Session Report

LLM とリアルな世界をつなぐ
「Vertex AI Extensions」のユースケースを紹介

LLM を強化する「Vertex AI Extensions」の活用・構築方法

Google Cloud
カスタマー エンジニア
下門 祐二

Google Cloud

Google Cloud

セッション レポート概要

「Vertex AI Extensions」を利用することで LLM を API に接続し、リアルタイムなデータ取得と現実世界のアクションという特長をもつ生成 AI が可能になり、企業は自社のアプリケーションを強化することができます。Google やパートナーが提供するエクステンション（拡張機能）の活用方法から、開発者が独自の拡張機能を構築、管理、デプロイする方法を紹介します。

プレゼンター紹介

Google Cloud
カスタマー エンジニア
下門 祐二

クラウドを活用したシステムのコンサルティングやテクニカル セールスを経験後、2017 年に Google Cloud に入社しました。ゲーム業界向けセールス担当を経て 2020 年より現職に就いています。現在は、デジタル ネイティブのお客様を中心に、Google Cloud のご提案と技術支援に注力しています。

目次

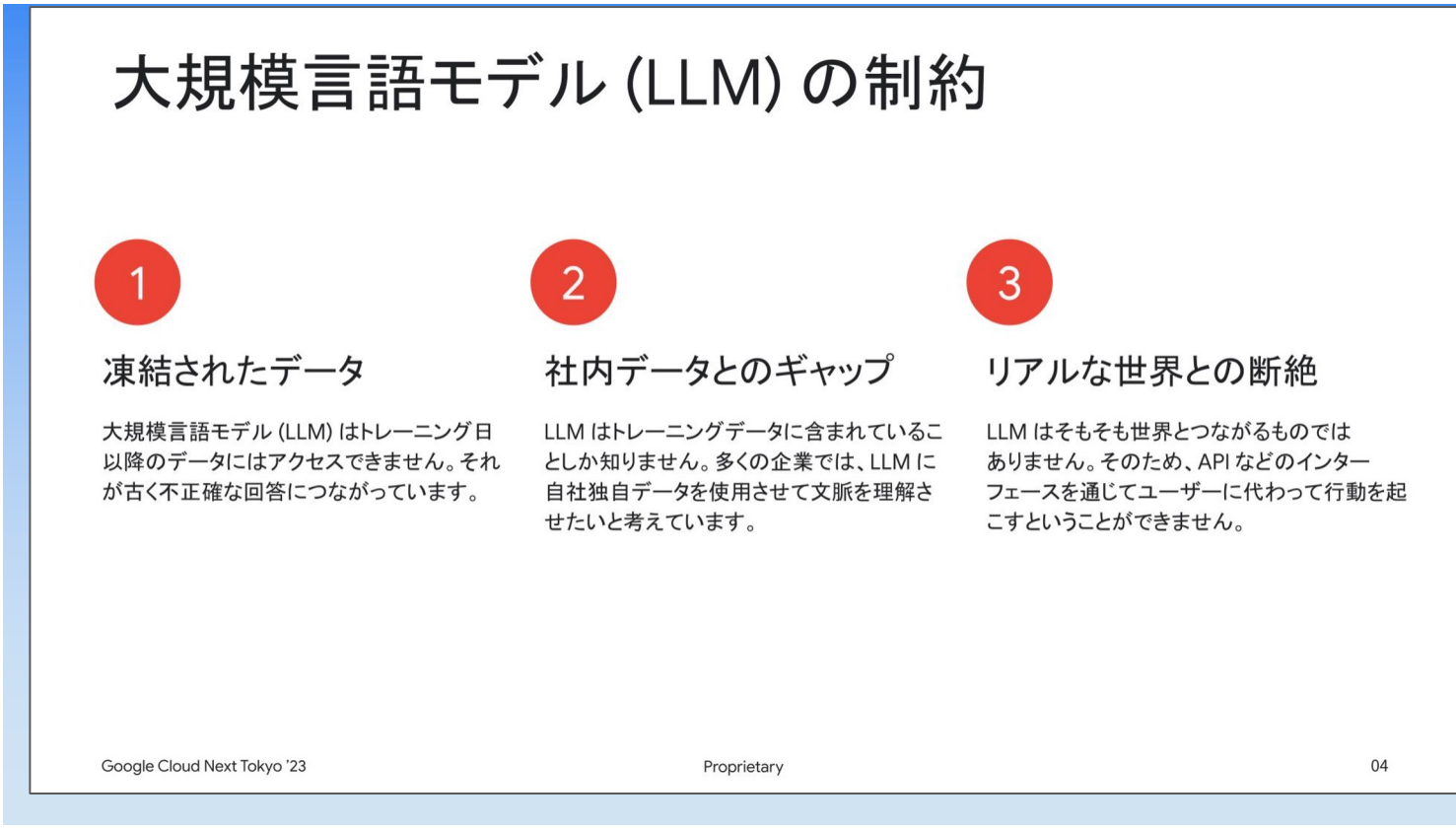
大規模言語モデル（LLM）の制約

LLM は非常に便利なテクノロジーですが、利用するうえでは「凍結されたデータ」「社内データとのギャップ」「リアルな世界との断絶」の 3 つの制約があります。

1 つ目の制約は、LLM が利用するデータは、あくまで過去のある時点で凍結されたデータであることです。LLM は、最後にトレーニングを受けた日以降の情報を得ることができず、それゆえに不正確な回答を返してしまうことがあります。

2 つ目の制約は、トレーニングを終えた LLM は、外部のデータにアクセスできないということです。一般的な基盤モデルは、公開データを用いてトレーニングされます。多くの企業では、収集した社内データを LLM に学習させたいと思いますが、LLM 単体では外部データとの動的な接続が難しく、社内データとのギャップが生じてしまいます。

3 つ目の制約は、LLM がリアルな世界と切り離されていることです。単体では API のような一般的なインターフェースを利用できず、ユーザーの代わりにアクションを起こせません。



LLM のよくある課題

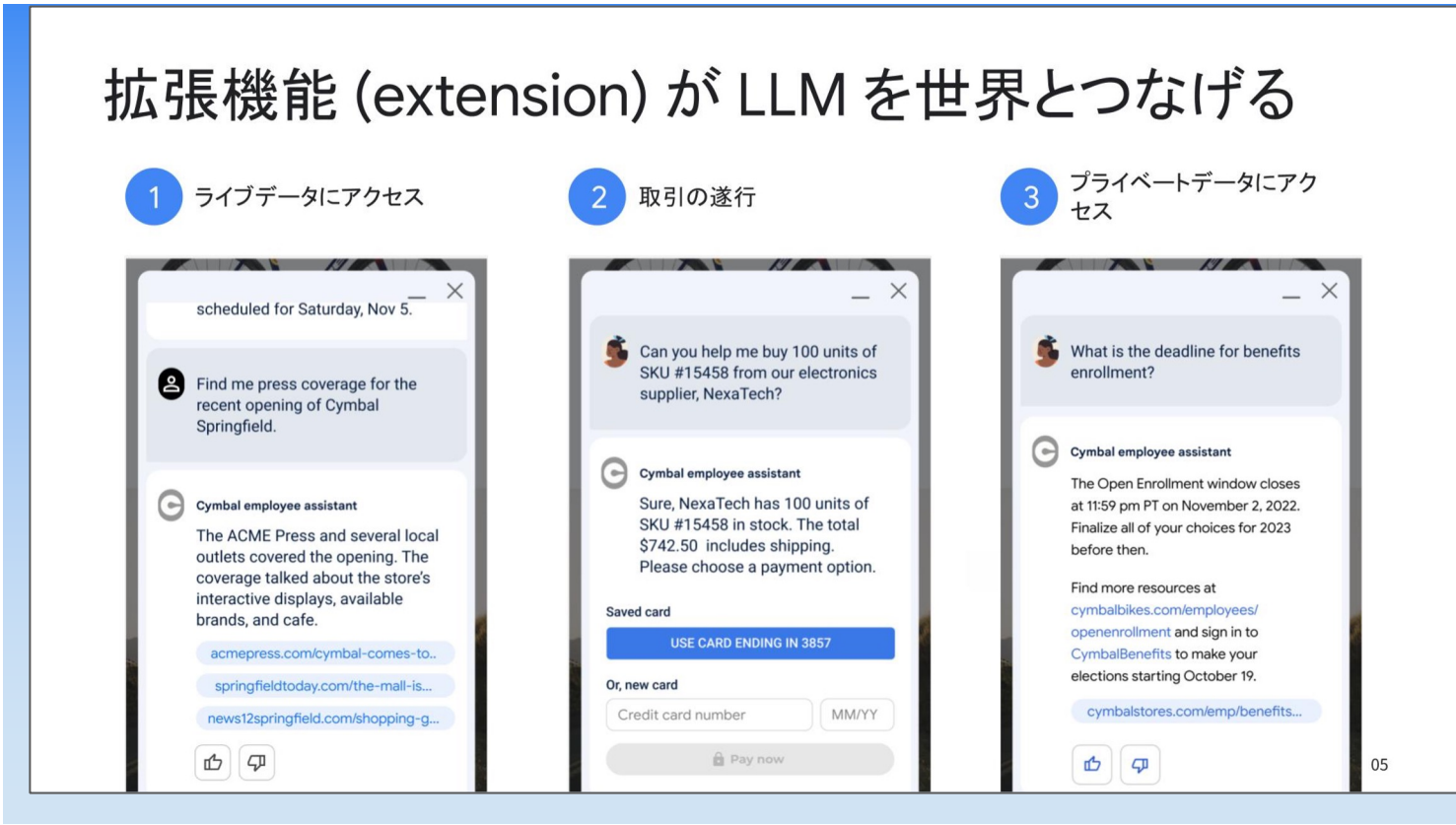
この問題に対処できるのが、一般的にプラグインとしても知られている「Extension（拡張機能）」です。拡張機能を使えば LLM と API をつなぎ、LLM とリアルな世界を効果的に結び付けることが可能となります。

LLM の拡張機能の活用例

活用例をいくつかご紹介しましょう。例えば、ある企業の最新のプレスリリースに問い合わせることで、リアルタイム検索が始まり、最新の情報を回答してくれるというシステムを構築できます。

ほかには、ある従業員が電子部品の購入に関してサポートを求めていると仮定しましょう。従業員の問い合わせによって、サプライチェーンの拡張機能がトリガーされ、リアルタイムの在庫検索が始まります。LLM がその部品の在庫の確認や価格を従業員に提示して、購入完了までスムーズに進められます。

さらに、福利厚生の申請期日を問い合わせた従業員に対して、企業のプライベート拡張機能により LLM と社内のナレッジベースが接続され、LLM は事実に基づいた回答を返す、ということも可能になります。

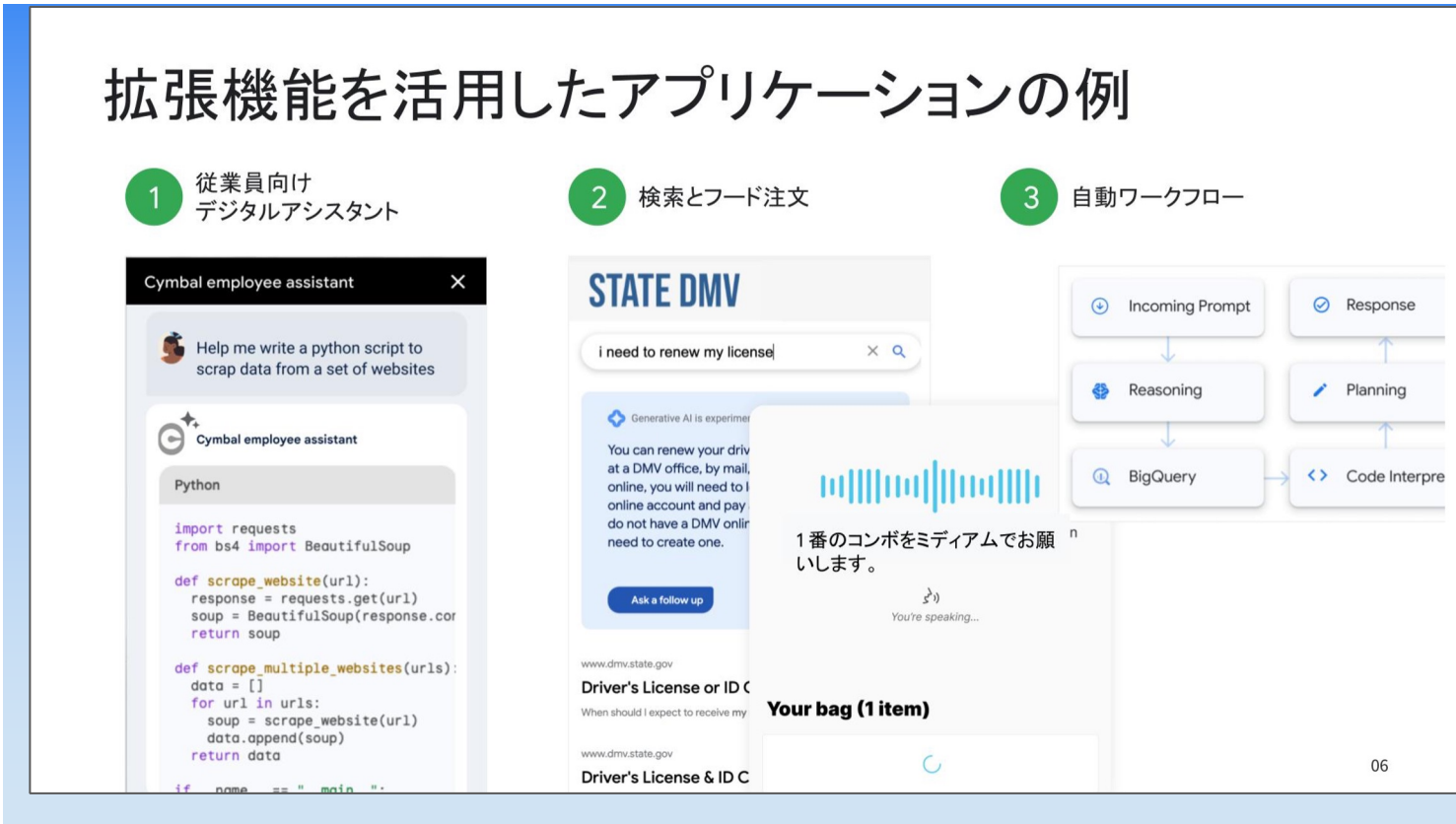


拡張機能によって LLM は最新の情報を返せるようになる

拡張機能を活用した具体的なアプリケーションとして、例えば従業員向けのデジタル アシスタントでは、文書やスプレッドシート、プログラム コードを LLM に作成してもらうことができます。

また、検索エンジンに LLM を活用した場合は、検索者のコンテキストに基づいて、高度にパーソナライズされた回答を返すことができるでしょう。

さらに、ワークフローを自動化するアプリケーションにおいても、自然言語で指示をすることで、データ サイエンスやサイバー セキュリティなど、さまざまな分野のタスクを自動で実行できます。



拡張機能は、さまざまなアプリケーションにも応用可能

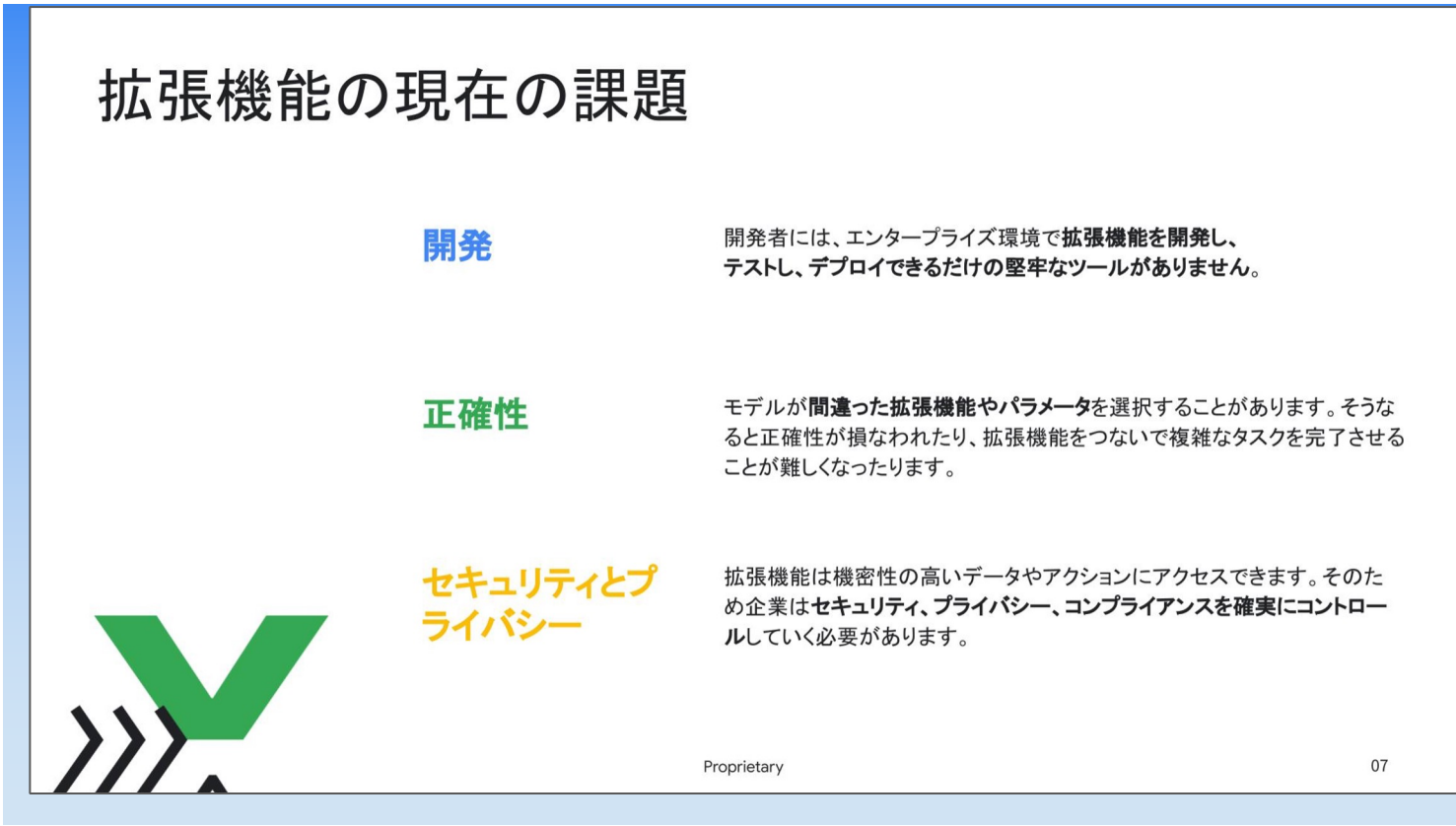
LLM の拡張機能の課題

LLM の制約を解消するための手段として拡張機能は非常に有効ですが、現時点ではいくつかの課題があります。

開発者には、エンタープライズ環境で拡張機能を開発・テスト・デプロイできるだけの効果的なツールがほとんどありません。デバッグや評価のためのツールを自社で開発されている方もいるかもしれませんが、あくまでも自社の特定の分野のユースケースに特化しており、汎用的なものではないことの方が多いでしょう。

また、拡張機能を活用したモデルの回答には、正確性に欠けるものもあります。クエリに対してモデルが誤った拡張機能を選択してしまったり異なるフォーマットを提示してしまったりすることもあります。

さらに、拡張機能は機密性の高いデータやアクションにアクセスできるため、それが新たなリスクにつながる可能性があります。そのため、拡張機能を利用する場合、セキュリティやプライバシー、コンプライアンスを確実に制御しなければなりません。



拡張機能には課題が残る

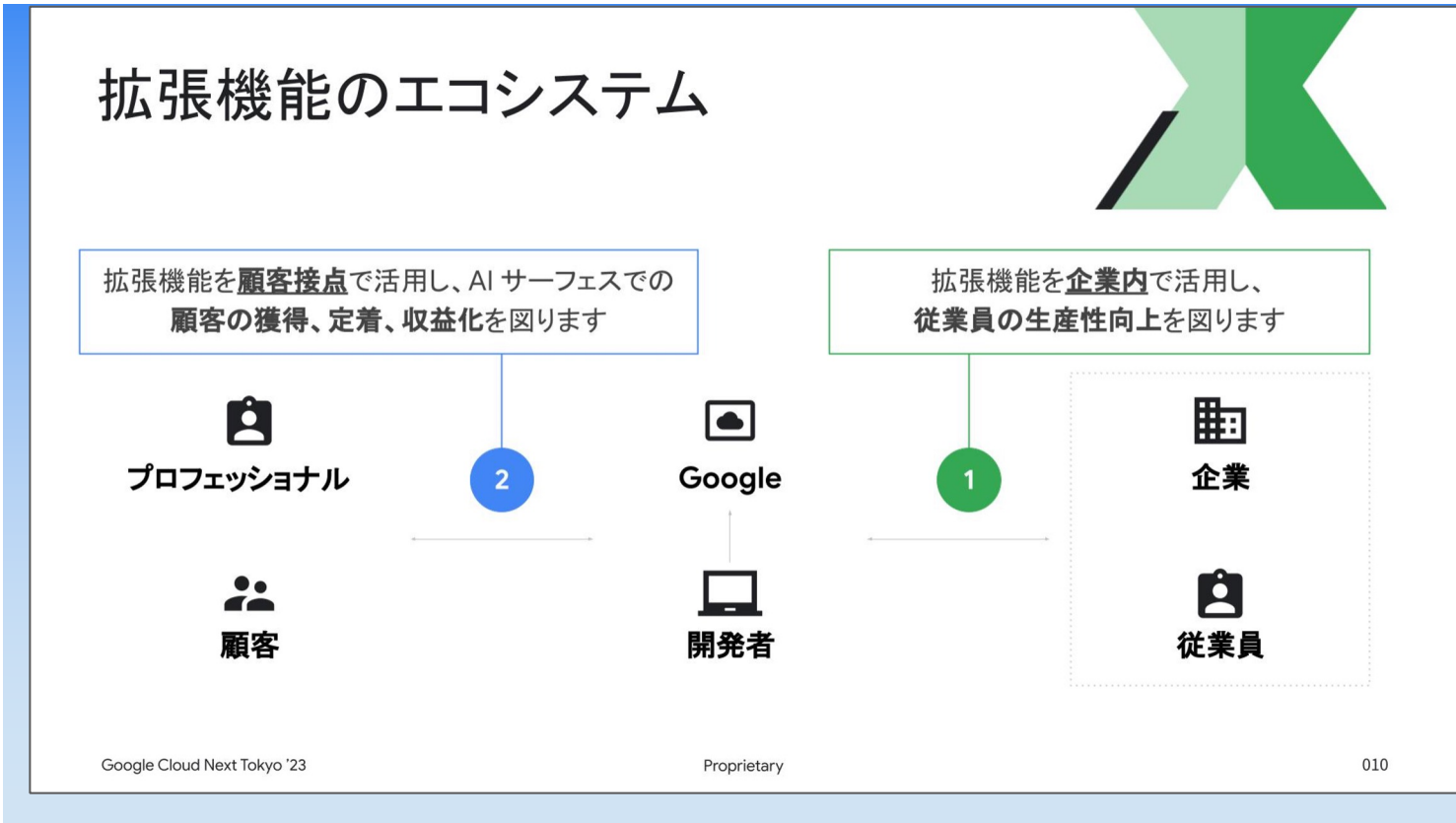
そこで、ここからはエンタープライズ レベルの拡張機能および拡張機能を活用したアプリケーションを構築するためのプラットフォームとして「Vertex AI Extensions」を紹介します。

Vertex AI Extensions とは？

Vertex AI Extensions は、LLM をリアルタイムなデータにセキュアに接続し、リアルな世界でのアクションを実行できる Google Cloud 上の拡張機能です。データアクセス、データの所在地、透明性、監査、社内ポリシーに基づくコンプライアンスをきめ細かに制御して、セキュアな拡張機能を構築できます。

Vertex AI Extensions では、拡張機能を用いたユースケース向けにチューニングされた最新モデルを利用できるという特徴があります。加えて、拡張機能に必要なデータを提供する「BigQuery」や「Vertex AI Search」など、Google Cloud の主要なサービスを統合し、必要な機能を網羅したアプリケーションの構築も可能です。

また、Vertex AI Extensions を中心とした拡張機能のエコシステムを形成できます。このエコシステムには、企業や従業員、開発者、エンドユーザー、一般顧客などが含まれます。



プロフェッショナルや一般ユーザーを含めたエコシステムを形成

拡張機能の作成方法

ここからは、拡張機能の作成方法を見ていきましょう。拡張機能は、主に「APIの選択」「マニフェスト・API 定義」「テスト・デバッグ・公開」「アプリケーションにデプロイ」の 4 つのステップで作成されます。

はじめに、LLM に利用させたい API を選択します。新しく API を作成する場合は、Google Cloud の API 管理サービスである「Apigee」も有効です。

続いて、拡張機能に必要な 2 つのファイルを作成します。

1 つ目は、拡張機能の名前や説明、認証方式など、拡張機能のメタデータを定義するマニフェストです。このファイルは、拡張機能や拡張機能内の API メソッドを、いつどのような場面で利用するのか、API リクエストのパラメーター、ユーザーへのレスポンスの定義などを自然言語で説明し、JSON 形式で保存します。

2 つ目は、LLM が利用できる API のメソッドを定義する API 仕様です。このファイルは、オープン API スペックに準拠した YAML 形式で定義されます。

2 つのファイルを作成したら、Vertex AI Extensions の開発者ツールにて、拡張機能のテスト・デバッグ・公開を行います。

最後に、チャットボットや自動ワークフローなど、自社のアプリケーションに拡張機能をデプロイします。

Vertex AI Extensions では、Google Cloud の API、またはクラウド コンソールから簡単に拡張機能を作成できます。ChatGPT の拡張機能とも互換性があるため、すでに ChatGPT の拡張機能向けに API 定義があれば、そのまま活用することも可能です。

Vertex AI Extensions が実行できる拡張機能

Vertex AI Extensions は、拡張機能の構築だけでなく、拡張機能の参照や活用も簡単に実行できます。主に実行できる拡張機能は 3 種類あります。

1 つ目が、自社の API を利用して独自に作成した拡張機能です。

2 つ目は、Google が作成・提供するファースト パーティの拡張機能です。BigQuery や Vertex AI Search など、広く活用されている Google Cloud のサービス向けに、拡張機能が提供されています。

3 つ目が、パートナーが開発・提供する拡張機能です。例えば、「DataStax」や「MongoDB」などのデータストアに保存されたデータの検索が可能になります。

また、LangChain との互換性があるため、LangChain 上のコードを簡単に Vertex AI Extensions にデプロイすることも可能です。

なお、Vertex AI Extensions では、拡張機能を探し、見つけ、テストができる「ギャラリー」も提供しています。ギャラリーを活用することで、開発にかかる時間を短縮できます。

実行できる拡張機能は多岐にわたる

拡張機能を選択した後は、拡張機能と LLM を関連付けて利用できるようにしましょう。Vertex AI Extensions のランタイムでは「エージェント」と呼ばれる機構が作成され、LLM 側で拡張機能を呼び出すタイミングや呼び出しの必要性を、リーズニングと呼ばれる推論プロセスとして実行します。

Vertex AI Extensionsでは、2 つのタイプのアプリケーション ランタイムを選択できます。

1 つ目は、Google が管理するプラットフォーム上でリーズニングを行うアプリケーション ランタイムです。このランタイムにより、正しい拡張機能が選択されるようになり、アウトプット形式も制御できます。

もう 1 つは、LangChain など、オープンソースのフレームワークを活用した、カスタマイズ可能なアプリケーション ランタイムです。具体的には、LangChain などのエージェントやチェーンなどを利用したカスタムコードを Vertex AI Extensions にデプロイすることで、より柔軟なリーズニングおよびワークフローを構築できるようになります。

2 つのランタイムがあることで、カスタマイズ性やコントロールのレベルを柔軟に選択しつつ、Vertex AI Extensions による一貫した管理が可能です。

Vertex AI Extensions のパフォーマンス評価と最適化

続いて、パフォーマンスの評価方法を解説します。

AI プロジェクトにおいて、単にモデルを動かすことは全体の道のりの一部でしかありません。本番環境レベルでアプリケーションを構築するには、パフォーマンス評価が重要です。そのためには、データセットを作成し、詳細な評価を行えるツールを確保しなければなりません。

Vertex AI Extensions なら、自分で用意したプロンプトのサンプルがあれば、すぐにテスト用のデータセットを作成できます。データセットがあれば、システムの問題が発生する場所を把握することも可能です。さらにはプロンプトのサンプルを集約し、高精度な拡張機能を構築できます。

さらに Vertex AI Extensions で提供されるプロンプト ツールを利用すれば、拡張機能のマニフェストおよび API 仕様のテストと最適化が可能です。問題になりそうな部分を特定できれば、エージェントの設定や最適化などの措置を講じることができます。また、さまざまな基盤モデルを拡張機能でテストして、最適な組み合わせを見つけられます。

これらのツールは、拡張機能を活用した LLM アプリケーションを本番環境に導入するうえで、精度の向上に役立つでしょう。

Vertex AI Extensions の活用例

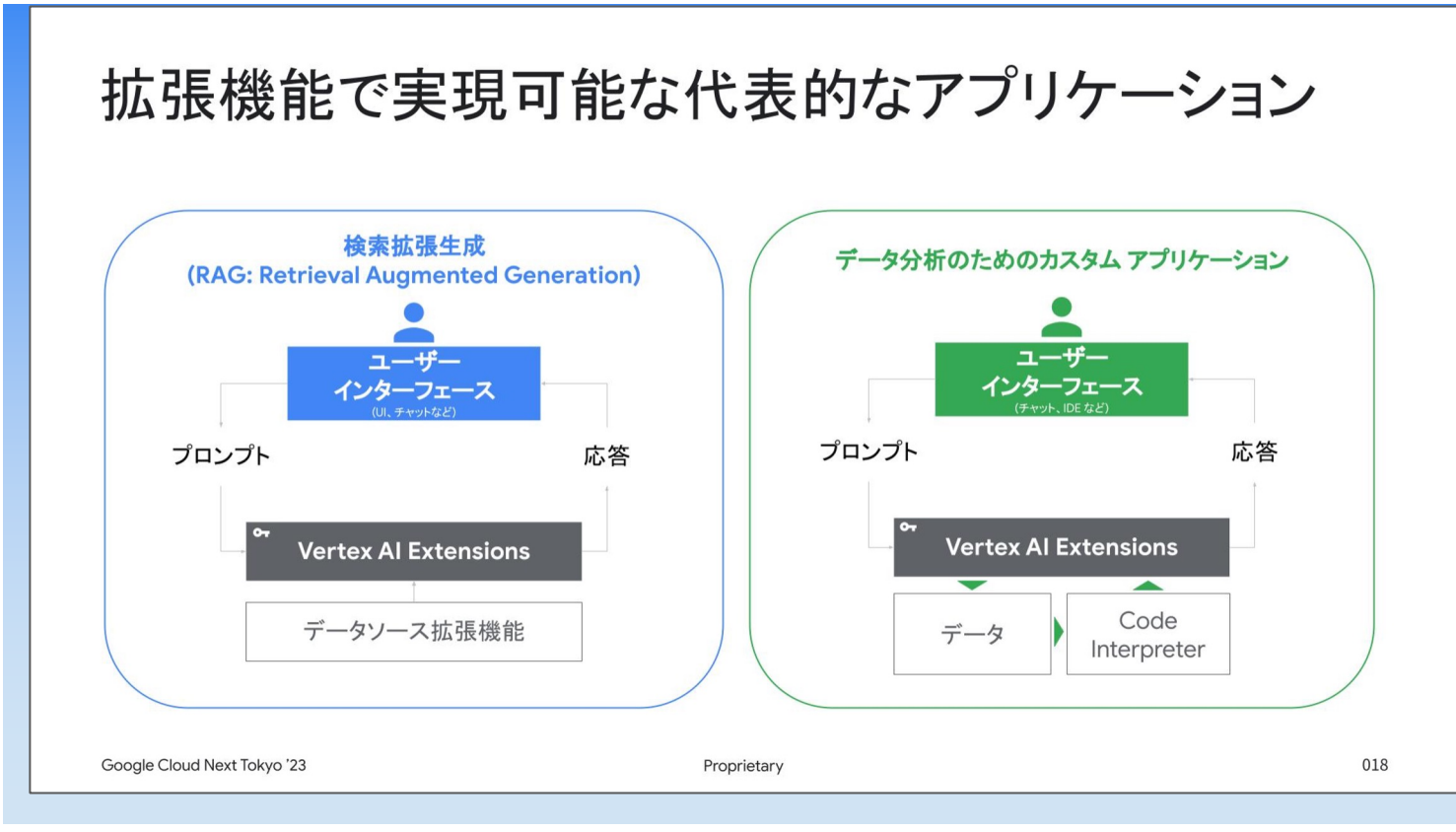
実際に Vertex AI Extensions を活用した 2 つの例を紹介します。

1 つ目に、外部ソースのデータを利用して、LLM で正しい回答を生成するアーキテクチャ「RAG（Retrieval-Augmented Generation：検索拡張生成）」を用いたユースケースを取り上げます。

RAG は自社のデータを利用するときに非常に効果的です。例えば、小売企業が顧客の情報を活用してターゲット型の広告キャンペーンを実施したい場合や、セキュリティ企業が脅威をモニタリング・分類したい場合も、拡張機能を使用することで、より正確で個々の状況に合わせた回答を生成できます。

2 つ目に、Code Interpreter という拡張機能を利用してデータ分析のためのカスタムアプリケーションを構築するユースケースを取り上げます。

Code Interpreter に自然言語のクエリを与えると、タスクを実行するための Python コードを自動的に生成・実行します。そのため、コードを自身で記述することなくデータの分析・可視化が可能になります。



Vertex AI Extensions をビジネスに活用する 2 つのユースケース

イベントでは、これらのユースケースのデモも行いました。ご興味のある方は、ぜひアーカイブ動画をご確認ください。

レースドライバー向けチャットボット開発に RAG を活用

フォーミュラ E の次世代レーシング電気自動車「Genbeta」は、7 月末に国内での自動車速度ギネス世界記録を樹立しました。このときに Google Cloud の生成 AI と RAG のアーキテクチャを採用し、自然言語で会話できるチャットボットをドライバー向けに開発しています。

このチャットボットは、レーシングカーに搭載された多数のセンサーを使って走行に関する膨大なテレメトリー データを取り込み、そこからスピードやパワー、グリップの状態を自然言語でやり取りできるものでした。

このチャットボットにより、ドライバーは自分が必要としているデータにリアルタイムにアクセスできるようになりました。

実際のアーキテクチャを見てみましょう。

まず、インターフェースとしては、ドライバー向け、ファン向け 2 つのチャットボットが提供されています。1 つ目のドライバー向けのボットでは、BigQuery上に蓄積されたテレメトリーデータに基づいて、レーサーが自然言語で必要な情報を質問できます。

もう 1 つが、Vertex AI Search にフォーミュラ E のさまざまな情報を読み込ませて、ファンが会話できるようにしたボットです。このアーキテクチャは「バージョン 1」に位置づけられており、Google Cloud のさまざまなコンポーネントと LangChain を使って構成されています。

この段階では、LangChain を実行するための GKE クラスターのデプロイメントと、検索の精度を確保するための複雑なプロンプトの管理が最大の課題でした。そこで、現在採用されている Vertex AI Extensions を採用した「バージョン 2」のアーキテクチャに移行しました。

LangChain コードと LLM の構成情報を、Vertex AI Extensions に簡単にデプロイするようにしたところ、アーキテクチャを簡略化できたうえに、デプロイの時間も大幅に削減できました。フォーミュラ E では、次のシーズンに向けてさらなるアーキテクチャの改善を計画しています。

業務体験を変える RAG の活用例

その他の事例として、Manhattan Associates や GitLab では、Vertex AI Extensions を活用して、業務において自社独自のワークフローを構築しています。

サプライチェーン ソリューションを提供する Manhattan Associates では、拡張機能を倉庫業務へ活用する取り組みを推進。同社は生成 AI により、サプライチェーンのマネージャーや担当者に、これまでにない有益なインサイトを自動で提供できると考えています。

しかし、そのためには、社内の API を取り込み、大規模なデータを業務に支障をきたすことなく LLM に渡す必要があります。

そこで、同社は Vertex AI Extensionsを活用。人員の稼働率や業務の空き状況、期限に関するリアルタイムの情報に基づいて人員をバランスよく割り当て、サプライチェーン プロセスを自動化していこうと考えています。

一方 GitLab の場合は、Vertex AI Extensions を活用して、開発者自身で脆弱性を解決できるようサポートできないか検討しています。同社では、セキュリティ向上に関するナレッジとユーザー固有のコードベースに LLM がアクセスできるような拡張機能の活用方法を検討中です。

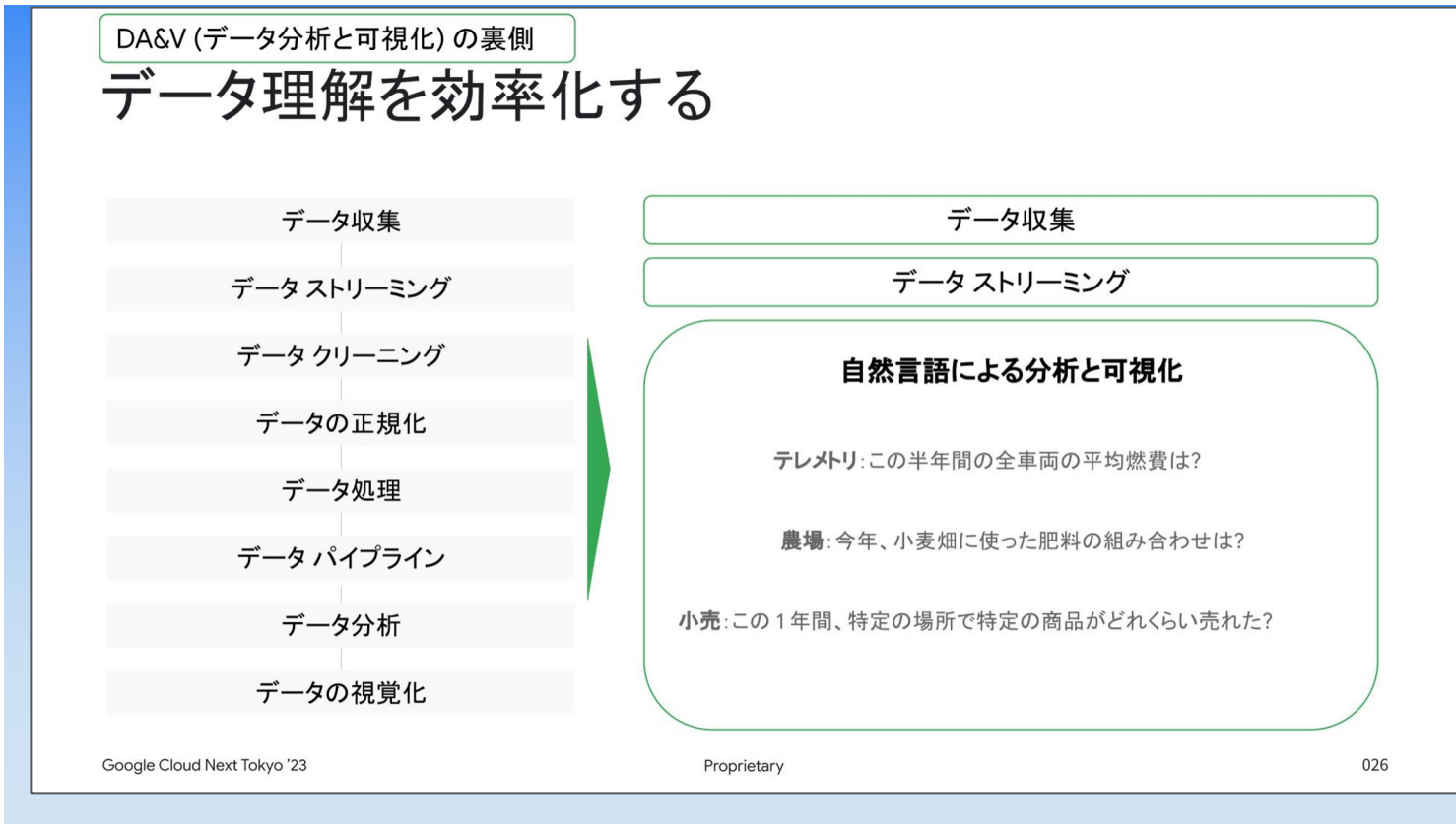
Manhattan Associates と GitLab における活用事例

拡張機能を活用したユースケースをもう1つ紹介します。

多くの企業には、長年をかけて収集してきた大量のデータがありますが、それらを活用するには、データ サイエンスや機械学習の知見を持つエンジニアの力が必要です。データのクリーニングや正規化、処理、パイプラインなど、複雑なプロセスが存在するからです。

Vertex AI Extensions を活用すれば、コードの書き方を知らない人物でも、業界を問わず自然言語での問いに答えられる環境を構築できます。例えば、テレメトリー分野では、この半年間の全車両の平均燃費を計算するということも可能です。農業分野では小麦畑に使った費用の組み合わせを視覚化できますし、小売分野では地域や商品を指定して、1年間でどの程度売れたのかを把握できます。

Vertex AI Extensions の使用によって、データの理解や分析・可視化が可能となるのです。



専門知識が不足している人物でもデータを活用できる

ここまで紹介してきた Vertex AI Extensions は、包括的なツールセットによって、エンタープライズ レベルかつ高品質な拡張機能を構築できる開発体験を提供します。具体的には、作成、構築、評価、最適化の 4 つのステップで、拡張機能を統合した LLM アプリケーションを迅速かつ容易に構築できます。

2024 年 1 月現在では、限定公開プレビュー版であり、引き続き正式ローンチに向けて開発を進めています。フィードバックなどを行える Trusted Tester Program も提供しているため、気になる方はぜひご登録ください。

参照リンク

1. Vertex AI 製品紹介ページ
2. LLM を強化する「Vertex AI Extensions」の活用・構築方法 アーカイブ視聴ページ

製品、サービスに関するお問い合わせ

goo.gl/CCZL78

Google Cloud の詳細については、上記 URL もしくは QR コードからアクセスしていただくか、同ページ「お問い合わせ」よりお問い合わせください。

Google Cloud