndromes, not intractable,
 without status epilepticus</description>
      <code_value>G40.409</code_value>
      <score>0.41382506489753723</score>
    </code>
  </entity>
  <entity>
    <text>headaches</text>
    <code>
      <description>Headache</description>
      <code_value>R51</code_value>
      <score>0.4066613018512726</score>
    </code>
  </entity>
  <entity>
    <text>Nausea</text>
    <code>
      <description>Nausea</description>
      <code_value>R11.0</code_value>
      <score>0.646083474159240 7</score>
    </code>
  </entity>
  <entity>
    <text>eye trouble</text>
    <code>
      <description>Unspecified disorder of eye and adnexa</description>
      <code_value>H57.9</code_value>
      <score>0.6780954599380493</score>
    </code>
    <code>
      <description>Unspecified visual disturbance</description>
      <code_value>H53.9</code_value>
      <score>0.5871203541755676</score>
    </code>
    <code>
```

```
      <description>Unspecified disorder of binocular vision</description>
      <code_value>H53.30</code_value>
      <score>0.5539672374725342</score>
    </code>
  </entity>
  <entity>
    <text>blurry vision</text>
    <code>
      <description>Other visual disturbances</description>
      <code_value>H53.8</code_value>
      <score>0.9001834392547607</score>
    </code>
  </entity>
</icd-10>
</comprehend_medical_results>

<prompt>
Given the patient note and Amazon Comprehend Medical ICD-10-CM code results above,
 please select the most relevant ICD-10-CM diagnosis codes for the patient.
For each selected code, provide a brief explanation of why it is relevant based on the
 information in the patient note.
</prompt>
```

Amazon Comprehend Medical を使用して医療 NLP タスクを拡張する

医療テキストを処理する場合、Amazon Comprehend Medical のコンテキストは LLM がより良いトークンを選択するのに役立ちます。この例では、診断症状を薬剤に一致させたいと考えています。また、血パネル検査に関連する用語など、医療検査に関連するテキストも検索できます。Amazon Comprehend Medical を使用して、エンティティと薬剤名を検出できます。この場合、Amazon Comprehend Medical の DetectEntitiesV2 および InferRxNorm APIs を使用します。

以下は、患者のメモの例です。

```
Carlie had a seizure 2 weeks ago. She is complaining of increased frequent headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
```

診断コードに焦点を当てるために、 タイプの MEDICAL_CONDITION に関連するエンティティのみがプロンプトでDX_NAME使用されます。関連性がないため、その他のメタデータは除外されます。薬剤エンティティの場合、薬剤名と抽出された属性が含まれます。Amazon Comprehend Medical

の他の薬剤エンティティメタデータは、関連性がないため除外されます。以下は、フィルタリングされた Amazon Comprehend Medical の結果を使用するプロンプトの例です。プロンプト は、DX_NAMEタイプを持つMEDICAL_CONDITIONエンティティに焦点を当てます。このプロンプトは、診断コードを薬剤とより正確にリンクし、より正確に医療指示テストを抽出するように設計されています。

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of increased freqeunt headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order
 prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
</patient_note>

<detect_entity_results>
<entity>
    <text>seizure</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>headaches</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>lyme disease</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>muscle ache</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>stiff neck</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
</detect_entity_results>

<rx_results>
```

```
<entity>
    <text>Topamax</text>
    <category>MEDICATION</category>
    <type>BRAND_NAME</type>
    <attributes>
        <attribute>
            <type>FREQUENCY</type>
            <text>at breakfast daily</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>50 mgs</text>
        </attribute>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
    </attributes>
</entity>
<entity>
    <text>Amoxicillan</text>
    <category>MEDICATION</category>
    <type>GENERIC_NAME</type>
    <attributes>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>25 mg</text>
        </attribute>
        <attribute>
            <type>FREQUENCY</type>
            <text>twice a day</text>
        </attribute>
    </attributes>
</entity>
</rx_results>

<prompt>
Based on the patient note and the detected entities, can you please:
1. Link the diagnosis symptoms with the medications prescribed.
Provide your reasoning for the linkages.
```

```
2. Extract any entities related to medical order tests mentioned in the note.
</prompt>
```

Amazon Comprehend Medical でガードレールを適用する

生成されたレスポンスを使用する前に、LLM と Amazon Comprehend Medical を使用してガードレールを作成できます。このワークフローは、未変更または後処理の医療テキストで実行できます。ユースケースには、保護された医療情報 (PHI) の対処、幻覚の検出、結果を発行するためのカスタムポリシーの実装などがあります。例えば、Amazon Comprehend Medical のコンテキストを使用して PHI データを識別し、LLM を使用してその PHI データを削除できます。

以下は、PHI を含む患者レコードからの情報の例です。

```
Patient name: John Doe
Patient SSN: 123-34-5678
Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
```

以下は、Amazon Comprehend Medical の結果をコンテキストとして含むプロンプトの例です。

```
<original_text>
Patient name: John Doe
Patient SSN: 123-34-5678 Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
</original_text>

<comprehend_medical_phi_entities>
<entity>
  <text>John Doe</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9967944025993347</score>
  <type>NAME</type>
</entity>
<entity>
  <text>123-34-5678</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9998034834861755</score>
  <type>ID</type>
</entity>
<entity>
```

```
  <text>01/01/2000</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9964448809623718</score>
  <type>DATE</type>
</entity>
</comprehend_medical_phi_entities>

<instructions>
Using the provided original text and the Amazon Comprehend Medical PHI entities
 detected, please analyze the text to determine if it contains any additional protected
 health information (PHI) beyond the entities already identified. If additional PHI is
 found, please list and categorize it. If no additional PHI is found, please state that
 explicitly.
In addition if PHI is found, generate updated text with the PHI removed.
</instructions>
```

ヘルスケアとライフサイエンスのユースケースでの大規模言語モデルの使用

ここでは、医療およびライフサイエンスアプリケーションに大規模言語モデル (LLMs) を使用する方法について説明します。一部のユースケースでは、生成 AI 機能に大規模言語モデルを使用する必要があります。state-of-the-art LLMs にも利点と制限があり、このセクションの推奨事項は、ターゲット結果の達成に役立つように設計されています。

決定パスを使用して、ドメインの知識や利用可能なトレーニングデータなどの要因を考慮して、ユースケースに適した LLM ソリューションを決定できます。さらに、このセクションでは、一般的な事前トレーニング済み医療 LLMsとその選択と使用に関するベストプラクティスについて説明します。また、複雑で高性能なソリューションと、よりシンプルで低コストなアプローチのトレードオフについても説明します。

LLM のユースケース

Amazon Comprehend Medical は、特定の NLP タスクを実行できます。詳細については、「Amazon Comprehend Medical のユースケース」を参照してください。

LLM の論理および生成 AI 機能は、次のような高度な医療およびライフサイエンスのユースケースで必要になる場合があります。

- カスタム医療エンティティまたはテキストカテゴリの分類
- 臨床的な質問に回答する

- メディカルレポートの概要
- 医療情報からのインサイトの生成と検出

カスタマイズアプローチ

LLMs実装方法を理解することが重要です。LLMs、多くのドメインからのトレーニングデータを含む数十億のパラメータでトレーニングされます。このトレーニングにより、LLM は最も一般化されたタスクに対処できます。ただし、ドメイン固有の知識が必要な場合に課題が発生することがよくあります。ヘルスケアとライフサイエンスの分野の知識の例としては、正確な回答を生成するために必要な臨床コード、医学用語、健康情報などがあります。したがって、これらのユースケースで LLM をそのまま使用すると (ドメインの知識を補足せずにゼロショットプロンプトを実行) 、結果が不正確になる可能性があります。この課題を克服するために使用できる一般的なアプローチには、プロンプトエンジニアリング、検索拡張生成 (RAG)、ファインチューニングなどがあります。

プロンプトエンジニアリング

プロンプトエンジニアリングは、LLM への入力を調整することで、生成 AI ソリューションをガイドして必要な出力を作成するプロセスです。関連するコンテキストで正確なプロンプトを作成することで、推論を必要とする専門的な医療タスクの完了に向けてモデルをガイドすることができます。効果的なプロンプトエンジニアリングにより、モデルの変更を必要とせずに、医療ユースケースのモデルパフォーマンスを大幅に向上させることができます。プロンプトエンジニアリングの詳細については、「Amazon Bedrock を使用した高度なプロンプトエンジニアリングの実装」(AWS ブログ記事) を参照してください。少数ショットプロンプトとchain-of-thoughtプロンプトは、プロンプトエンジニアリングで使用できる手法です。

数ショットプロンプト

少数ショットプロンプトは、LLM に同様のタスクの実行を求める前に、必要な入出力のいくつかの例を提供する手法です。医療の文脈では、このアプローチは、医療エンティティの認識や臨床メモの要約などの特殊なタスクに特に役立ちます。プロンプトに 3～5 個の高品質の例を含めることで、医学用語とドメイン固有のパターンに関するモデルの理解を大幅に向上させることができます。数ショットプロンプトの例については、「Amazon Bedrock での LLMs」(AWS ブログ記事) を参照してください。

例えば、臨床ノートから薬剤の投与量を抽出する場合、医療専門家が処方を記録する方法のバリエーションをモデルが認識するのに役立つさまざまな表記スタイルの例を提供できます。このアプローチは、標準化されたドキュメント形式を使用する場合や、データに一貫したパターンが存在する場合に特に効果的です。

Chain-of-thoughtプロンプト

Chain-of-thought (CoT) プロンプトは、LLM をステップstep-by-stepの推論プロセスでガイドします。これにより、複雑な医療上の意思決定のサポートや診断の推論タスクに役立ちます。臨床シナリオを分析するときにモデルに「ステップバイステップ」を明示的に指示することで、医療推論プロトコルに従う能力を向上させ、診断エラーを減らすことができます。

この手法は、臨床推論に差分診断や治療計画などの複数の論理的なステップが必要な場合に優れています。ただし、モデルのトレーニングデータ外で高度に専門的な医療知識を扱う場合や、クリティカルケアの決定に絶対精度が必要な場合、このアプローチには制限があります。

このような場合、CoT を別のアプローチと組み合わせると、より良い結果が得られます。1 つのオプションは、CoT と自己整合性プロンプト を組み合わせることです。詳細については、「Amazon Bedrock での自己整合性プロンプトによる生成言語モデルのパフォーマンスの向上」(AWS ブログ記事) を参照してください。もう 1 つのオプションは、ReAct プロンプトなどの推論フレームワークを RAG と組み合わせることです。詳細については、「RAG および ReAct プロンプトを使用して高度な生成 AI チャットベースのアシスタントを開発する」(AWS 規範ガイダンス) を参照してください。

検索拡張生成

Retrieval Augmented Generation (RAG) は、LLM がレスポンスを生成する前にトレーニングデータソースの外部にある信頼できるデータソースを参照する生成 AI テクノロジーです。RAG システムは、ナレッジソースから医療オントロジー情報 (疾病の国際分類、国の医薬品ファイル、医療対象者の見出しなど) を取得できます。これにより、医療 NLP タスクをサポートするために LLM に追加のコンテキストが提供されます。

Amazon Comprehend Medical と大規模言語モデルの組み合わせ セクションで説明したように、RAG アプローチを使用して Amazon Comprehend Medical からコンテキストを取得できます。その他の一般的なナレッジソースには、Amazon OpenSearch Service、Amazon Kendra、Amazon Aurora などのデータベースサービスに保存されている医療ドメインデータが含まれます。これらのナレッジソースから情報を抽出すると、特にベクトルデータベースを使用するセマンティッククエリでは、取得パフォーマンスに影響する可能性があります。

ドメイン固有の知識を保存および取得するもう 1 つのオプションは、RAG ワークフローで Amazon Q Business を使用することです。Amazon Q Business は、内部ドキュメントリポジトリまたは公開ウェブサイト (ICD-10 データの場合は CMS.gov など) のインデックスを作成できます。Amazon Q Business は、クエリを LLM に渡す前に、これらのソースから関連情報を抽出できます。

カスタム RAG ワークフローを構築する方法は複数あります。たとえば、ナレッジソースからデータを取得する方法は多数あります。簡単にするために、Amazon OpenSearch Service などのベクトル

データベースを使用して知識を埋め込みとして保存する一般的な取得アプローチをお勧めします。そのためには、文トランスフォーマーなどの埋め込みモデルを使用して、クエリとベクトルデータベースに保存されているナレッジの埋め込みを生成する必要があります。

フルマネージド型およびカスタム RAG アプローチの詳細については、「Retrieval Augmented Generation options and architectures on AWS」を参照してください。

ファインチューニング

既存のモデルを微調整するには、Amazon Titan、Mistral、Llama モデルなどの LLM を取得し、カスタムデータにモデルを適応させる必要があります。ファインチューニングにはさまざまな手法があり、そのほとんどはモデル内のすべてのパラメータを変更するのではなく、少数のパラメータのみを変更するものです。これは、パラメータ効率の高い微調整 (PEFT) と呼ばれます。詳細については、GitHub の「Hugging Face PEFT」を参照してください。

以下は、医療 NLP タスクの LLM を微調整する場合の 2 つの一般的なユースケースです。

- 生成タスク – デコーダーベースのモデルは生成 AI タスクを実行します。AI/ML 実務者はグラウンドトゥルースデータを使用して既存の LLM を微調整します。例えば、公的医療質問回答データセットである MedQuAD を使用して LLM をトレーニングできます。ファインチューニングされた LLM にクエリを呼び出す場合、LLM に追加のコンテキストを提供するために RAG アプローチは必要ありません。
- 埋め込み – エンコーダベースのモデルは、テキストを数値ベクトルに変換することで埋め込みを生成します。これらのエンコーダーベースのモデルは通常、埋め込みモデルと呼ばれます。文変換モデルは、文用に最適化された特定のタイプの埋め込みモデルです。目的は、入力テキストから埋め込みを生成することです。その後、埋め込みはセマンティック分析または取得タスクに使用されます。埋め込みモデルを微調整するには、トレーニングデータとして使用できるドキュメントなどの医療知識のコーパスが必要です。これは、文トランスフォーマーモデルを微調整するための類似性または感情に基づくテキストのペアで実現されます。詳細については、Hugging Face の「Training and Finetuning Embedding Models with Sentence Transformers v3」を参照してください。

Amazon SageMaker Ground Truth を使用して、ラベル付きの高品質のトレーニングデータセットを構築できます。Ground Truth のラベル付きデータセット出力を使用して、独自のモデルをトレーニングできます。出力を Amazon SageMaker AI モデルをトレーニングデータセットとして使用することもできます。名前付きエンティティ認識、単一ラベルテキスト分類、マルチラベルテキスト分類の詳細については、Amazon SageMaker AI ドキュメントの「Ground Truth を使用したテキストラベル付け」を参照してください。

ファインチューニングの詳細については、このガイドヘルスケアにおける大規模言語モデルのファインチューニングの「」を参照してください。

LLM の選択

Amazon Bedrock は、高性能 LLMs。詳細については、「Amazon Bedrock でサポートされている基盤モデル」を参照してください。Amazon Bedrock のモデル評価ジョブを使用して、複数の出力からの出力を比較し、ユースケースに最適なモデルを選択できます。詳細については、Amazon Bedrock ドキュメントの「Amazon Bedrock 評価を使用して最もパフォーマンスの高いモデルを選択する」を参照してください。

一部の LLMsでは、医療ドメインデータに関するトレーニングが制限されています。ユースケースで Amazon Bedrock がサポートしていない LLM または LLM の微調整が必要な場合は、Amazon SageMaker AI の使用を検討してください。SageMaker AI では、微調整された LLM を使用するか、医療ドメインデータでトレーニングされたカスタム LLM を選択できます。

次の表LLMs の一覧です。

LLM	タスク	ナレッジ	アーキテクチャ
BioBERT	情報の取得、テキスト分類、および名前付きエンティティの認識	PubMed からの抜粋、PubMedCentral からの全文記事、一般的なドメイン知識	エンコーダー
ClinicalBERT	情報の取得、テキスト分類、および名前付きエンティティの認識	電子ヘルスレコード (EHR) システムから 3,000,000 を超える患者レコードを含む大規模な多施設データセット	エンコーダー
ClinicalGPT	要約、質疑応答、テキスト生成	医療記録、ドメイン固有の知識、マルチラウンド対話の相談など、広範で多様な医療データセット	デコーダー

GatorTron-OG	要約、質疑応答、テキスト生成、情報取得	臨床ノートと生体医学の文献	エンコーダー
Med-BERT	情報の取得、テキスト分類、および名前付きエンティティの認識	医療テキスト、臨床ノート、研究論文、医療関連ドキュメントの大規模なデータセット	エンコーダー
Med-PaLM	医療目的の質疑応答	医療テキストとバイオメディカルテキストのデータセット	デコーダー
medAlpaca	質問への回答と医療対話タスク	医療フラッシュカード、Wiki、ダイアログデータセットなどのリソースを含むさまざまな医療テキスト	デコーダー
BiomedBERT	情報の取得、テキスト分類、および名前付きエンティティの認識	PubMed および PubMedCentral からの全文記事から排他的に抽象化	エンコーダー
BioMedLM	要約、質疑応答、テキスト生成	PubMed ナレッジソースからの生物医学文献	デコーダー

以下は、事前トレーニング済みの医療 LLMs。

- トレーニングデータとその医療 NLP タスクとの関連性を理解します。
- LLM アーキテクチャとその目的を特定します。エンコーダーは、埋め込みおよび NLP タスクに適しています。デコーダーは生成タスク用です。
- 事前トレーニング済みの医療 LLM をホストするためのインフラストラクチャ、パフォーマンス、コスト要件を評価します。
- 微調整が必要な場合は、トレーニングデータの正確なグラウンドトゥルースまたは知識を確保します。個人を特定できる情報 (PII) または保護された医療情報 (PHI) をマスクまたは編集してください。

実際の医療 NLP タスクは、知識や意図したユースケースの点で、事前トレーニング済みの LLMs とは異なる場合があります。ドメイン固有の LLM が評価ベンチマークを満たさない場合は、独自のデータセットを使用して LLM を微調整することも、新しい基盤モデルをトレーニングすることもできます。新しい基盤モデルのトレーニングは、野心的で、多くの場合、コストがかかります。ほとんどのユースケースでは、既存のモデルを微調整することをお勧めします。

事前トレーニング済みの医療 LLM を使用または微調整する場合は、インフラストラクチャ、セキュリティ、ガードレールに対応することが重要です。

インフラストラクチャ

オンデマンドまたはバッチ推論に Amazon Bedrock を使用する場合と比較して、事前トレーニング済みの医療 LLMs (Hugging Face からのみ) をホストするには、大量のリソースが必要です。事前トレーニング済みの医療 LLMs をホストするには、Amazon Elastic Compute Cloud (Amazon EC2) インスタンスで実行される Amazon SageMaker AI イメージを、高速コンピューティング用の ml.g5 インスタンスや ml.inf2 インスタンスなどの 1 つ以上の GPUs とともに使用するのが一般的です AWS Inferentia。これは、LLMs消費するためです。

セキュリティとガードレール

ビジネスコンプライアンス要件に応じて、Amazon Comprehend と Amazon Comprehend Medical を使用して、トレーニングデータから個人を特定できる情報 (PII) と保護医療情報 (PHI) をマスクまたは編集することを検討してください。これにより、LLM がレスポンスを生成するときに機密データを使用するのを防ぐことができます。

生成 AI アプリケーションのバイアス、公平性、幻覚を考慮し、評価することをお勧めします。既存の LLM を使用している場合でも、ファインチューニングを使用している場合でも、有害な応答を防ぐためにガードレールを実装します。ガードレールは、生成 AI アプリケーションの要件と責任ある AI ポリシーに合わせてカスタマイズする保護手段です。例えば、Amazon Bedrock ガードレールを使用できます。

ヘルスケアにおける大規模言語モデルのファインチューニング

このセクションで説明するファインチューニングアプローチは、倫理および規制ガイドラインへの準拠をサポートし、医療における AI システムの責任ある使用を促進します。これは、正確でプライベートなインサイトを生成するように設計されています。生成 AI は医療の提供に変革をもたらしていますがoff-the-shelfモデルが不足することがよくあります。ドメイン固有のデータを使用して基盤モデルを微調整することで、このギャップを埋めることができます。これは、厳格な規制基準に準拠

しながら、医学の言語を話す AI システムを作成するのに役立ちます。ただし、ファインチューニングを成功させるには、機密データの保護、測定可能な成果を伴う AI 投資の正当化、急速に進化する医療環境での臨床的な関連性の維持など、医療固有の課題を慎重に把握する必要があります。

軽量アプローチが制限に達すると、ファインチューニングは戦略的投資になります。精度、レイテンシー、または運用効率の向上により、必要なコンピューティングとエンジニアリングの大幅なコストが相殺されることが期待されます。基盤モデルの進行速度は速いため、微調整されたモデルの利点は、次のメジャーモデルリリースまで続く可能性があることに注意してください。

このセクションでは、ヘルスケア業界の AWS お客様からの次の 2 つの影響の大きいユースケースについて説明します。

- 臨床決定支援システム – 複雑な患者の履歴と進化するガイドラインを理解するモデルを通じて診断精度を向上させます。微調整は、モデルが複雑な患者の履歴を深く理解し、特殊なガイドラインを統合するのに役立ちます。これにより、モデル予測エラーを減らすことができます。ただし、これらの利益と、大規模で機密性の高いデータセットのトレーニングコスト、および高リスクの臨床アプリケーションに必要なインフラストラクチャを比較検討する必要があります。特に新しいモデルが頻繁にリリースされる場合、精度とコンテキスト認識の向上は投資を正当化しますか?
- 医療文書分析 – 医療保険の相互運用性と説明責任に関する法律 (HIPAA) コンプライアンスを維持しながら、臨床記録、画像レポート、保険文書の処理を自動化します。ここでは、微調整により、モデルが一意の形式、特殊な略語、規制要件をより効果的に処理できる場合があります。ペイオフは、手動レビュー時間の短縮とコンプライアンスの向上でよく見られます。それでも、これらの改善が微調整リソースを正当化するのに十分な大きさであるかどうかを評価することが重要です。プロンプトエンジニアリングとワークフローオーケストレーションがニーズを満たすことができるかどうかを判断します。

これらの実世界のシナリオは、初期の実験からモデルデプロイまでのファインチューニングジャーニーを示しながら、あらゆる段階で医療固有の要件に対処します。

コストと投資収益率の見積もり

以下は、LLM をファインチューニングするときに考慮すべきコスト要因です。

- モデルサイズ – モデルが大きいほど微調整にコストがかかります
- データセットサイズ – ファインチューニング用のデータセットのサイズに伴うコンピューティングコストと時間の増加
- ファインチューニング戦略 – パラメータ効率の高い方法では、パラメータの完全な更新と比較してコストを削減できます。

投資収益率 (ROI) を計算するときは、選択したメトリクス (精度など) にリクエストの量 (モデルが使用される頻度) を掛けた改善と、モデルが新しいバージョンで超過するまでの予想期間を考慮してください。

また、基本 LLM の有効期間も考慮してください。6～12 か月ごとに新しいベースモデルが登場します。希少疾患ディテクターの微調整と検証に 8 か月かかる場合、新しいモデルがギャップを埋めるまでに 4 か月しか優れたパフォーマンスが得られない可能性があります。

ユースケースのコスト、ROI、および潜在的な存続期間を計算することで、データ駆動型の意思決定を行うことができます。例えば、臨床意思決定サポートモデルを微調整すると、年間数千のケースで診断エラーが測定可能な程度に減少する場合、投資はすぐに報われる可能性があります。逆に、プロンプトエンジニアリングだけでドキュメント分析ワークフローが目標精度に近づく場合、次世代のモデルが到着するまでファインチューニングを延期することをお勧めします。

ファインチューニングは one-size-fits-allではありません。微調整を行う場合、適切なアプローチはユースケース、データ、リソースによって異なります。

ファインチューニング戦略の選択

ファインチューニングが医療ユースケースに適したアプローチであると判断したら、次のステップとして最も適切なファインチューニング戦略を選択します。利用可能なアプローチはいくつかあります。各 には、ヘルスケアアプリケーションに固有の利点とトレードオフがあります。これらの方法の選択は、特定の目標、利用可能なデータ、リソースの制約によって異なります。

トレーニングの目的

ドメイン適応事前トレーニング (DAPT) は、ドメイン固有のラベル付けされていない大量のテキスト (何百万もの医療文書など) でモデルを事前トレーニングする、教師なしの方法です。このアプローチは、放射線学者、神経学者、その他の専門プロバイダーが使用する医療専門分野の略語と用語を理解するモデルの能力を向上させるのに適しています。ただし、DAPT には大量のデータが必要であり、特定のタスク出力には対応しません。

教師ありファインチューニング (SFT) では、構造化された入出力例を使用して、明示的な指示に従うようにモデルに指示します。このアプローチは、ドキュメントの要約や臨床コーディングなどの医療ドキュメント分析ワークフローに適しています。命令チューニングは、目的の出力とペアになった明示的な命令を含む例でモデルをトレーニングする SFT の一般的な形式です。これにより、モデルの多様なユーザープロンプトを理解し、それに従う能力が向上します。この手法は、特定の臨床例でモデルをトレーニングするため、医療環境で特に役立ちます。主な欠点は、慎重にラベル付けされた例が必要であることです。さらに、微調整されたモデルは、例のないエッジケースで苦労する

可能性があります。Amazon SageMaker Jumpstart を使用したファインチューニングの手順については、Amazon SageMaker Jumpstart を使用した FLAN T5 XL のファインチューニング手順」(AWS ブログ記事) を参照してください。

人間のフィードバックからの強化学習 (RLHF) は、専門家のフィードバックと好みに基づいてモデルの動作を最適化します。プロキシマルポリシー最適化 (PPO) や直接設定最適化 (DPO) など、人間の好みや方法に基づいてトレーニングされた報酬モデルを使用して、破壊的な更新を防止しながらモデルを最適化します。RLHF は、出力を臨床ガイドラインに合わせて調整し、レコメンデーションが承認されたプロトコル内に収まるようにするのに最適です。このアプローチでは、臨床医のフィードバックにかなりの時間が必要であり、複雑なトレーニングパイプラインが必要です。ただし、RLHF は、医療の専門家が AI システムの通信方法やレコメンデーションを行う方法を形成するのに役立つため、医療において特に重要です。例えば、臨床医はフィードバックを提供して、モデルが適切な脇道を維持し、いつ不確実性を表現すべきかを理解し、臨床ガイドラインに従っていることを確認できます。PPO などの手法は、主要な医療知識を保持するためにパラメータの更新を制限しながら、専門家のフィードバックに基づいてモデルの動作を繰り返し最適化します。これにより、モデルは患者にとってわかりやすい言語で複雑な診断を伝達しながら、直ちに医療を受けるための重大な条件にフラグを付けることができます。これは、精度とコミュニケーションスタイルの両方が重要な医療にとって重要です。RLHF の詳細については、「Fine-tune large language models with reinforcement learning from human or AI feedback」(AWS ブログ記事) を参照してください。

実装方法

完全なパラメータ更新では、トレーニング中にすべてのモデルパラメータを更新する必要があります。このアプローチは、患者の履歴、検査結果、進化するガイドラインの深い統合を必要とする臨床意思決定サポートシステムに最適です。欠点には、データセットが大きく多様でない場合の、高いコンピューティングコストと過剰適合のリスクが含まれます。

パラメータ効率の高いファインチューニング (PEFT) メソッドは、オーバーフィットや言語機能の壊滅的な損失を防ぐために、パラメータのサブセットのみを更新します。タイプには、低ランク適応 (LoRA)、アダプター、プレフィックスチューニングなどがあります。PEFT メソッドは、計算コストが低く、トレーニングが速くなり、臨床決定サポートモデルを新しい病院のプロトコルや用語に適応させるなどの実験に最適です。主な制限は、完全なパラメータ更新と比較してパフォーマンスが低下する可能性があることです。

ファインチューニング方法の詳細については、Amazon SageMakerの高度なファインチューニング方法」(AWS ブログ記事) を参照してください。

ファインチューニングデータセットの構築

ファインチューニングデータセットの品質と多様性は、モデルのパフォーマンス、安全性、バイアス防止にとって重要です。このデータセットを構築する際に考慮すべき 3 つの重要な領域を次に示します。

- ファインチューニングアプローチに基づくボリューム
- ドメインエキスパートからのデータ注釈
- データセットの多様性

次の表に示すように、ファインチューニングのデータセットサイズ要件は、実行されるファインチューニングのタイプによって異なります。

ファインチューニング戦略	データセットサイズ
ドメイン適応事前トレーニング	100,000 を超えるドメインテキスト
教師ありファインチューニング	10,000 以上のラベル付きペア
人間のフィードバックによる学習の強化	1,000 を超えるエキスパート設定ペア

AWS Glue、Amazon EMR、Amazon SageMaker Data Wrangler を使用して、データ抽出と変換プロセスを自動化し、所有するデータセットをキュレートできます。十分なサイズのデータセットをキュレートできない場合は、 AWS アカウント を使用してデータセットを検出し、 に直接ダウンロードできますAWS Data Exchange。サードパーティーのデータセットを利用する前に、法律顧問に相談してください。

医療データと生物学的データのニュアンスをモデル出力に組み込むには、医学者、バイオロジスト、化学者などのドメイン知識を持つ専門家のアノテーターをデータキュレーションプロセスの一部にする必要があります。Amazon SageMaker Ground Truth は、エキスパートがデータセットに注釈を付けるためのローコードユーザーインターフェイスを提供します。

人間の母集団を表すデータセットは、バイアスを防ぎ、実際の結果を反映するために、ヘルスケアやライフサイエンスのファインチューニングのユースケースに不可欠です。 AWS Glue インタラクティブセッションまたは Amazon SageMaker ノートブックインスタンスは、Jupyter 互換ノートブックを使用してデータセットを繰り返し探索し、変換をファインチューニングする強力な方法を提供します。インタラクティブセッションを使用すると、ローカル環境で一般的な統合開発環境 (IDEs

を選択できます。または、 を使用して AWS Glue または Amazon SageMaker Studio ノートブックを使用することもできます AWS マネジメントコンソール。

モデルの微調整

AWS は、ファインチューニングを成功させるために不可欠な Amazon Amazon SageMaker AI や Amazon Bedrock などのサービスを提供します。

SageMaker AI は、開発者やデータサイエンティストが ML モデルを迅速に構築、トレーニング、デプロイできるようにするフルマネージド型の機械学習サービスです。SageMaker AI の微調整に役立つ 3 つの機能は次のとおりです。

- SageMaker トレーニング – 幅広いモデルを大規模に効率的にトレーニングするのに役立つフルマネージド ML 機能
- SageMaker JumpStart – SageMaker トレーニングジョブ上に構築され、ML タスク用の事前トレーニング済みモデル、組み込みアルゴリズム、ソリューションテンプレートを提供する機能
- SageMaker HyperPod – 基盤モデルと LLMs の分散トレーニング専用のインフラストラクチャソリューション

Amazon Bedrock は、セキュリティ、プライバシー、スケーラビリティ機能が組み込まれた、API を通じて高性能な基盤モデルへのアクセスを提供するフルマネージドサービスです。このサービスは、利用可能ないくつかの基本モデルを微調整する機能を提供します。詳細については、Amazon Bedrock ドキュメントの「ファインチューニングと継続的な事前トレーニングでサポートされているモデルとリージョン」を参照してください。

いずれかのサービスでファインチューニングプロセスに近づくときは、ベースモデル、ファインチューニング戦略、インフラストラクチャを検討してください。

基本モデルの選択

Anthropic Claude、Meta Llama、Amazon Nova などのクローズドソースモデルは、マネージドコンプライアンスでout-of-the-box使える強力なパフォーマンスを提供しますが、Amazon Bedrock などのマネージド APIs などのプロバイダーがサポートするオプションにファインチューニングの柔軟性を制限します。これにより、特に規制された医療ユースケースのカスタマイズ可能性が制約されます。対照的に、Meta Llama などのオープンソースモデルは、Amazon SageMaker AI サービス全体で完全な制御と柔軟性を提供するため、モデルを特定のデータまたはワークフロー要件に合わせてカスタマイズ、監査、または深く適応させる必要がある場合に最適です。

ファインチューニング戦略

簡単な命令チューニングは、Amazon Bedrock モデルのカスタマイズまたは Amazon SageMaker JumpStart で処理できます。LoRA やアダプターなどの複雑な PEFT アプローチには、Amazon Bedrock の SageMaker トレーニングジョブまたはカスタム微調整機能が必要です。非常に大規模なモデルの分散トレーニングは、SageMaker HyperPod でサポートされています。

インフラストラクチャのスケールと制御

Amazon Bedrock などのフルマネージドサービスは、インフラストラクチャ管理を最小限に抑え、使いやすさとコンプライアンスを優先する組織に最適です。SageMaker JumpStart などのセミマネージドオプションは、複雑さを抑えながら柔軟性を提供します。これらのオプションは、ラピッドプロトタイピングや、構築済みのワークフローを使用する場合に適しています。フルコントロールとカスタマイズには SageMaker トレーニングジョブと HyperPod が付属していますが、これらにはより多くの専門知識が必要であり、大規模なデータセットのスケールアップやカスタムパイプラインが必要な場合に最適です。

微調整されたモデルのモニタリング

ヘルスケアとライフサイエンスでは、LLM ファインチューニングをモニタリングするには、複数の主要業績評価指標を追跡する必要があります。精度はベースライン測定を提供しますが、特に誤分類が重大な結果をもたらすアプリケーションでは、精度と再現率のバランスを取る必要があります。F1-scoreは、医療データセットでよく見られるクラスの不均衡問題に対処するのに役立ちます。詳細については、このガイドの「ヘルスケアおよびライフサイエンスアプリケーション用の LLMs の評価」を参照してください。

キャリブレーションメトリクスは、モデルの信頼レベルが実際の確率と一致することを確認するのに役立ちます。公平性メトリクスは、さまざまな患者属性にわたる潜在的なバイアスを検出するのに役立ちます。

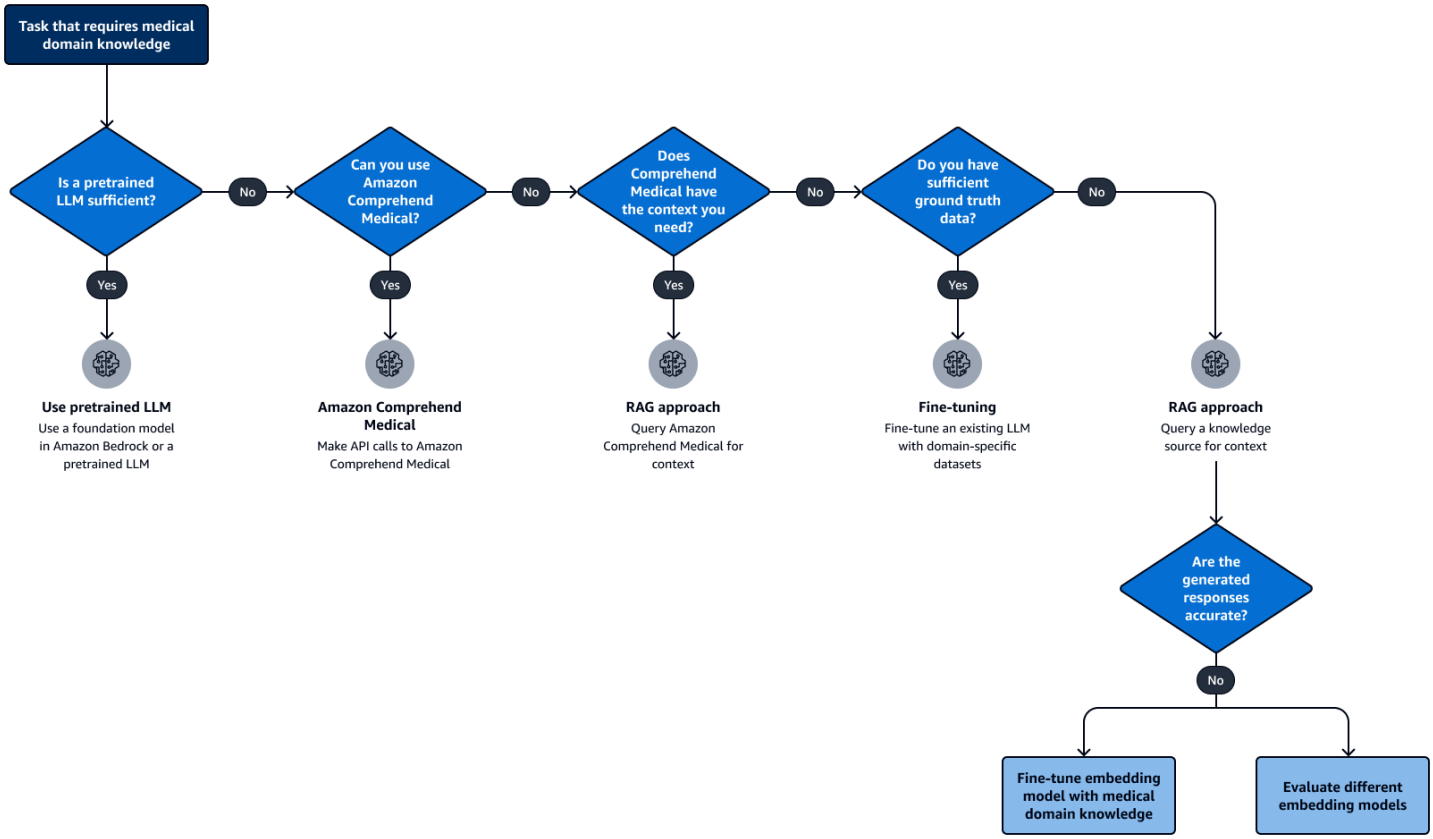
MLflow は、ファインチューニング実験の追跡に役立つオープンソースソリューションです。MLflow は Amazon SageMaker AI 内でネイティブにサポートされており、トレーニング実行のメトリクスを視覚的に比較するのに役立ちます。Amazon Bedrock の ファインチューニングジョブの場合、メトリクスは Amazon CloudWatch にストリーミングされ、CloudWatch コンソールでメトリクスを視覚化できます。

ヘルスケアとライフサイエンスの NLP アプローチの選択

ヘルスケアとライフサイエンスのための生成 AI と NLP アプローチ このセクションでは、ヘルスケアおよびライフサイエンスアプリケーションの自然言語処理 (NLP) タスクに対処するための以下のアプローチについて説明します。

- Amazon Comprehend Medical の使用
- 検索拡張生成 (RAG) ワークフローでの Amazon Comprehend Medical と LLM の組み合わせ
- 微調整された LLM の使用
- RAG ワークフローの使用

医療ドメインタスクの LLMs の既知の制限とユースケースを評価することで、タスクに最適なアプローチを選択できます。次の決定木は、医療 NLP タスクの LLM アプローチを選択するのに役立ちます。



この図表は、次のワークフローを示しています:

1. ヘルスケアとライフサイエンスのユースケースでは、NLP タスクに特定のドメイン知識が必要かどうかを特定します。必要に応じて、対象分野の専門家 (SMEs。
2. 一般的な LLM または医療データセットでトレーニングされたモデルを使用できる場合は、Amazon Bedrock または事前トレーニング済みの LLM で利用可能な基盤モデルを使用します。詳細については、このガイドの「LLM の選択」を参照してください。
3. Amazon Comprehend Medical のエンティティ検出およびオントロジーリンク機能がユースケースに対応する場合は、Amazon Comprehend Medical APIsを使用します。詳細については、このガイドの「Amazon Comprehend Medical の使用」を参照してください。
4. Amazon Comprehend Medical には必要なコンテキストがありますが、ユースケースをサポートしていない場合があります。たとえば、さまざまなエンティティ定義が必要になったり、膨大な数の結果を受け取ったり、カスタムエンティティが必要になったり、カスタム NLP タスクが必要になったりする場合があります。この場合、RAG アプローチを使用して Amazon Comprehend Medical にコンテキストをクエリします。詳細については、このガイドの「Amazon Comprehend Medical と大規模言語モデルの組み合わせ」を参照してください。
5. 十分な量のグラウンドトゥルースデータがある場合は、既存の LLM を微調整します。詳細については、このガイドの「カスタマイズアプローチ」を参照してください。
6. 他のアプローチが NLP タスクの目標を満たしていない場合は、RAG ソリューションを実装します。詳細については、このガイドの「カスタマイズアプローチ」を参照してください。
7. RAG ソリューションを実装した後、生成されたレスポンスが正確かどうかを評価します。詳細については、このガイドの「ヘルスケアおよびライフサイエンスアプリケーション用の LLMs の評価」を参照してください。Amazon Titan Text Embeddings モデルまたは all-MiniLM-L6-v2 などの一般的な文トランスフォーマーモデルから始めるのが一般的です。ただし、ドメインコンテキストがないため、これらのモデルはテキストの医学用語をキャプチャしない可能性があります。必要に応じて、以下の調整を検討してください。

   a. 他の埋め込みモデルを評価する

   b. ドメイン固有のデータセットを使用して埋め込みモデルを微調整する

ビジネスの成熟度に関する考慮事項

LLM ソリューションをヘルスケアやライフサイエンスアプリケーションに適応させるには、ビジネスの成熟度が不可欠です。これらの組織は、受け入れ基準に応じて、LLMs を実装する際にさまざまなレベルの複雑さに直面します。多くの場合、AI/ML リソースがない組織は、LLM ソリューションを構築するための請負業者のサポートに投資します。このような状況では、次のトレードオフを理解することが重要です。

- 高コストとメンテナンスのための高パフォーマンス – 厳格なパフォーマンス基準を満たすために、微調整された LLM またはカスタム LLMs を含む複雑なソリューションが必要になる場合があります。ただし、これにはコストとメンテナンス要件が高くなります。これらの高度なソリューションを維持するために、専門のリソースを雇用したり、請負業者と提携したりする必要がある場合があります。これにより、開発が遅くなる可能性があります。
- 低コストとメンテナンスに適したパフォーマンス - または、Amazon Bedrock や Amazon Comprehend Medical などのサービスが許容可能なパフォーマンスを提供する場合があります。これらの LLMs またはアプローチは完全な結果を提供する可能性がありますが、これらのソリューションは多くの場合、一貫した高品質の結果を提供することができます。これらのソリューションは低コストで、メンテナンスの負担を軽減します。これにより、開発が加速する可能性があります。

よりシンプルで低コストのアプローチが、受け入れ基準を満たす高品質の結果を一貫して提供する場合は、パフォーマンスの向上がコスト、メンテナンス、時間のトレードオフに値するかどうかを検討してください。ただし、シンプルなソリューションが目標パフォーマンスを大幅に下回り、組織が複雑なソリューションとそのメンテナンス要件に対する投資能力がない場合は、より多くのリソースや代替ソリューションが利用可能になるまで AI/ML 開発を延期することを検討してください。

さらに、LLM に依存する医療 NLP ソリューションの場合は、継続的なモニタリングと評価を実行することをお勧めします。時間の経過とともにユーザーからのフィードバックを評価し、定期的に評価を実施して、ソリューションが引き続きビジネス目標を達成していることを確認します。

ヘルスケアおよびライフサイエンスアプリケーション用の LLMs の評価

このセクションでは、ヘルスケアとライフサイエンスのユースケースで大規模言語モデル (LLMs) を評価するための要件と考慮事項の包括的な概要を説明します。

グラウンドトゥルースデータと SME フィードバックを使用してバイアスを軽減し、LLM 生成レスポンスの精度を検証することが重要です。このセクションでは、トレーニングおよびテストデータを収集およびキュレートするためのベストプラクティスについて説明します。また、ガードレールを実装し、データのバイアスと公平性を測定するのに役立ちます。また、テキスト分類、名前付きエンティティ認識、テキスト生成などの一般的な医療自然言語処理 (NLP) タスク、および関連する評価メトリクスについても説明します。

また、トレーニング実験フェーズとポストプロダクションフェーズ中に LLM 評価を実行するワークフローも示します。モデルモニタリングと LLM オペレーションは、この評価プロセスの重要な要素です。

医療 NLP タスクのトレーニングデータとテストデータ

医療 NLP タスクは、通常、医療法人 (PubMed など) または患者情報 (患者の訪問メモなど) を使用して、インサイトを分類、要約、生成します。医療担当者、医療管理者、技術者は、専門知識と視点が異なります。これらの医療担当者間の主観性により、トレーニングデータセットとテストデータセットが小さくなると、バイアスのリスクが生じます。このリスクを軽減するには、次のベストプラクティスをお勧めします。

- 事前トレーニング済みの LLM ソリューションを使用する場合は、十分な量のテストデータがあることを確認してください。テストデータは、実際の医療データによく似ている必要があります。タスクに応じて、20～100 レコードまでの範囲になります。
- LLM をファインチューニングする場合は、対象となる医療ドメインのさまざまな SMEs から十分な数のラベル付き (グラウンドトゥルース) レコードを収集します。一般的な開始点は、少なくとも 100 個の高品質のレコードです。ただし、タスクの複雑さと精度の許容基準を考慮すると、より多くのレコードが必要になる場合があります。
- 医療ユースケースで必要な場合は、ガードレールを実装し、データのバイアスと公平性を測定します。例えば、LLM が患者の人種的プロファイルによる誤診断を防止していることを確認してください。詳細については、このガイドの「セキュリティとガードレール」セクションを参照してください。

Anthropic などの多くの AI 研究および開発企業は、毒性を回避するために基盤モデルにガードレールを既に実装しています。毒性検出を使用して、LLMs からの入力プロンプトと出力レスポンスを確認できます。詳細については、Amazon Comprehend ドキュメントの「毒性検出」および Amazon Bedrock ドキュメントの「ガードレール」を参照してください。

生成 AI タスクでは、幻覚のリスクがあります。このリスクを軽減するには、分類などの NLP タスクを実行します。テキスト類似度メトリクスなど、より高度な手法を使用することもできます。BertScore は一般的に採用されているテキスト類似度メトリクスです。幻覚を軽減するために使用できる手法の詳細については、「大規模言語モデルにおける幻覚緩和手法の包括的な調査」を参照してください。

医療 NLP タスクのメトリクス

グラウンドトゥルースデータと SME が提供するトレーニングとテスト用のラベルを確立した後、定量化可能なメトリクスを作成できます。ストレステストや LLM 結果の確認などの定性的なプロセスによる品質チェックは、迅速な開発に役立ちます。ただし、メトリクスは将来の LLM オペレーションをサポートする定量的ベンチマークとして機能し、各本番リリースのパフォーマンスベンチマークとして機能します。

医療タスクを理解することは重要です。メトリクスは通常、次のいずれかの一般的な NLP タスクにマッピングされます。

- テキスト分類 – LLM は、入力プロンプトと提供されたコンテキストに基づいて、テキストを 1 つ以上の事前定義されたカテゴリに分類します。たとえば、ペインスケールを使用してペインカテゴリを分類します。テキスト分類メトリクスの例は次のとおりです。
  - 精度
  - 精度、マクロ精度とも呼ばれます
  - マクロリコールとも呼ばれるリコール
  - マクロ F1 スコアとも呼ばれる F1 スコア
  - ハミング損失
- 名前付きエンティティ認識 (NER) – テキスト抽出とも呼ばれる名前付きエンティティ認識は、非構造化テキストで記述されている名前付きエンティティを検索して事前定義されたカテゴリに分類するプロセスです。たとえば、患者レコードから薬剤名を抽出します。NER メトリクスの例は次のとおりです。
  - 精度
  - 精度

  - リコール
  - F1 スコア
  - ハミング損失
- 生成 – LLM は、プロンプトと提供されたコンテキストを処理して新しいテキストを生成します。生成には、要約タスクまたは質疑応答タスクが含まれます。生成メトリクスの例は次のとおりです。
  - ギスト評価のためのリコール指向の研究中 (ROUGE)
  - 明示的な ORdering を使用した翻訳の評価メトリクス (METEOR)
  - 研究中のバイリンガル評価 (BLEU) (翻訳用)
  - コサイン類似度とも呼ばれる文字列距離

ヘルスケアとライフサイエンスのユースケースに関するよくある質問

以下は、医療 NLP タスクでの Amazon Comprehend Medical または LLMs の使用に関するよくある質問です。

Amazon Comprehend Medical と LLM のどちらを選択するか

医療テキスト内の医療エンティティを検出するタスクの場合は、Amazon Comprehend Medical のドキュメントを参照して、抽出できる医療エンティティと、ユースケースに対応するオントロジーがあるかどうかを確認してください。そうでない場合は、LLM の使用を検討してください。詳細については、このガイドの「Amazon Comprehend Medical のユースケース」および「LLM のユースケース」を参照してください。

Amazon Comprehend Medical の結果を LLM に提供するにはどうすればよいですか?

Amazon Comprehend Medical の結果を LLM プロンプト内のコンテキストとして組み込むことができます。これにより、LLM に追加の医療知識と用語が提供されます。提供されたコンテキストは、エンティティ認識、要約、質問への回答などのタスクに対する LLM のパフォーマンスを向上させることができます。このガイドでは、Amazon Comprehend Medical の結果を使用してプロンプトを構築する方法の例をいくつか紹介します。詳細については、このガイドの「Amazon Comprehend Medical と大規模言語モデルの組み合わせ」を参照してください。

LLMs で Amazon Comprehend Medical を使用する際のベストプラクティスは何ですか?

Amazon Comprehend Medical 信頼スコアを使用して、プロンプト内のエンティティをフィルタリングまたは優先順位付けすることをお勧めします。また、特定のデータに対するパフォーマンスを評価し、エンティティ定義が要件と一致していることを確認することも重要です。Amazon Comprehend Medical をドメイン固有のナレッジソースと組み合わせることで、LLM のパフォーマンスをさらに向上させることができます。詳細については、このガイドの「RAG ワークフローで Amazon Comprehend Medical を使用するためのベストプラクティス」を参照してください。

医療ユースケースでは、事前トレーニング済みの医療 LLM を使用するか、一般的な LLM を微調整する必要がありますか?

決定は、特定の要件と高品質のトレーニングデータの可用性によって異なります。事前トレーニング済みの医療 LLMs は、良い出発点となります。ただし、ドメイン固有のデータを使用して微調整する必要がある場合があります。十分なラベル付きデータがある場合は、一般的な LLM の微調整が有効なオプションである可能性があります。詳細については、このガイドの「LLM の選択」および「ヘルスケアとライフサイエンスの NLP アプローチの選択」を参照してください。

医療 NLP タスクの LLMs のパフォーマンスを評価するにはどうすればよいですか?

テキスト分類や名前付きエンティティ認識タスクには、精度、精度、再現率、F1 スコアなどの定量的メトリクスを使用することをお勧めします。テキスト生成タスクには ROUGE と METEOR を使用できます。対象分野のエキスパートがラベル付けした信頼性の高いグラウンドトゥルースデータを用意し、時間の経過とともにモデルのパフォーマンスをモニタリングするプロセスを実装することが重要です。詳細については、このガイドの「ヘルスケアおよびライフサイエンスアプリケーション用の LLMs の評価」を参照してください。

高複雑度と低複雑度の LLM ソリューションのトレードオフは何ですか?

LLM の微調整またはカスタム LLM の構築は、非常に複雑なソリューションです。これらのアプローチはパフォーマンスを向上させることができますが、コストとメンテナンス要件は高くなります。事前トレーニング済みの LLMs や Amazon Comprehend Medical を使用するなど、よりシンプルなソリューションは、低コストでより高速な開発サイクルで許容可能なパフォーマンスを提供する可能性があります。ただし、これらのアプローチは、一部のユースケースで厳格な精度要件を満たしていない場合があります。詳細については、このガイドの「ビジネスの成熟度に関する考慮事項」を参照してください。

次のステップとリソース

このガイドは AWS のサービス 、 を使用して、実稼働環境での実際のアプリケーションの医療 NLP および生成 AI タスクを自動化するのに役立ちます。ここでは、Amazon Comprehend Medical、Amazon Bedrock でサポートされている LLMs、事前トレーニング済みの医療 LLMs、または微調整された LLMs を使用して、ヘルスケアとライフサイエンスのビジネス目標を達成する方法について説明します。このガイドでは、以下のアプローチの利点と制限について説明します。

- Amazon Comprehend Medical を個別に使用する
- Amazon Comprehend Medical の結果を LLM に提供する
- 取得拡張生成 (RAG) アプローチでの事前トレーニング済みの一般的な LLM または医療 LLM の使用
- 一般的な LLM または医療 LLM の微調整

このガイドのディシジョンツリーとビジネス成熟度に関する考慮事項を使用して、組織の AI/ML 成熟度レベルに基づいてこれらのアプローチから選択します。Amazon Comprehend Medical および Amazon Bedrock LLMs は強力な機能を提供しますが、それらを適切に実装して評価する場合にのみ成功します。このガイドで説明されている評価情報とメトリクスを使用して、ソリューションのパフォーマンスを検証します。

次のステップでは、ヘルスケア IT マネージャー、アーキテクト、技術リーダーが AI/ML 実務者と連携し、NLP 医療タスクを特定することをお勧めします。このガイドを使用して開発パスを選択し、適切な AWS のサービス および 機能を使用して自動化ソリューションを正常に実装します AWS。

AWS リソース

- Amazon Comprehend Medical ドキュメント:
  - デベロッパーガイド
  - API リファレンス
- Amazon Bedrock ドキュメント
  - Amazon Bedrock モデル評価
  - Amazon Bedrock でのファインチューニング
- Amazon SageMaker AI でモデルをファインチューニングする
- Amazon SageMaker Ground Truth

- Amazon Comprehend 毒性検出
- AWS ヘルスケアコンピテンシーパートナー

その他のリソース

- Medical-LLM リーダーボードを開く
- ヘルスケアの大規模言語モデルの調査: データ、テクノロジー、アプリケーションから説明責任と健全性へ
- 大規模言語モデルは不十分な医療コーダー — 医療コードクエリのベンチマーク
- From Beginner to Expert: Modeling Medical Knowledge into General LLMs

寄稿者

オーサリング

- Joeking、 AWS シニアデータサイエンティスト
- Ankith Ede、 AWS ソリューションアーキテクト
- Clement Perrot、 AWS シニア生成 AI ストラテジスト
- Jillian Forde、 AWS シニアソリューションアーキテクト
- Rajesh Sitaraman、 AWS シニアデリバリーコンサルタント
- Ross Claytor、 AWS プリンシパル応用サイエンティスト
- Shivesh Ummat、 AWS ソリューションアーキテクト

レビューアー

- Dilshad Raihan Akkam Veettil、 AWS シニアデータサイエンティスト
- Joseph Cottingham、 AWS 深層学習アーキテクト

テクニカルライター

- " AbouHarb, AWS Senior Technical Writer

ドキュメント履歴

以下の表は、本ガイドの重要な変更点について説明したものです。今後の更新に関する通知を受け取る場合は、RSS フィードをサブスクライブできます。

変更	説明	日付
新しいセクション	ヘルスケアセクションの「大規模言語モデルのファインチューニング」と「プロンプトエンジニアリング」セクションを追加しました。	2025 年 12 月 5 日
初版発行	—	2024 年 12 月 16 日

AWS 規範ガイダンスの用語集

以下は、 AWS 規範ガイダンスによって提供される戦略、ガイド、パターンで一般的に使用される用語です。エントリを提案するには、用語集の最後のフィードバックの提供リンクを使用します。

数字

7 Rs

アプリケーションをクラウドに移行するための 7 つの一般的な移行戦略。これらの戦略は、ガートナーが 2011 年に特定した 5 Rs に基づいて構築され、以下で構成されています。

- リファクタリング/アーキテクチャの再設計 — クラウドネイティブ特徴を最大限に活用して、俊敏性、パフォーマンス、スケーラビリティを向上させ、アプリケーションを移動させ、アーキテクチャを変更します。これには、通常、オペレーティングシステムとデータベースの移植が含まれます。例: オンプレミスの Oracle データベースを Amazon Aurora PostgreSQL 互換エディションに移行する。
- リプラットフォーム (リフトアンドリシェイプ) — アプリケーションをクラウドに移行し、クラウド機能を活用するための最適化レベルを導入します。例: お客様のオンプレミスの Oracle データベースを AWS クラウドの Oracle 用の Amazon Relational Database Service (Amazon RDS) に移行する。
- 再購入 (ドロップアンドショップ) — 通常、従来のライセンスから SaaS モデルに移行して、別の製品に切り替えます。例: 顧客関係管理 (CRM) システムを Salesforce.com に移行する。
- リホスト (リフトアンドシフト) — クラウド機能を活用するための変更を加えずに、アプリケーションをクラウドに移行します。例: お客様のオンプレミスの Oracle データベースを AWS クラウドの EC2 インスタンス上の Oracle に移行する。
- 再配置 (ハイパーバイザーレベルのリフトアンドシフト) – 新しいハードウェアを購入したり、アプリケーションを書き換えたり、既存の運用を変更したりすることなく、インフラストラクチャをクラウドに移行できます。オンプレミスプラットフォームから同じプラットフォームのクラウドサービスにサーバーを移行します。例： Microsoft Hyper-Vアプリケーションを に移行します AWS。
- 保持 (再アクセス) — アプリケーションをお客様のソース環境で保持します。これには、主要なリファクタリングを必要とするアプリケーションや、お客様がその作業を後日まで延期したいアプリケーション、およびそれらを移行するためのビジネス上の正当性がないため、お客様が保持するレガシーアプリケーションなどがあります。
- 廃止 — お客様のソース環境で不要になったアプリケーションを停止または削除します。

A

ABAC

「属性ベースのアクセス制御」をご覧ください。

抽象化されたサービス

「マネージドユーザー」をご覧ください。

ACID

「原子性、一貫性、分離性、耐久性 (ACID)」をご覧ください。

アクティブ/アクティブ移行

(双方向レプリケーションツールまたは二重書き込み操作を使用して) ソースデータベースとターゲットデータベースを同期させ、移行中に両方のデータベースが接続アプリケーションからのトランザクションを処理するデータベース移行方法。この方法では、1 回限りのカットオーバーの必要がなく、管理された小規模なバッチで移行できます。アクティブ/パッシブ移行よりも柔軟な方法ですが、さらに多くの作業が必要となります。

アクティブ/パッシブ移行

ソースデータベースとターゲットデータベースを同期させながら、データがターゲットデータベースにレプリケートされている間、接続しているアプリケーションからのトランザクションをソースデータベースのみで処理するデータベース移行方法。移行中、ターゲットデータベースはトランザクションを受け付けません。

集計関数

複数行に処理を行い、グループ全体を対象に単一の戻り値を計算する SQL 関数。集計関数の例としては、SUM や MAX などがあります。

AI

「人工知能」をご覧ください。

AIOps

「AI オペレーション」をご覧ください。

匿名化

データセット内の個人情報を完全に削除するプロセス。匿名化は個人のプライバシー保護に役立ちます。匿名化されたデータは、もはや個人データとは見なされません。

アンチパターン

繰り返し起こる問題に対して頻繁に用いられる解決策で、その解決策が逆効果であったり、効果がなかったり、代替案よりも効果が低かったりするもの。

アプリケーション制御

マルウェアからシステムを保護するために、承認されたアプリケーションのみを使用できるようにするセキュリティアプローチ。

アプリケーションポートフォリオ

アプリケーションの構築と維持にかかるコスト、およびそのビジネス価値を含む、組織が使用する各アプリケーションに関する詳細情報の集まり。この情報は、ポートフォリオの検出と分析プロセス の重要な要素であり、移行、モダナイズ、最適化するアプリケーションを特定し、優先順位を付けるのに役立ちます。

人工知能 (AI)

コンピューティングテクノロジーを使用し、学習、問題の解決、パターンの認識など、通常は人間に関連づけられる認知機能の実行に特化したコンピュータサイエンスの分野。詳細については、「人工知能 (AI) とは何ですか?」をご覧ください。

AI オペレーション (AIOps)

機械学習技術を使用して運用上の問題を解決し、運用上のインシデントと人の介入を減らし、サービス品質を向上させるプロセス。 AWS 移行戦略での AIOps の使用方法については、オペレーション統合ガイド を参照してください。

非対称暗号化

暗号化用のパブリックキーと復号用のプライベートキーから成る 1 組のキーを使用した、暗号化のアルゴリズム。パブリックキーは復号には使用されないため共有しても問題ありませんが、プライベートキーの利用は厳しく制限する必要があります。

原子性、一貫性、分離性、耐久性 (ACID)

エラー、停電、その他の問題が発生した場合でも、データベースのデータ有効性と運用上の信頼性を保証する一連のソフトウェアプロパティ。

属性ベースのアクセス制御 (ABAC)

部署、役職、チーム名など、ユーザーの属性に基づいてアクセス許可をきめ細かく設定する方法。詳細については、 AWS Identity and Access Management (IAM) ドキュメントの「 の ABAC AWS」を参照してください。

信頼できるデータソース

最も信頼性のある情報源とされるデータのプライマリーバージョンを保存する場所。匿名化、編集、仮名化など、データを処理または変更する目的で、信頼できるデータソースから他の場所にデータをコピーすることができます。

アベイラビリティーゾーン (AZ)

他のアベイラビリティーゾーンの障害から AWS リージョン 隔離され、同じリージョン内の他のアベイラビリティーゾーンへの低コストで低レイテンシーのネットワーク接続を提供する 内の別の場所。

AWS クラウド導入フレームワーク (AWS CAF)

組織がクラウドへの移行を成功させるための効率的で効果的な計画を立て AWS るための、 のガイドラインとベストプラクティスのフレームワークです。 AWS CAF は、ビジネス、人材、ガバナンス、プラットフォーム、セキュリティ、運用という 6 つの重点分野にガイダンスを整理しています。ビジネス、人材、ガバナンスの観点では、ビジネススキルとプロセスに重点を置き、プラットフォーム、セキュリティ、オペレーションの視点は技術的なスキルとプロセスに焦点を当てています。例えば、人材の観点では、人事 (HR)、人材派遣機能、および人材管理を扱うステークホルダーを対象としています。この観点から、 AWS CAF は、クラウド導入を成功させるための組織の準備に役立つ人材開発、トレーニング、コミュニケーションに関するガイダンスを提供します。詳細については、AWS CAF ウェブサイトと AWS CAF のホワイトペーパー を参照してください。

AWS ワークロード認定フレームワーク (AWS WQF)

データベース移行ワークロードを評価し、移行戦略を推奨し、作業見積もりを提供するツール。AWS WQF は AWS Schema Conversion Tool (AWS SCT) に含まれています。データベーススキーマとコードオブジェクト、アプリケーションコード、依存関係、およびパフォーマンス特性を分析し、評価レポートを提供します。

B

不正なボット

個人や組織に混乱や損害を与えることを目的としたボット。

BCP

「ビジネス継続性計画 (BCP)」をご覧ください。

動作グラフ

リソースの動作とインタラクションを経時的に示した、一元的なインタラクティブビュー。Amazon Detective の動作グラフを使用すると、失敗したログオンの試行、不審な API 呼び出し、その他同様のアクションを調べることができます。詳細については、Detective ドキュメントの「動作グラフのデータ」を参照してください。

ビッグエンディアンシステム

最上位バイトを最初に格納するシステム。「エンディアン性」もご覧ください。

二項分類

バイナリ結果 (2 つの可能なクラスのうちの 1 つ) を予測するプロセス。例えば、お客様の機械学習モデルで「この E メールはスパムですか、それともスパムではありませんか」などの問題を予測する必要があるかもしれません。または「この製品は書籍ですか、車ですか」などの問題を予測する必要があるかもしれません。

ブルームフィルター

要素がセットのメンバーであるかどうかをテストするために使用される、確率的でメモリ効率の高いデータ構造。

ブルー/グリーンデプロイ

それぞれが独立しているが、同一の環境を 2 つ作成するデプロイ戦略。現在のアプリケーションバージョンを 1 つの環境 (ブルー) で実行し、新しいアプリケーションバージョンを別の環境 (グリーン) で実行します。この戦略は、最小限の影響で迅速にロールバックするのに役立ちます。

ボット

インターネット経由で自動タスクを実行し、人間のアクティビティややり取りをシミュレートするソフトウェアアプリケーション。インターネット上の情報のインデックスを作成するウェブクローラーなど、一部のボットは有用または有益です。悪質なボットと呼ばれる他のボットの中には、個人や組織を混乱させたり、損害を与えたりすることを意図したものもあります。

ボットネット

マルウェアに感染しており、ボットハーダーまたはボットオペレーターと呼ばれる単一の当事者によって制御されているボットのネットワーク。ボットネットは、ボットとその影響力を拡大する仕組みとして、非常によく知られています。

ブランチ

コードリポジトリに含まれる領域。リポジトリに最初に作成するブランチは、メインブランチといいます。既存のブランチから新しいブランチを作成し、その新しいブランチで機能を開発した

り、バグを修正したりできます。機能を構築するために作成するブランチは、通常、機能ブランチと呼ばれます。機能をリリースする準備ができたら、機能ブランチをメインブランチに統合します。詳細については、「ブランチの概要」(GitHub ドキュメント) を参照してください。

ブレークグラスアクセス

例外的な状況では、承認されたプロセスを通じて、ユーザーが AWS アカウント 通常アクセス許可を持たない にすばやくアクセスできるようにします。詳細については、 AWS Well-Architected ガイダンスの「ブレークグラス手順の実装」インジケータを参照してください。

ブラウンフィールド戦略

環境の既存インフラストラクチャ。システムアーキテクチャにブラウンフィールド戦略を導入する場合、現在のシステムとインフラストラクチャの制約に基づいてアーキテクチャを設計します。既存のインフラストラクチャを拡張している場合は、ブラウンフィールド戦略とグリーンフィールド戦略を融合させることもできます。

バッファキャッシュ

アクセス頻度が最も高いデータが保存されるメモリ領域。

ビジネス能力

価値を生み出すためにビジネスが行うこと (営業、カスタマーサービス、マーケティングなど)。マイクロサービスのアーキテクチャと開発の決定は、ビジネス能力によって推進できます。詳細については、AWSでのコンテナ化されたマイクロサービスの実行ホワイトペーパーの「ビジネス機能を中心に組織化」セクションを参照してください。

ビジネス継続性計画 (BCP)

大規模移行など、中断を伴うイベントが運用に与える潜在的な影響に対処し、ビジネスを迅速に再開できるようにする計画。

C

CAF

「AWS クラウド導入フレームワーク」を参照してください

カナリアデプロイ

エンドユーザーへのバージョンリリースを、時間をかけて段階的に行うこと。確信が持てたら新規バージョンをデプロイして、現在のバージョン全体を置き換えます。

CCoE

「Cloud Center of Excellence」を参照してください。

CDC

「変更データキャプチャ」を参照してください。

変更データキャプチャ (CDC)

データソース (データベーステーブルなど) の変更を追跡し、その変更に関するメタデータを記録するプロセス。CDC は、ターゲットシステムでの変更を監査またはレプリケートして同期を維持するなど、さまざまな目的に使用できます。

カオスエンジニアリング

障害や破壊的なイベントを意図的に導入して、システムの耐障害性をテストすること。AWS Fault Injection Service (AWS FIS) を使用して、 AWS ワークロードにストレスを与え、その応答を評価する実験を実行できます。

CI/CD

「継続的インテグレーションと継続的デリバリー」を参照してください。

分類

予測を生成するのに役立つ分類プロセス。分類問題の機械学習モデルは、離散値を予測します。離散値は、常に互いに区別されます。例えば、モデルがイメージ内に車があるかどうかを評価する必要がある場合があります。

クライアント側の暗号化

ターゲットがデータ AWS のサービス を受信する前のローカルでのデータの暗号化。

Cloud Center of Excellence (CCoE)

クラウドのベストプラクティスの作成、リソースの移動、移行のタイムラインの確立、大規模変革を通じて組織をリードするなど、組織全体のクラウド導入の取り組みを推進する学際的なチーム。詳細については、 AWS クラウド エンタープライズ戦略ブログの CCoE 投稿を参照してください。

クラウドコンピューティング

リモートデータストレージと IoT デバイス管理に通常使用されるクラウドテクノロジー。クラウドコンピューティングは、一般的に、エッジコンピューティングに接続されています。

クラウド運用モデル

IT 組織において、1 つ以上のクラウド環境を構築、成熟、最適化するために使用される運用モデル。詳細については、「クラウド運用モデルの構築」を参照してください。

導入のクラウドステージ

組織が、 AWS クラウドへの移行時に通常実行する 4 つの段階。

- プロジェクト — 概念実証と学習を目的として、クラウド関連のプロジェクトをいくつか実行する
- 基礎固め — お客様のクラウドの導入を拡大するための基礎的な投資 (ランディングゾーンの作成、CCoE の定義、運用モデルの確立など)
- 移行 — 個々のアプリケーションの移行
- 再発明 — 製品とサービスの最適化、クラウドでのイノベーション

これらのステージは、 AWS クラウド エンタープライズ戦略ブログのブログ記事「クラウドファーストへのジャーニー」と「導入のステージ」で Stephen Orban によって定義されました。移行戦略との関連性については、 AWS 「移行準備ガイド」を参照してください。

CMDB

「構成管理データベース (CMDB)」を参照してください。

コードリポジトリ

ソースコードやその他の資産 (ドキュメント、サンプル、スクリプトなど) が保存され、バージョン管理プロセスを通じて更新される場所。一般的なクラウドリポジトリには、GitHub や Bitbucket Cloud があります。コードの各バージョンはブランチと呼ばれます。マイクロサービスの構造では、各リポジトリは 1 つの機能専用です。1 つの CI/CD パイプラインで複数のリポジトリを使用できます。

コールドキャッシュ

空である、または、かなり空きがある、もしくは、古いデータや無関係なデータが含まれているバッファキャッシュ。データベースインスタンスはメインメモリまたはディスクから読み取る必要があり、バッファキャッシュから読み取るよりも時間がかかるため、パフォーマンスに影響します。

コールドデータ

めったにアクセスされず、通常は過去のデータです。この種類のデータをクエリする場合、通常は低速なクエリでも問題ありません。このデータを低パフォーマンスで安価なストレージ階層またはクラスに移動すると、コストを削減することができます。

コンピュータビジョン (CV)

機械学習を使用してデジタルイメージやビデオといった、ビジュアル形式の情報を分析および抽出する AI の分野。例えば、Amazon SageMaker AI では、CV 用の画像処理アルゴリズムを利用できます。

設定ドリフト

ワークロードにおいて、設定が想定した状態から変化すること。これによって、ワークロードが非準拠になる可能性がありますが、この状態は、徐々に生じ、意図的なものではありません。

構成管理データベース (CMDB)

データベースとその IT 環境 (ハードウェアとソフトウェアの両方のコンポーネントとその設定を含む) に関する情報を保存、管理するリポジトリ。通常、CMDB のデータは、移行のポートフォリオの検出と分析の段階で使用します。

コンフォーマンスパック

コンプライアンスチェックとセキュリティチェックをカスタマイズするためにアセンブルできる AWS Config ルールと修復アクションのコレクション。YAML テンプレートを使用して、コンフォーマンスパックを AWS アカウント および リージョンの単一のエンティティとしてデプロイすることも、組織全体にデプロイすることもできます。詳細については、 AWS Config ドキュメントの「コンフォーマンスパック」を参照してください。

継続的インテグレーションと継続的デリバリー (CI/CD)

ソフトウェアリリースプロセスのソース、ビルド、テスト、ステージング、本番の各ステージを自動化するプロセス。CI/CD は一般的にパイプラインと呼ばれます。プロセスの自動化、生産性の向上、コード品質の向上、配信の加速化を可能にします。詳細については、「継続的デリバリーの利点」を参照してください。CD は継続的デプロイ (Continuous Deployment) の略語でもあります。詳細については「継続的デリバリーと継続的なデプロイ」を参照してください。

CV

「コンピュータビジョン」を参照してください。

D

保管中のデータ

ストレージ内にあるデータなど、常に自社のネットワーク内にあるデータ。

データ分類

ネットワーク内のデータを重要度と機密性に基づいて識別、分類するプロセス。データに適した保護および保持のコントロールを判断する際に役立つため、あらゆるサイバーセキュリティのリスク管理戦略において重要な要素です。データ分類は、 AWS Well-Architected フレームワークのセキュリティの柱のコンポーネントです。詳細については、「データ分類」を参照してください。

データドリフト

実稼働データと ML モデルのトレーニングに使用されたデータとの間に有意な差異が生じたり、入力データが時間の経過と共に有意に変化したりすることです。データドリフトは、ML モデル予測の全体的な品質、精度、公平性を低下させる可能性があります。

転送中のデータ

ネットワーク内 (ネットワークリソース間など) を活発に移動するデータ。

データメッシュ

非一元的で分散型のデータ所有権を持つとともに、一元的な管理およびガバナンスを行えるアーキテクチャフレームワーク。

データ最小化

厳密に必要なデータのみを収集し、処理するという原則。でデータ最小化を実践 AWS クラウドすることで、プライバシーリスク、コスト、分析のカーボンフットプリントを削減できます。

データ境界

AWS 環境内の一連の予防ガードレール。信頼された ID のみが、期待されるネットワークから信頼されたリソースにアクセスできるようにします。詳細については、「 でのデータ境界の構築 AWS」を参照してください。

データの前処理

raw データをお客様の機械学習モデルで簡単に解析できる形式に変換すること。データの前処理とは、特定の列または行を削除して、欠落している、矛盾している、または重複する値に対処することを意味します。

データ出所

データの生成、送信、保存の方法など、データのライフサイクル全体を通じてデータの出所と履歴を追跡するプロセス。

データ件名

データを収集、処理している個人。

データウェアハウス

分析などのビジネスインテリジェンスをサポートするデータ管理システム。データウェアハウスには、一般的に、大量の履歴データが含まれており、多くの場合、それらはクエリや分析に使用されます。

データベース定義言語 (DDL)

データベース内のテーブルやオブジェクトの構造を作成または変更するためのステートメントまたはコマンド。

データベース操作言語 (DML)

データベース内の情報を変更 (挿入、更新、削除) するためのステートメントまたはコマンド。

DDL

「データベース定義言語」を参照してください。

ディープアンサンブル

予測のために複数の深層学習モデルを組み合わせます。ディープアンサンブルを使用して、より正確な予測を取得したり、予測の不確実性を推定したりできます。

深層学習

人工ニューラルネットワークの複数層を使用して、入力データと対象のターゲット変数の間のマッピングを識別する機械学習サブフィールド。

多層防御

一連のセキュリティメカニズムとコントロールをコンピュータネットワーク全体に層状に重ねて、ネットワークとその内部にあるデータの機密性、整合性、可用性を保護する情報セキュリティの手法。この戦略を に採用するときは AWS、リソースの保護に役立つように、 AWS Organizations 構造の異なるレイヤーに複数のコントロールを追加します。たとえば、多層防御アプローチでは、多要素認証、ネットワークセグメンテーション、暗号化を組み合わせることができます。

委任管理者

では AWS Organizations、互換性のあるサービスが AWS メンバーアカウントを登録して組織のアカウントを管理し、そのサービスのアクセス許可を管理できます。このアカウントを、そのサービスの委任管理者と呼びます。詳細、および互換性のあるサービスの一覧は、 AWS

Organizations ドキュメントの「AWS Organizationsで使用できるサービス」を参照してください。

トラブルシューティング

アプリケーション、新機能、コードの修正をターゲットの環境で利用できるようにするプロセス。デプロイでは、コードベースに変更を施した後、アプリケーションの環境でそのコードベースを構築して実行します。

開発環境

「環境」を参照してください。

検出管理

イベントが発生したときに、検出、ログ記録、警告を行うように設計されたセキュリティコントロール。これらのコントロールは副次的な防衛手段であり、実行中の予防的コントロールをすり抜けたセキュリティイベントをユーザーに警告します。詳細については、「 AWSでのセキュリティコントロールの実装」の「検出的コントロール」を参照してください。

開発バリューストリームマッピング (DVSM)

ソフトウェア開発ライフサイクルのスピードと品質に悪影響を及ぼす制約を特定し、優先順位を付けるために使用されるプロセス。DVSM は、もともとリーンマニュファクチャリング・プラクティスのために設計されたバリューストリームマッピング・プロセスを拡張したものです。ソフトウェア開発プロセスを通じて価値を創造し、動かすために必要なステップとチームに焦点を当てています。

デジタルツイン

建物、工場、産業機器、生産ラインなど、現実世界のシステムを仮想的に表現したものです。デジタルツインは、予知保全、リモートモニタリング、生産最適化をサポートします。

ディメンションテーブル

スタースキーマにおいて、ファクトテーブルの定量データに関するデータ属性が含まれる小さいテーブル。ディメンションテーブルの属性は、通常、テキストフィールド、またはテキストのように扱える個別の数値で示されます。これらの属性は、一般的に、クエリの制約、フィルタリング、結果セットのラベル付けに使用されます。

ディザスタ

ワークロードまたはシステムが、導入されている主要な場所でのビジネス目標の達成を妨げるイベント。これらのイベントは、自然災害、技術的障害、または意図しない設定ミスやマルウェア攻撃などの人間の行動の結果である場合があります。

ディザスタリカバリ (DR)

ディザスタによるダウンタイムとデータ損失を最小限に抑えるための戦略とプロセス。詳細については、 AWS Well-Architected フレームワークの「 でのワークロードのディザスタリカバリ AWS: クラウドでのリカバリ」を参照してください。

DML

「データベース操作言語」を参照してください。

ドメイン駆動型設計

各コンポーネントが提供している変化を続けるドメイン、またはコアビジネス目標にコンポーネントを接続して、複雑なソフトウェアシステムを開発するアプローチ。この概念は、エリック・エヴァンスの著書、Domain-Driven Design: Tackling Complexity in the Heart of Software (ドメイン駆動設計:ソフトウェアの中心における複雑さへの取り組み) で紹介されています (ボストン: Addison-Wesley Professional、2003)。strangler fig パターンでドメイン駆動型設計を使用する方法の詳細については、「コンテナと Amazon API Gateway を使用して、従来の Microsoft ASP.NET (ASMX) ウェブサービスを段階的にモダナイズ」を参照してください。

DR

「ディザスタリカバリ」を参照してください。

ドリフト検出

ベースライン設定からの偏差を追跡します。たとえば、 AWS CloudFormation を使用してシステムリソースのドリフトを検出したり、 を使用して AWS Control Tower 、ガバナンス要件への準拠に影響するランディングゾーンの変更を検出したりできます。

DVSM

「開発バリューストリームマッピング」を参照してください。

E

EDA

「探索的データ分析」を参照してください。

EDI

「電子データ交換」を参照してください。

エッジコンピューティング

IoT ネットワークのエッジにあるスマートデバイスの計算能力を高めるテクノロジー。クラウドコンピューティングと比較すると、エッジコンピューティングは通信レイテンシーを短縮し、応答時間を改善できます。

電子データ交換 (EDI)

組織間で行う、ビジネスドキュメントの自動交換。詳細については、「電子データ交換とは」を参照してください。

暗号化

人間が読み取り可能なプレーンテキストデータを暗号文に変換するコンピューティング処理。

暗号化キー

暗号化アルゴリズムが生成した、ランダム化されたビットからなる暗号文字列。キーの長さは決まっておらず、各キーは予測できないように、一意になるように設計されています。

エンディアン

コンピュータメモリにバイトが格納される順序。ビッグエンディアンシステムでは、最上位バイトが最初に格納されます。リトルエンディアンシステムでは、最下位バイトが最初に格納されます。

エンドポイント

「サービスエンドポイント」を参照してください。

エンドポイントサービス

仮想プライベートクラウド (VPC) 内でホストして、他のユーザーと共有できるサービス。を使用してエンドポイントサービスを作成し AWS PrivateLink 、他の AWS アカウント または AWS Identity and Access Management (IAM) プリンシパルにアクセス許可を付与できます。これらのアカウントまたはプリンシパルは、インターフェイス VPC エンドポイントを作成することで、エンドポイントサービスにプライベートに接続できます。詳細については、Amazon Virtual Private Cloud (Amazon VPC) ドキュメントの「エンドポイントサービスを作成する」を参照してください。

エンタープライズリソースプランニング (ERP)

エンタープライズの主要なビジネスプロセス (会計、MES、プロジェクト管理など) を自動化および管理するシステム。

エンベロープ暗号化

暗号化キーを、別の暗号化キーを使用して暗号化するプロセス。詳細については、 AWS Key Management Service (AWS KMS) ドキュメントの「エンベロープ暗号化」を参照してください。

環境

実行中のアプリケーションのインスタンス。クラウドコンピューティングにおける一般的な環境の種類は以下のとおりです。

- 開発環境 — アプリケーションのメンテナンスを担当するコアチームのみが使用できる、実行中のアプリケーションのインスタンス。開発環境は、上位の環境に昇格させる変更をテストするときに使用します。このタイプの環境は、テスト環境と呼ばれることもあります。
- 下位環境 — 初期ビルドやテストに使用される環境など、アプリケーションのすべての開発環境。
- 本番環境 — エンドユーザーがアクセスできる、実行中のアプリケーションのインスタンス。CI/CD パイプラインでは、本番環境が最後のデプロイ環境になります。
- 上位環境 — コア開発チーム以外のユーザーがアクセスできるすべての環境。これには、本番環境、本番前環境、ユーザー承認テスト環境などが含まれます。

エピック

アジャイル方法論で、お客様の作業の整理と優先順位付けに役立つ機能カテゴリ。エピックでは、要件と実装タスクの概要についてハイレベルな説明を提供します。例えば、 AWS CAF セキュリティエピックには、ID とアクセスの管理、検出コントロール、インフラストラクチャセキュリティ、データ保護、インシデント対応が含まれます。 AWS 移行戦略のエピックの詳細については、プログラム実装ガイドを参照してください。

ERP

「エンタープライズリソース計画」を参照してください。

探索的データ分析 (EDA)

データセットを分析してその主な特性を理解するプロセス。お客様は、データを収集または集計してから、パターンの検出、異常の検出、および前提条件のチェックのための初期調査を実行します。EDA は、統計の概要を計算し、データの可視化を作成することによって実行されます。

F

ファクトテーブル

スタースキーマの中央にあるテーブル。ビジネスオペレーションに関する定量的データが保存されます。一般的に、ファクトテーブルは、2 種類の列で構成されます。1 つは測定値が含まれる列、もう 1 つはディメンションテーブルへの外部キーが含まれる列です。

フェイルファスト

開発ライフサイクルを短縮するために、頻繁かつ段階的にテストを行う哲学であり、アジャイルアプローチでは、この考え方がきわめて重要です。

障害分離境界

では AWS クラウド、障害の影響を制限し、ワークロードの耐障害性を高めるのに役立つアベイラビリティーゾーン AWS リージョン、コントロールプレーン、データプレーンなどの境界。詳細については、「AWS 障害分離境界」を参照してください。

機能ブランチ

「ブランチ」を参照してください。

特徴量

お客様が予測に使用する入力データ。例えば、製造コンテキストでは、特徴量は製造ラインから定期的にキャプチャされるイメージの可能性もあります。

特徴量重要度

モデルの予測に対する特徴量の重要性。これは通常、Shapley Additive Deskonations (SHAP) や積分勾配など、さまざまな手法で計算できる数値スコアで表されます。詳細については、「 を使用した機械学習モデルの解釈可能性 AWS」を参照してください。

機能変換

追加のソースによるデータのエンリッチ化、値のスケーリング、単一のデータフィールドからの複数の情報セットの抽出など、機械学習プロセスのデータを最適化すること。これにより、機械学習モデルはデータの恩恵を受けることができます。例えば、「2021-05-27 00:15:37」の日付を「2021 年」、「5 月」、「木」、「15」に分解すると、学習アルゴリズムがさまざまなデータコンポーネントに関連する微妙に異なるパターンを学習するのに役立ちます。

数ショットプロンプト

LLM に、タスクと望ましい出力を示す例を少数提示した後に、類似のタスクを実行させること。この手法は、プロンプトに記述された例 (ショット) からモデルが学習する「インコンテキスト学

習」の一種です。数ショットプロンプトは、特定のフォーマット、推論、専門知識が必要なタスクに効果的です。「ゼロショットプロンプト」も参照してください。

FGAC

「きめ細かなアクセス制御」を参照してください。

きめ細かなアクセス制御 (FGAC)

複数の条件を使用してアクセス要求を許可または拒否すること。

フラッシュカット移行

変更データのキャプチャによる継続的なデータ複製を利用して、段階的なアプローチではなく、可能な限り短時間でデータを移行するデータベース移行方法。目的はダウンタイムを最小限に抑えることです。

FM

「基盤モデル」を参照してください。

基盤モデル (FM)

大規模な深層学習ニューラルネットワークであり、一般化およびラベル付けされていないデータからなる大規模データセットでトレーニングされています。FM により、言語理解、テキストおよび画像生成、自然言語での会話といった、一般的な各種タスクを実行できます。詳細については、「基盤モデルとは何ですか?」を参照してください。

G

生成 AI

AI モデルのサブセット。大量のデータでトレーニングされており、シンプルなテキストプロンプトを使用して、画像、動画、テキスト、オーディオなどの新しいコンテンツやアーティファクトを作成できます。詳細については、「生成 AI とは何ですか?」を参照してください。

ジオブロッキング

「地理的制限」を参照してください。

地理的制限 (ジオブロッキング)

特定の国のユーザーがコンテンツ配信にアクセスできないようにするための、Amazon CloudFront のオプション。アクセスを許可する国と禁止する国は、許可リストまたは禁止リスト

を使って指定します。詳細については、CloudFront ドキュメントの「コンテンツの地理的ディストリビューションの制限」を参照してください。

Gitflow ワークフロー

下位環境と上位環境が、ソースコードリポジトリでそれぞれ異なるブランチを使用する方法。Gitflow ワークフローは古いと見なされている方法であり、トランクベースのワークフローは推奨されている新しい方法です。

ゴールデンイメージ

システムまたはソフトウェアのスナップショットであり、システムまたはソフトウェアの新規インスタンスをデプロイするテンプレートとして使用されます。製造の例で言えば、ゴールデンイメージを使用すると、複数のデバイスにソフトウェアをプロビジョニングして、デバイス製造オペレーションの速度、スケーラビリティ、生産性を向上させることができます。

グリーンフィールド戦略

新しい環境に既存のインフラストラクチャが存在しないこと。システムアーキテクチャにグリーンフィールド戦略を導入する場合、既存のインフラストラクチャ (別名ブラウンフィールド) との互換性の制約を受けることなく、あらゆる新しいテクノロジーを選択できます。既存のインフラストラクチャを拡張している場合は、ブラウンフィールド戦略とグリーンフィールド戦略を融合させることもできます。

ガードレール

組織単位 (OU) 全般のリソース、ポリシー、コンプライアンスを管理するのに役立つ概略的なルール。予防ガードレールは、コンプライアンス基準に一致するようにポリシーを実施します。これらは、サービスコントロールポリシーと IAM アクセス許可の境界を使用して実装されます。検出ガードレールは、ポリシー違反やコンプライアンス上の問題を検出し、修復のためのアラートを発信します。これらは AWS Config、、 AWS Security Hub CSPM、Amazon GuardDuty、、 AWS Trusted Advisor Amazon Inspector、およびカスタム AWS Lambda チェックを使用して実装されます。

H

HA

「高可用性」を参照してください。

異種混在データベースの移行

別のデータベースエンジンを使用するターゲットデータベースへお客様の出典データベースの移行 (例えば、Oracle から Amazon Aurora)。異種間移行は通常、アーキテクチャの再設計作業の一部であり、スキーマの変換は複雑なタスクになる可能性があります。AWS は、スキーマの変換に役立つ AWS SCTを提供します。

高可用性 (HA)

課題や災害が発生した場合に、介入なしにワークロードを継続的に運用できること。HA システムは、自動的にフェイルオーバーし、一貫して高品質のパフォーマンスを提供し、パフォーマンスへの影響を最小限に抑えながらさまざまな負荷や障害を処理するように設計されています。

ヒストリアンのモダナイゼーション

製造業のニーズによりよく応えるために、オペレーションテクノロジー (OT) システムをモダナイズし、アップグレードするためのアプローチ。ヒストリアンは、工場内のさまざまなソースからデータを収集して保存するために使用されるデータベースの一種です。

ホールドアウトデータ

機械学習モデルのトレーニング用データセットから保留される、ラベル付き履歴データの一部。ホールドアウトデータを使用すると、モデル予測をホールドアウトデータと比較して、モデルのパフォーマンスを評価できます。

同種データベースの移行

お客様の出典データベースを、同じデータベースエンジンを共有するターゲットデータベース (Microsoft SQL Server から Amazon RDS for SQL Server など) に移行する。同種間移行は、通常、リホストまたはリプラットフォーム化の作業の一部です。ネイティブデータベースユーティリティを使用して、スキーマを移行できます。

ホットデータ

リアルタイムデータや最近の翻訳データなど、頻繁にアクセスされるデータ。通常、このデータには高速なクエリ応答を提供する高性能なストレージ階層またはクラスが必要です。

ホットフィックス

本番環境の重大な問題を修正するために緊急で配布されるプログラム。緊急性が高いため、通常の DevOps のリリースワークフローからは外れた形で実施されます。

ハイパーケア期間

カットオーバー直後、移行したアプリケーションを移行チームがクラウドで管理、監視して問題に対処する期間。通常、この期間は 1～4 日です。ハイパーケア期間が終了すると、アプリケーションに対する責任は一般的に移行チームからクラウドオペレーションチームに移ります。

I

IaC

「Infrastructure as Code」を参照してください。

ID ベースのポリシー

AWS クラウド 環境内のアクセス許可を定義する 1 つ以上の IAM プリンシパルにアタッチされたポリシー。

アイドル状態のアプリケーション

90 日間の平均的な CPU およびメモリ使用率が 5～20% のアプリケーション。移行プロジェクトでは、これらのアプリケーションを廃止するか、オンプレミスに保持するのが一般的です。

IIoT

「インダストリアル IoT」を参照してください。

イミュータブルインフラストラクチャ

既存インフラストラクチャの更新、パッチ適用、変更などを行わずに、本番環境ワークロードに使用する新規インフラストラクチャをデプロイするモデル。本質的に、イミュータブルインフラストラクチャは、ミュータブルインフラストラクチャよりも一貫性、信頼性、予測性に優れています。詳細については、 AWS Well-Architected フレームワークにある「イミュータブルインフラストラクチャを使用してデプロイする」のベストプラクティスを参照してください。

インバウンド (受信) VPC

AWS マルチアカウントアーキテクチャでは、アプリケーションの外部からネットワーク接続を受け入れ、検査し、ルーティングする VPC。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

増分移行

アプリケーションを 1 回ですべてカットオーバーするのではなく、小さい要素に分けて移行するカットオーバー戦略。例えば、最初は少数のマイクロサービスまたはユーザーのみを新しいシステムに移行する場合があります。すべてが正常に機能することを確認できたら、残りのマイクロサービスやユーザーを段階的に移行し、レガシーシステムを廃止できるようにします。この戦略により、大規模な移行に伴うリスクが軽減されます。

インダストリー 4.0

2016 年に Klaus Schwab 氏が提唱した用語で、接続、リアルタイムデータ、オートメーション、分析、AI/ML の進歩による、ビジネスプロセスのモダナイズを意味します。

インフラストラクチャ

アプリケーションの環境に含まれるすべてのリソースとアセット。

Infrastructure as Code (IaC)

アプリケーションのインフラストラクチャを一連の設定ファイルを使用してプロビジョニングし、管理するプロセス。IaC は、新しい環境を再現可能で信頼性が高く、一貫性のあるものにするため、インフラストラクチャを一元的に管理し、リソースを標準化し、スケールを迅速に行えるように設計されています。

インダストリアル IoT (IIoT)

製造、エネルギー、自動車、ヘルスケア、ライフサイエンス、農業などの産業部門におけるインターネットに接続されたセンサーやデバイスの使用。詳細については、「インダストリアル IoT (IIoT) デジタルトランスフォーメーション戦略の構築」を参照してください。

インスペクション VPC

AWS マルチアカウントアーキテクチャでは、VPC (同一または異なる 内 AWS リージョン)、インターネット、オンプレミスネットワーク間のネットワークトラフィックの検査を管理する一元化された VPCs。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

IoT

インターネットまたはローカル通信ネットワークを介して他のデバイスやシステムと通信する、センサーまたはプロセッサが組み込まれた接続済み物理オブジェクトのネットワーク。詳細については、「IoT とは」を参照してください。

解釈可能性

機械学習モデルの特性で、モデルの予測がその入力にどのように依存するかを人間が理解できる度合いを表します。詳細については、「 を使用した機械学習モデルの解釈可能性 AWS」を参照してください。

IoT

「IoT」を参照してください。

IT 情報ライブラリ (ITIL)

IT サービスを提供し、これらのサービスをビジネス要件に合わせるための一連のベストプラクティス。ITIL は ITSM の基盤を提供します。

IT サービス管理 (ITSM)

組織の IT サービスの設計、実装、管理、およびサポートに関連する活動。クラウドオペレーションと ITSM ツールの統合については、オペレーション統合ガイドを参照してください。

ITIL

「IT 情報ライブラリ」を参照してください。

ITSM

「IT サービス管理」を参照してください。

L

ラベルベースアクセス制御 (LBAC)

強制アクセス制御 (MAC) の実装で、ユーザーとデータ自体にそれぞれセキュリティラベル値が明示的に割り当てられます。ユーザーセキュリティラベルとデータセキュリティラベルが交差する部分によって、ユーザーに表示される行と列が決まります。

ランディングゾーン

ランディングゾーンは、スケーラブルで安全な、適切に設計されたマルチアカウント AWS 環境です。これは、組織がセキュリティおよびインフラストラクチャ環境に自信を持ってワークロードとアプリケーションを迅速に起動してデプロイできる出発点です。ランディングゾーンの詳細については、「安全でスケーラブルなマルチアカウント AWS 環境のセットアップ」を参照してください。

大規模言語モデル (LLM)

大量のデータで事前トレーニングされた深層学習 AI モデル。LLM では、質問への回答、ドキュメントの要約、他言語へのテキスト翻訳、文を完成させるなど、さまざまなタスクを実行できます。詳細については、「大規模言語モデル (LLM) とは何ですか?」を参照してください。

大規模な移行

300 台以上のサーバの移行。

LBAC

「ラベルベースアクセス制御」を参照してください。

最小特権

タスクの実行には必要最低限の権限を付与するという、セキュリティのベストプラクティス。詳細については、IAM ドキュメントの「最小特権アクセス許可を適用する」を参照してください。

リフトアンドシフト

「7 Rs」を参照してください。

リトルエンディアンシステム

最下位バイトを最初に格納するシステム。「エンディアン性」もご覧ください。

LLM

「大規模言語モデル」を参照してください。

下位環境

「環境」を参照してください。

M

機械学習 (ML)

パターン認識と学習にアルゴリズムと手法を使用する人工知能の一種。ML は、モノのインターネット (IoT) データなどの記録されたデータを分析して学習し、パターンに基づく統計モデルを生成します。詳細については、「機械学習」を参照してください。

メインブランチ

「ブランチ」を参照してください。

マルウェア

コンピュータのセキュリティやプライバシーを侵害するように設計されたソフトウェア。マルウェアは、コンピュータシステムの中断、機密情報の漏洩、不正アクセスを招く可能性があります。マルウェアの例には、ウイルス、ワーム、ランサムウェア、トロイの木馬、スパイウェア、キーロガーなどがあります。

マネージドサービス

AWS のサービス はインフラストラクチャレイヤー、オペレーティングシステム、プラットフォーム AWS を運用し、エンドポイントにアクセスしてデータを保存および取得します。マネージドサービスの例として、Amazon Simple Storage Service (Amazon S3) と Amazon DynamoDB が挙げられます。このサービスは、抽象化されたサービスとも呼ばれます。

製造実行システム (MES)

生産プロセスを追跡、モニタリング、文書化、制御するソフトウェアシステムであり、工場では、これによって、原材料から製品を完成させます。

MAP

「Migration Acceleration Program」を参照してください。

メカニズム

ツールを作成してその導入を推進し、導入結果を調べて調整を行うための包括的なプロセス。メカニズムとは、運用中にそれ自体を強化し改善するサイクルを意味します。詳細については、AWS 「 Well-Architected フレームワーク」の「メカニズムの構築」を参照してください。

メンバーアカウント

組織の一部である管理アカウント AWS アカウント 以外のすべて AWS Organizations。 アカウントが 組織のメンバーになることができるのは、一度に 1 つのみです。

MES

「製造実行システム」を参照してください。

Message Queuing Telemetry Transport (MQTT)

発行/サブスクライブのパターンに基づく、軽量のマシンツーマシン (M2M) 通信プロトコルであり、リソースに限りのある IoT デバイスに使用されます。

マイクロサービス

明確に定義された API を介して通信し、通常は小規模な自己完結型のチームが所有する、小規模で独立したサービスです。例えば、保険システムには、販売やマーケティングなどのビジネス

機能、または購買、請求、分析などのサブドメインにマッピングするマイクロサービスが含まれる場合があります。マイクロサービスの利点には、俊敏性、柔軟なスケーリング、容易なデプロイ、再利用可能なコード、回復力などがあります。詳細については、AWS「サーバーレスサービスを使用したマイクロサービスの統合」を参照してください。

マイクロサービスアーキテクチャ

各アプリケーションプロセスをマイクロサービスとして実行する独立したコンポーネントを使用してアプリケーションを構築するアプローチ。これらのマイクロサービスは、軽量 API を使用して、明確に定義されたインターフェイスを介して通信します。このアーキテクチャの各マイクロサービスは、アプリケーションの特定の機能に対する需要を満たすように更新、デプロイ、およびスケーリングできます。詳細については、「 でのマイクロサービスの実装 AWS」を参照してください。

Migration Acceleration Program (MAP)

組織がクラウドに移行するための強力な運用基盤を構築し、移行の初期コストを相殺するのに役立つコンサルティングサポート、トレーニング、サービスを提供する AWS プログラム。MAP には、組織的な方法でレガシー移行を実行するための移行方法論と、一般的な移行シナリオを自動化および高速化する一連のツールが含まれています。

大規模な移行

アプリケーションポートフォリオの大部分を次々にクラウドに移行し、各ウェーブでより多くのアプリケーションを高速に移動させるプロセス。この段階では、以前の段階から学んだベストプラクティスと教訓を使用して、移行ファクトリー チーム、ツール、プロセスのうち、オートメーションとアジャイルデリバリーによってワークロードの移行を合理化します。これは、AWS 移行戦略 の第 3 段階です。

移行ファクトリー

自動化された俊敏性のあるアプローチにより、ワークロードの移行を合理化する部門横断的なチーム。移行ファクトリーチームには、通常、運用、ビジネスアナリストおよび所有者、移行エンジニア、デベロッパー、およびスプリントで作業する DevOps プロフェッショナルが含まれます。エンタープライズアプリケーションポートフォリオの 20～50% は、ファクトリーのアプローチによって最適化できる反復パターンで構成されています。詳細については、このコンテンツセットの移行ファクトリーに関する解説とCloud Migration Factory ガイドを参照してください。

移行メタデータ

移行を完了するために必要なアプリケーションおよびサーバーに関する情報。移行パターンごとに、異なる一連の移行メタデータが必要です。移行メタデータの例としては、ターゲットサブネット、セキュリティグループ、 AWS アカウントなどがあります。

移行パターン

移行戦略、移行先、および使用する移行アプリケーションまたはサービスを詳述する、反復可能な移行タスク。例: AWS Application Migration Service を使用して Amazon EC2 への移行をリホストします。

Migration Portfolio Assessment (MPA)

オンラインツール。これによって、 AWS クラウドに移行するビジネスケースの検証に必要な情報を得られます。MPA は、詳細なポートフォリオ評価 (サーバーの適切なサイジング、価格設定、TCO 比較、移行コスト分析) および移行プラン (アプリケーションデータの分析とデータ収集、アプリケーションのグループ化、移行の優先順位付け、およびウェーブプランニング) を提供します。MPA ツール (ログインが必要) は、すべての AWS コンサルタントと APN パートナーコンサルタントが無料で利用できます。

移行準備状況評価 (MRA)

AWS CAF を使用して、組織のクラウド準備状況に関するインサイトを取得し、長所と短所を特定し、特定されたギャップを埋めるためのアクションプランを構築するプロセス。詳細については、移行準備状況ガイドを参照してください。MRA は、AWS 移行戦略の第一段階です。

移行戦略

ワークロードを AWS クラウドに移行するために使用するアプローチ。詳細については、この用語集の 7 Rs エントリと、「組織を動員して大規模な移行を加速する」を参照してください。

ML

「機械学習」を参照してください。

モダナイゼーション

古い (レガシーまたはモノリシック) アプリケーションとそのインフラストラクチャをクラウド内の俊敏で弾力性のある高可用性システムに変換して、コストを削減し、効率を高め、イノベーションを活用します。詳細については、「AWS クラウドでのアプリケーションのモダナイズ戦略」を参照してください。

モダナイゼーション準備状況評価

組織のアプリケーションのモダナイゼーションの準備状況を判断し、利点、リスク、依存関係を特定し、組織がこれらのアプリケーションの将来の状態をどの程度適切にサポートできるかを決定するのに役立つ評価。評価の結果として、ターゲットアーキテクチャのブループリント、モダナイゼーションプロセスの開発段階とマイルストーンを詳述したロードマップ、特定されたギャップに対処するためのアクションプランが得られます。詳細については、「AWS クラウドでのアプリケーションのモダナイゼーションの準備状況を評価する」を参照してください。

モノリシックアプリケーション (モノリス)

緊密に結合されたプロセスを持つ単一のサービスとして実行されるアプリケーション。モノリシックアプリケーションにはいくつかの欠点があります。1 つのアプリケーション機能エクスペリエンスの需要が急増する場合は、アーキテクチャ全体をスケーリングする必要があります。モノリシックアプリケーションの特徴を追加または改善することは、コードベースが大きくなると複雑になります。これらの問題に対処するには、マイクロサービスアーキテクチャを使用できます。詳細については、「モノリスをマイクロサービスに分解する」を参照してください。

MPA

「Migration Portfolio Assessment」を参照してください。

MQTT

「Message Queuing Telemetry Transport」を参照してください。

多クラス分類

複数のクラスの予測を生成するプロセス (2 つ以上の結果の 1 つを予測します)。例えば、機械学習モデルが、「この製品は書籍、自動車、電話のいずれですか?」 または、「このお客様にとって最も関心のある商品のカテゴリはどれですか?」と聞くかもしれません。

ミュータブルなインフラストラクチャ

本番ワークロードに使用する既存のインフラストラクチャを更新および変更するためのモデル。Well-Architected AWS フレームワークでは、一貫性、信頼性、予測可能性を向上させるために、イミュータブルインフラストラクチャの使用をベストプラクティスとして推奨しています。

O

OAC

「オリジンアクセス制御」を参照してください。

OAI

「オリジンアクセスアイデンティティ」を参照してください。

OCM

「組織変更管理」を参照してください。

オフライン移行

移行プロセス中にソースワークロードを停止させる移行方法。この方法はダウンタイムが長くなるため、通常は重要ではない小規模なワークロードに使用されます。

OI

「オペレーション統合」を参照してください。

Ola

「オペレーショナルレベルアグリーメント」を参照してください。

オンライン移行

ソースワークロードをオフラインにせずにターゲットシステムにコピーする移行方法。ワークロードに接続されているアプリケーションは、移行中も動作し続けることができます。この方法はダウンタイムがゼロから最小限で済むため、通常は重要な本番稼働環境のワークロードに使用されます。

OPC-UA

「Open Process Communications - Unified Architecture」を参照してください。

Open Process Communications - Unified Architecture (OPC-UA)

産業オートメーション用のマシンツーマシン (M2M) 通信プロトコル。OPC-UA により、相互運用の際に、データ暗号化、認証、認可の各スキームを標準化できます。

オペレーショナルレベルアグリーメント (OLA)

サービスレベルアグリーメント (SLA) をサポートするために、どの機能的 IT グループが互いに提供することを約束するかを明確にする契約。

運用準備状況レビュー (ORR)

質問と関連するベストプラクティスのチェックリスト。インシデントや起こり得る障害を理解、評価、防止したり、その範囲を縮小したりする際に役立ちます。詳細については、 AWS Well-Architected フレームワークの「Operational Readiness Reviews (ORR)」を参照してください。

運用テクノロジー (OT)

産業オペレーション、機器、インフラストラクチャを制御するために物理環境と連携させるハードウェアおよびソフトウェアシステム。製造分野では、Industry 4.0 への変革を進める上で、OT と情報技術 (IT) システムの統合に焦点が当てられています。

オペレーション統合 (OI)

クラウドでオペレーションをモダナイズするプロセスには、準備計画、オートメーション、統合が含まれます。詳細については、オペレーション統合ガイドを参照してください。

組織の証跡

組織 AWS アカウント 内のすべての のすべてのイベント AWS CloudTrail を ログに記録する によって作成された証跡 AWS Organizations。証跡は、組織に含まれている各 AWS アカウント に作成され、各アカウントのアクティビティを追跡します。詳細については、CloudTrail ドキュメントの「組織の証跡の作成」を参照してください。

組織変更管理 (OCM)

人材、文化、リーダーシップの観点から、主要な破壊的なビジネス変革を管理するためのフレームワーク。OCM は、変化の導入を加速し、移行問題に対処し、文化や組織の変化を推進することで、組織が新しいシステムと戦略の準備と移行するのを支援します。 AWS 移行戦略では、クラウド導入プロジェクトに必要な変化のスピードにより、このフレームワークは人材アクセラレーションと呼ばれます。詳細については、OCM ガイドを参照してください。

オリジンアクセス制御 (OAC)

Amazon Simple Storage Service (Amazon S3) コンテンツを保護するための、CloudFront のアクセス制限の強化オプション。OAC は AWS リージョン、すべての S3 バケット、 AWS KMS (SSE-KMS) によるサーバー側の暗号化、S3 バケットへの動的 PUT および DELETEリクエストをサポートします。

オリジンアクセスアイデンティティ (OAI)

CloudFront の、Amazon S3 コンテンツを保護するためのアクセス制限オプション。OAI を使用すると、CloudFront が、Amazon S3 に認証可能なプリンシパルを作成します。認証されたプリンシパルは、S3 バケット内のコンテンツに、特定の CloudFront ディストリビューションを介してのみアクセスできます。OAC も併せて参照してください。OAC では、より詳細な、強化されたアクセス制御が可能です。

ORR

「運用準備状況レビュー」を参照してください。

OT

「運用テクノロジー」を参照してください。

アウトバウンド (送信) VPC

AWS マルチアカウントアーキテクチャでは、アプリケーション内から開始されたネットワーク接続を処理する VPC。AWS Security Reference Architecture では、アプリケーションとより広範なインターネット間の双方向のインターフェイスを保護するために、インバウンド、アウトバウンド、インスペクションの各 VPC を使用してネットワークアカウントを設定することを推奨しています。

P

アクセス許可の境界

ユーザーまたはロールが使用できるアクセス許可の上限を設定する、IAM プリンシパルにアタッチされる IAM 管理ポリシー。詳細については、IAM ドキュメントのアクセス許可の境界を参照してください。

個人を特定できる情報 (PII)

直接閲覧した場合、または他の関連データと組み合わせた場合に、個人の身元を合理的に推測するために使用できる情報。PII の例には、氏名、住所、連絡先情報などがあります。

PII

「個人を特定できる情報」を参照してください。

プレイブック

クラウドでのコアオペレーション機能の提供など、移行に関連する作業を取り込む、事前定義された一連のステップ。プレイブックは、スクリプト、自動ランブック、またはお客様のモダナイズされた環境を運用するために必要なプロセスや手順の要約などの形式をとることができます。

PLC

「プログラマブルロジックコントローラー」を参照してください。

PLM

「製品ライフサイクル管理」を参照してください。

ポリシー

次の操作を可能にするオブジェクト: アクセス許可を定義する (ID ベースのポリシーを参照)。アクセス条件を指定する (リソースベースのポリシーを参照)。 AWS Organizations の組織における全アカウントにアクセス許可の上限を定義する (サービスコントロールポリシーを参照)。

多言語の永続性

データアクセスパターンやその他の要件に基づいて、マイクロサービスのデータストレージテクノロジーを個別に選択します。マイクロサービスが同じデータストレージテクノロジーを使用している場合、実装上の問題が発生したり、パフォーマンスが低下する可能性があります。マイクロサービスは、要件に最も適合したデータストアを使用すると、より簡単に実装でき、パフォーマンスとスケーラビリティが向上します。

ポートフォリオ評価

移行を計画するために、アプリケーションポートフォリオの検出、分析、優先順位付けを行うプロセス。詳細については、「移行の準備状況の評価」を参照してください。

述語

`true` または `false` を返すためのクエリ条件。一般的に、`WHERE` 句に記述されます。

述語プッシュダウン

データベースクエリを最適化する手法。これによって、転送前にクエリ内のデータをフィルタリングします。この手法を取ると、リレーショナルデータベースから取得し処理する必要のあるデータの量が減少するため、クエリのパフォーマンスが向上します。

予防的コントロール

イベントの発生を防ぐように設計されたセキュリティコントロール。このコントロールは、ネットワークへの不正アクセスや好ましくない変更を防ぐ最前線の防御です。詳細については、「AWSでのセキュリティコントロールの実装」の「予防的コントロール」を参照してください。

プリンシパル

アクションを実行し AWS 、リソースにアクセスできる のエンティティ。このエンティティは通常、、IAM AWS アカウントロール、または ユーザーのルートユーザーです。詳細については、IAM ドキュメントの「ロールに関する用語と概念」にあるプリンシパルを参照してください。

プライバシーバイデザイン

開発プロセス全体を通してプライバシーが考慮されているシステムエンジニアリングのアプローチ。

プライベートホストゾーン

1 つ以上の VPC 内のドメインとそのサブドメインへの DNS クエリに対し、Amazon Route 53 がどのように応答するかに関する情報を保持するコンテナ。詳細については、Route 53 ドキュメントの「プライベートホストゾーンの使用」を参照してください。

プロアクティブコントロール

非準拠リソースのデプロイ防止を目的としたセキュリティコントロール。このコントロールにより、プロビジョニング前にリソースをスキャンします。コントロールに準拠していないリソースは、プロビジョニングされません。詳細については、 AWS Control Tower ドキュメントの「コントロールリファレンスガイド」および「 セキュリティコントロールの実装」の「プロアクティブコントロール」を参照してください。 AWS

製品ライフサイクル管理 (PLM)

製品の設計、開発、発売から、成長、成熟、衰退、廃棄に至る、製品のライフサイクル全体を通してデータとプロセスを管理すること。

本番環境

「環境」を参照してください。

プログラマブルロジックコントローラー (PLC)

製造分野で使用される、信頼性と適応性に優れたコンピュータであり、これによって、マシンをモニタリングするとともに、製造プロセスを自動化します。

プロンプトチェイニング

1 つの LLM プロンプトによる出力を次のプロンプトの入力に使用して、より良いレスポンスを生成します。この手法を使用すると、複雑なタスクをサブタスクに分割したり、事前レスポンスを繰り返し改良または拡張したりできます。これによって、モデルのレスポンスの精度と関連性が向上し、粒度の高いパーソナライズされた結果を得られます。

仮名化

データセット内の個人識別子をプレースホルダー値に置き換えるプロセス。仮名化は個人のプライバシー保護に役立ちます。仮名化されたデータは、依然として個人データとみなされます。

発行/サブスクライブ (pub/sub)

マイクロサービス間の非同期通信を可能にするパターン。これにより、スケーラビリティと応答性を向上させます。例えば、マイクロサービスベースの MES の場合、マイクロサービスは、他のマイクロサービスがサブスクライブ可能なチャネルにイベントメッセージを発行できます。このシステムでは、発行サービスの変更なしに、新規マイクロサービスを追加できます。

Q

クエリプラン

手順などの一連のステップであり、SQL リレーショナルデータベースシステムのデータにアクセスするために使用されます。

クエリプランのリグレッション

データベースサービスのオプティマイザーが、データベース環境に特定の変更が加えられる前に選択されたプランよりも最適性の低いプランを選択すること。これは、統計、制限事項、環境設定、クエリパラメータのバインディングの変更、およびデータベースエンジンの更新などが原因である可能性があります。

R

RACI マトリックス

「実行責任者、説明責任者、協業先、報告先 (RACI)」を参照してください。

RAG

「検索拡張生成」を参照してください。

ランサムウェア

決済が完了するまでコンピュータシステムまたはデータへのアクセスをブロックするように設計された、悪意のあるソフトウェア。

RASCI マトリックス

「実行責任者、説明責任者、協業先、報告先 (RACI)」を参照してください。

RCAC

「行と列のアクセス制御」を参照してください。

リードレプリカ

読み取り専用に使用されるデータベースのコピー。クエリをリードレプリカにルーティングして、プライマリデータベースへの負荷を軽減できます。

リアーキテクト

「7 Rs」を参照してください。

目標復旧時点 (RPO)

最後のデータリカバリポイントからの最大許容時間です。これにより、最後の回復時点からサービスが中断されるまでの間に許容できるデータ損失の程度が決まります。

目標復旧時間 (RTO)

サービスの中断から復旧までの最大許容遅延時間。

リファクタリング

「7 Rs」を参照してください。

リージョン

地理的エリア内の AWS リソースのコレクション。各 AWS リージョン は、耐障害性、安定性、耐障害性を提供するために、他の から分離され、独立しています。詳細については、「アカウントが使用できる AWS リージョン を指定する」を参照してください。

リグレッション

数値を予測する機械学習手法。例えば、「この家はどれくらいの値段で売れるでしょうか?」という問題を解決するために、機械学習モデルは、線形回帰モデルを使用して、この家に関する既知の事実 (平方フィートなど) に基づいて家の販売価格を予測できます。

リホスト

「7 Rs」を参照してください。

リリース

デプロイプロセスで、変更を本番環境に昇格させること。

再配置

「7 Rs」を参照してください。

リプラットフォーム

「7 Rs」を参照してください。

再購入

「7 Rs」を参照してください。

回復性

中断に抵抗または中断から回復するアプリケーションの機能。 AWS クラウドでの回復力を計画する際には、一般的に、高可用性とディザスタリカバリが考慮されます。詳細については、「AWS クラウド の耐障害性」を参照してください。

リソースベースのポリシー

Amazon S3 バケット、エンドポイント、暗号化キーなどのリソースにアタッチされたポリシー。このタイプのポリシーは、アクセスが許可されているプリンシパル、サポートされているアクション、その他の満たすべき条件を指定します。

実行責任者、説明責任者、協業先、報告先 (RACI) に基づくマトリックス

移行活動とクラウド運用に関わるすべての関係者の役割と責任を定義したマトリックス。マトリックスの名前は、マトリックスで定義されている責任の種類、すなわち責任 (R) 、説明責任 (A) 、協議 (C) 、情報提供 (I) に由来します。サポート (S) タイプはオプションです。サポートが含まれる場合は RASCI マトリックスと呼ばれ、含まれない場合は RACI マトリックスと呼ばれます。

レスポンシブコントロール

有害事象やセキュリティベースラインからの逸脱について、修復を促すように設計されたセキュリティコントロール。詳細については、「 AWSでのセキュリティコントロールの実装」の「レスポンシブコントロール」を参照してください。

保持

「7 Rs」を参照してください。

廃止

「7 Rs」を参照してください。

検索拡張生成 (RAG)

生成 AI の技術。これにより、LLM では、レスポンスの生成前に、トレーニングデータソースの外部にある信頼できるデータソースが参照されます。例えば、RAG モデルによって、組織のナレッジベースまたはカスタムデータのセマンティック検索を実行できる場合があります。細については、「RAG (検索拡張生成) とは何ですか?」を参照してください。

ローテーション

定期的にシークレット情報を更新して、攻撃者が認証情報にアクセスするのをより困難にするプロセス。

行と列のアクセス制御 (RCAC)

アクセスルールが定義された、基本的で柔軟な SQL 表現の使用。RCAC は行権限と列マスクで構成されています。

RPO

「目標復旧時点」を参照してください。

RTO

「目標復旧時間」を参照してください。

ランブック

特定のタスクを実行するために必要な手動または自動化された一連の手順。これらは通常、エラー率の高い反復操作や手順を合理化するために構築されています。

S

SAML 2.0

多くの ID プロバイダー (IdP) が使用しているオープンスタンダード。この機能を使用すると、フェデレーティッドシングルサインオン (SSO) が有効になるため、ユーザーは組織内のすべてのユーザーを IAM で作成しなくても、 AWS マネジメントコンソール にログインしたり AWS 、API オペレーションを呼び出すことができます。SAML 2.0 ベースのフェデレーションの詳細については、IAM ドキュメントの「SAML 2.0 ベースのフェデレーションについて」を参照してください。

SCADA

「監視制御とデータ取得」を参照してください。

SCP

「サービスコントロールポリシー」を参照してください。

シークレット

暗号化された形式で保存する AWS Secrets Managerパスワードやユーザー認証情報などの機密情報または制限付き情報。シークレット値とそのメタデータで構成されます。シークレット値には、バイナリ、1 つの文字列、複数の文字列を指定できます。詳細については、Secrets Manager ドキュメントの「Secrets Manager シークレットの概要」を参照してください。

セキュリティバイデザイン

開発プロセス全体を通してセキュリティが考慮されているシステムエンジニアリングのアプローチ。

セキュリティコントロール

脅威アクターによるセキュリティ脆弱性の悪用を防止、検出、軽減するための、技術上または管理上のガードレール。セキュリティコントロールには、主に 4 つの種類があります。4 つとは、予防、検出、レスポンシブ、プロアクティブです。

セキュリティ強化

アタックサーフェスを狭めて攻撃への耐性を高めるプロセス。このプロセスには、不要になったリソースの削除、最小特権を付与するセキュリティのベストプラクティスの実装、設定ファイル内の不要な機能の無効化、といったアクションが含まれています。

Security Information and Event Management (SIEM) システム

セキュリティ情報管理 (SIM) とセキュリティイベント管理 (SEM) のシステムを組み合わせたツールとサービス。SIEM システムは、サーバー、ネットワーク、デバイス、その他ソースからデータを収集、モニタリング、分析して、脅威やセキュリティ違反を検出し、アラートを発信します。

セキュリティレスポンスの自動化

セキュリティイベントへの自動レスポンスまたは自動修復を目的として、事前定義およびプログラムされたアクション。これらの自動化は、セキュリティのベストプラクティスを実装するのに役立つ検出的または応答的な AWS セキュリティコントロールとして機能します。自動レスポンスアクションの例には、VPC セキュリティグループの変更、Amazon EC2 インスタンスへのパッチ適用、認証情報の更新などがあります。

サーバー側の暗号化

送信先で、それ AWS のサービス を受け取る によるデータの暗号化。

サービスコントロールポリシー (SCP)

AWS Organizationsの組織内の、すべてのアカウントのアクセス許可を一元的に管理するポリシー。SCP は、管理者がユーザーまたはロールに委任するアクションに、ガードレールを定義したり、アクションの制限を設定したりします。SCP は、許可リストまたは拒否リストとして、許可または禁止するサービスやアクションを指定する際に使用できます。詳細については、 AWS Organizations ドキュメントの「サービスコントロールポリシー」を参照してください。

サービスエンドポイント

のエントリポイントの URL AWS のサービス。ターゲットサービスにプログラムで接続するには、エンドポイントを使用します。詳細については、「AWS 全般のリファレンス」の「AWS のサービス エンドポイント」を参照してください。

サービスレベルアグリーメント (SLA)

サービスのアップタイムやパフォーマンスなど、IT チームがお客様に提供すると約束したものを明示した合意書。

サービスレベルインジケータ (SLI)

エラー率、可用性、スループットといった、サービスパフォーマンス面の指標。

サービスレベル目標 (SLO)

サービスレベルインジケータによって測定され、サービスの状態を表すターゲットメトリクス。

責任共有モデル

クラウドのセキュリティとコンプライアンス AWS について と共有する責任を説明するモデル。AWS はクラウドのセキュリティを担当しますが、 はクラウドのセキュリティを担当します。詳細については、「責任共有モデル」を参照してください。

SIEM

「Security Information and Event Management システム」を参照してください。

単一障害点 (SPOF)

特定のアプリケーションを構成する単一の重要なコンポーネントで発生し、システム稼働に支障をきたす可能性のある障害。

SLA

「サービスレベルアグリーメント」を参照してください。

SLI

「サービスレベルインジケータ」を参照してください。

SLO

「サービスレベルの目標」を参照してください。

スプリットアンドシードモデル

モダナイゼーションプロジェクトのスケーリングと加速のためのパターン。新機能と製品リリースが定義されると、コアチームは解放されて新しい製品チームを作成します。これにより、お客様の組織の能力とサービスの拡張、デベロッパーの生産性の向上、迅速なイノベーションのサポートに役立ちます。詳細については、「AWS クラウドでのアプリケーションをモダナイズするための段階的アプローチ」を参照してください。

SPOF

「単一障害点」を参照してください。

スタースキーマ

データベースの編成構造を意味し、1 つの大きいファクトテーブルにトランザクションデータまたは測定データが保存され、1 つ以上の小さいディメンションテーブルにデータ属性が保存されます。この構造は、データウェアハウスやビジネスインテリジェンスを用途とするように設計されています。

strangler fig パターン

レガシーシステムが廃止されるまで、システム機能を段階的に書き換えて置き換えることにより、モノリシックシステムをモダナイズするアプローチ。このパターンは、宿主の樹木から根を成長させ、最終的にその宿主を包み込み、宿主に取って代わるイチジクのつるを例えています。そのパターンは、モノリシックシステムを書き換えるときのリスクを管理する方法として Martin Fowler により提唱されました。このパターンの適用方法の例については、「コンテナと Amazon API Gateway を使用して、従来の Microsoft ASP.NET (ASMX) ウェブサービスを段階的にモダナイズ」を参照してください。

サブネット

VPC 内の IP アドレスの範囲。サブネットは、1 つのアベイラビリティーゾーンに存在する必要があります。

監視制御とデータ取得 (SCADA)

製造分野において、ハードウェアとソフトウェアを使用して物理アセットと本番運用をモニタリングするシステム。

対称暗号化

データの暗号化と復号に同じキーを使用する暗号化のアルゴリズム。

合成テスト

ユーザーとのやり取りをシミュレートして、起こり得る問題を検出したり、パフォーマンスをモニタリングしたりすることで、システムをテストします。Amazon CloudWatch Synthetics を使用すると、こうしたテストを作成できます。

システムプロンプト

コンテキスト、指示、ガイドラインなどを提示して、LLM に動作を指示する手法。システムプロンプトは、コンテキストを設定して、ユーザーとやり取りするルールを確立するのに有用です。

T

タグ

AWS リソースを整理するためのメタデータとして機能するキーと値のペア。タグは、 リソースの管理、識別、整理、検索、フィルタリングに役立ちます。詳細については、「AWS リソースのタグ付け」を参照してください。

ターゲット変数

監督された機械学習でお客様が予測しようとしている値。これは、結果変数のことも指します。例えば、製造設定では、ターゲット変数が製品の欠陥である可能性があります。

タスクリスト

ランブックの進行状況を追跡するために使用されるツール。タスクリストには、ランブックの概要と完了する必要のある一般的なタスクのリストが含まれています。各一般的なタスクには、推定所要時間、所有者、進捗状況が含まれています。

テスト環境

「環境」を参照してください。

トレーニング

お客様の機械学習モデルに学習するデータを提供すること。トレーニングデータには正しい答えが含まれている必要があります。学習アルゴリズムは入力データ属性をターゲット (お客様が予測したい答え) にマッピングするトレーニングデータのパターンを検出します。これらのパターンをキャプチャする機械学習モデルを出力します。そして、お客様が機械学習モデルを使用して、ターゲットがわからない新しいデータでターゲットを予測できます。

トランジットゲートウェイ

VPC とオンプレミスネットワークを相互接続するために使用できる、ネットワークの中継ハブ。詳細については、 AWS Transit Gateway ドキュメントの「トランジットゲートウェイとは」を参照してください。

トランクベースのワークフロー

デベロッパーが機能ブランチで機能をローカルにビルドしてテストし、その変更をメインブランチにマージするアプローチ。メインブランチはその後、開発環境、本番前環境、本番環境に合わせて順次構築されます。

信頼されたアクセス

ユーザーに代わって AWS Organizations およびそのアカウントで組織内でタスクを実行するために指定したサービスにアクセス許可を付与します。信頼されたサービスは、サービスにリンクされたロールを必要なときに各アカウントに作成し、ユーザーに代わって管理タスクを実行します。詳細については、 ドキュメントの「Using AWS Organizations with other AWS services AWS Organizations 」を参照してください。

チューニング

機械学習モデルの精度を向上させるために、お客様のトレーニングプロセスの側面を変更する。例えば、お客様が機械学習モデルをトレーニングするには、ラベル付けセットを生成し、ラベルを追加します。これらのステップを、異なる設定で複数回繰り返して、モデルを最適化します。

ツーピザチーム

2 枚のピザを分け合えることができるくらい小さな DevOps チーム。ツーピザチームの規模では、ソフトウェア開発におけるコラボレーションに最適な機会が確保されます。

U

不確実性

予測機械学習モデルの信頼性を損なう可能性がある、不正確、不完全、または未知の情報を指す概念。不確実性には、次の 2 つのタイプがあります。認識論的不確実性は、限られた、不完全なデータによって引き起こされ、弁論的不確実性 は、データに固有のノイズとランダム性によって引き起こされます。

未分化なタスク

ヘビーリフティングとも呼ばれ、アプリケーションの作成と運用には必要だが、エンドユーザーに直接的な価値をもたらさなかったり、競争上の優位性をもたらしたりしない作業です。未分化なタスクの例としては、調達、メンテナンス、キャパシティプランニングなどがあります。

上位環境

「環境」を参照してください。

V

バキューミング

ストレージを再利用してパフォーマンスを向上させるために、増分更新後にクリーンアップを行うデータベースのメンテナンス操作。

バージョンコントロール

リポジトリ内のソースコードへの変更など、変更を追跡するプロセスとツール。

VPC ピアリング

プライベート IP アドレスを使用してトラフィックをルーティングできる、2 つの VPC 間の接続。詳細については、Amazon VPC ドキュメントの「VPC ピア機能とは」を参照してください。

脆弱性

システムのセキュリティを脅かすソフトウェアまたはハードウェアの欠陥。

W

ウォームキャッシュ

頻繁にアクセスされる最新の関連データを含むバッファキャッシュ。データベースインスタンスはバッファキャッシュから、メインメモリまたはディスクからよりも短い時間で読み取りを行うことができます。

ウォームデータ

アクセス頻度の低いデータ。この種類のデータをクエリする場合、通常は適度に遅いクエリでも問題ありません。

ウィンドウ関数

現在のレコードに何らかの形で関連している行のグループに計算を実行する SQL 関数。ウィンドウ関数は、移動平均を計算したり、現在の行の相対位置に基づいて他の行の値にアクセスするといったタスクの処理に役立ちます。

ワークロード

ビジネス価値をもたらすリソースとコード (顧客向けアプリケーションやバックエンドプロセスなど) の総称。

ワークストリーム

特定のタスクセットを担当する移行プロジェクト内の機能グループ。各ワークストリームは独立していますが、プロジェクト内の他のワークストリームをサポートしています。たとえば、ポートフォリオワークストリームは、アプリケーションの優先順位付け、ウェーブ計画、および移行メタデータの収集を担当します。ポートフォリオワークストリームは、これらの設備を移行ワークストリームで実現し、サーバーとアプリケーションを移行します。

WORM

「Write-Once-Read-Many」を参照してください。

WQF

「AWS ワークロード資格フレームワーク」を参照してください

Write-Once-Read-Many (WORM)

データを 1 回のみ書き込むことで、データの削除や変更を防ぐストレージモデル。承認済みユーザーは、必要な回数だけデータを読み取ることができますが、変更することはできません。このデータストレージインフラストラクチャは、イミュータブルと見なされます。

Z

ゼロデイエクスプロイト

ゼロデイ脆弱性を悪用した攻撃 (一般的にマルウェアによる)。

ゼロデイ脆弱性

実稼働システムにおける未解決の欠陥または脆弱性。脅威アクターは、このような脆弱性を利用してシステムを攻撃する可能性があります。開発者は、よく攻撃の結果で脆弱性に気付きます。

ゼロショットプロンプト

LLM にタスク実行の手順は提示するが、実行のガイドとして役立つ例 (ショット) は提示しない方法。LLM は、事前トレーニング済みの知識を使用してタスクを処理する必要があります。ゼロショットプロンプトの有効性は、タスクの複雑さとプロンプトの品質によって異なります。「数ショットプロンプト」も参照してください。

ゾンビアプリケーション

平均 CPU およびメモリ使用率が 5% 未満のアプリケーション。移行プロジェクトでは、これらのアプリケーションを廃止するのが一般的です。

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。