Session Report

Vertex AI で実装する
「ビジネス活用できる」生成 AI ソリューション

生成 AIソリューションを実装するためのVertex AIの最新情報

Google Cloud
ソリューション&テクノロジー AI/ML スペシャリスト
牧 允皓

Google Cloud

Google Cloud

セッション レポート概要

Vertex AI は、これまでの機械学習と同じく、生成 AI ソリューションの実装にも有用です。実験を繰り返したものの、ビジネス活用にいたらずに終わってしまうリスクを最小限に抑えられるよう、最初から本番を想定した環境で実験できます。今回は、生成 AI ソリューションを実装するための Vertex AI の最新情報について紹介します。

プレゼンター紹介

Google Cloud
ソリューション&テクノロジー AI/ML スペシャリスト
牧 允皓

Google Cloud の AI/ML スペシャリストを務めています。これまで構造化データの分析や、機械学習システムのビジネス実装に携わってきました。Google Cloud での専門領域は Vertex AI で実装する生成 AI や MLOps です。

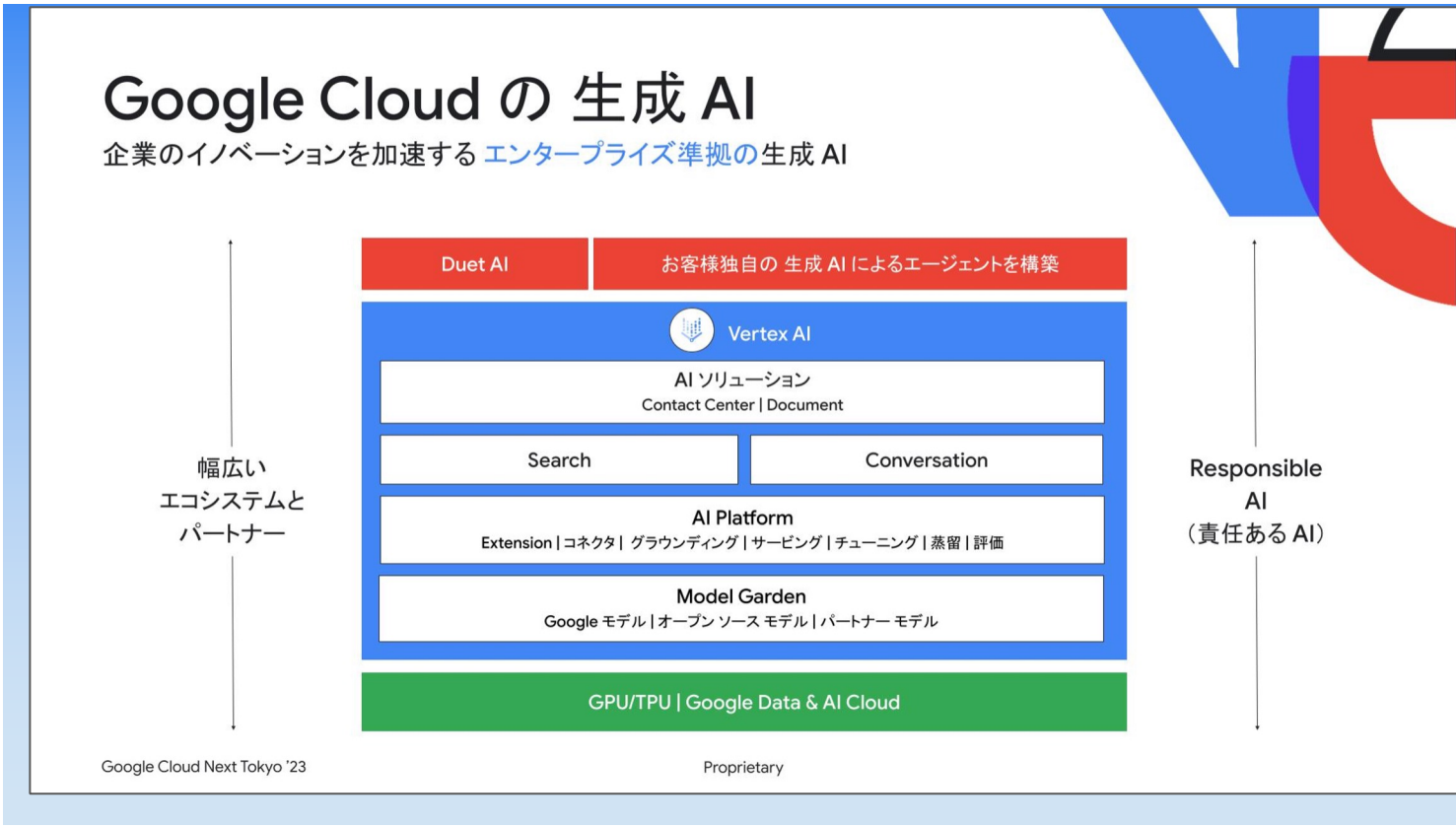
目次

- **生成 AI アプリケーションの構築を容易にする「Vertex AI」** 3
- **ユーザーの違いから見る「Vertex AI」のソリューション群** 5
- **事前学習モデルを有効利用する** 6
- **生成 AI のハルシネーション回避には別の技術が必要** 7
- **生成 AI でアプリケーションを開発するときのマネージド サービス** 8
- **モデルのカスタマイズのステップ** 10
- **LLM カスタマイズに役立つツール** 12
- **生成 AI を本番環境へ移行するには？** 13

生成 AI アプリケーションの構築を容易にする「Vertex AI」

2023 年は生成 AIが話題になりました。実際に触ってみてビジネス活用の価値を考えた方が多いのではないでしょうか。Google Cloud でも、生成 AI アプリケーションを構築できるソリューションとして「Vertex AI」を提供しています。

Vertex AI は、生成 AI が話題になる以前の 2021 年 5 月に発表されました。従来の AI/ML（機械学習）において、幅広いシナリオを実装できるプラットフォームとしてリリースされており、また「責任のある AI」や「幅広いエコシステムのパートナー」という理念を強く意識して開発されています。



Vertex AI のポートフォリオ

上図は、Vertex AI の簡単な全体像です。具体的なシナリオが既に決まっている場合、例えばコンタクトセンターを AI で自動化したい場合などは、そのシナリオに適した技術として、Contact Center や Document などが AI ソリューションとして提供されます。

上から 2 段目の Search と Conversation は生成 AI とつながりが非常に強く、事実を検索するための技術と、対話形式で情報を提供するための技術が用意されています。

上から 3 段目の AI Platform は、従来から Vertex AI が提供している古典的な機械学習技術にも当てはまりますが、AI を自身が所有するデータやアルゴリズムに当てはめるための機能が提供されています。

一番下にある Model Garden では、Google が開発してきたファーム モデルやオープンソースの Llama2 のようなモデルが提供されています。Google Cloud は、オープンであることを強く意識しているので、100 を超える数の OSS のモデルを提供しています。もちろん、チューニングも可能です。

こうした Vertex AI のさまざまな機能は、Google Cloud のインフラ上で動いています。私たちは AI が流行した初期から、今後大量の数値計算が必要になることを予測し、GPU や行列演算に特化したGoogle 独自のチップセットに投資してきました。

ユーザーの違いから見る「Vertex AI」のソリューション群

ここまで説明してきた機能は、どのようなお客様を想定して作られているのでしょうか。まずは下図をご覧ください。

ソリューション群によって AI 開発がより容易に

一番左側の Out of the box は、「箱から出して」という意味が示すように、そのまますぐに使える AI であり、事前学習 API や AI エージェントの機能を提供します。データやアルゴリズムは Google 側が提供します。

これまでは、例えば「画像の中の車」を検知するためには、大量のデータと大量の計算リソースが必要でした。しかし、それらをGoogle が事前に準備しているため、開発者は API をコールするだけでこれらを利用できます。

逆に一番右側の DIY は、データ サイエンティストや ML エンジニアが、自身のデータや知識を用いて独自の ML アプリケーションを開発するためのプラットフォームです。データおよびアルゴリズムをカスタマイズできるプラットフォームも提供されています。

中央にあるのは、両者の中間的なものであり、BigQuery MLやAutoMLなどが該当し、DIYよりは簡易にAIを構築する方法になっています。例えば、アルゴリズムや技術はGoogle が提供する物を使うがデータはユーザー自身のものを使っていただくというイメージです。

事前学習モデルを有効利用する

このように、Google Cloud ではすぐにお使いいただける AI ソリューションを実装しているため、大量のデータを用意したり、モデルを開発したりする必要はありません。ここで重要になるのが「非構造化データ」です。

画像、音声、テキストなど表形式で表しにくい非構造化データは、データの収集やラベリングにコストがかかります。もしここに事前学習モデルが使えるとしたら、コストを抑えて多様な検証を行うことができます。ビジネス活用における価値の有無を検証する際にも、初期投資を低減することができます。

実際に、どのような事前学習モデルが使えるかの種類を知っておけば、ビジネス上の課題解決を目指すときに、例えば「簡単な文章分類だから Vertex AI の AutoML を用いてテキスト分類すれば済む」というような発想を得られます。生成 AI を用いるより簡単に、より低コストで解決が見込める場合は、この事前学習モデルの利用を検討してください。

利用できる事前学習モデルの 1 つ目は「画像」です。AutoML Vision、Video Intelligenceなどがあります。2 つ目が「自然言語」。Translation、AutoML Translation、Natural Language などが提供されています。3 つ目は「会話」で、Dialogflow や Speech-to-Text などが一般に提供されています。

一般提供されている事前学習モデルの例

この事前学習モデルにおいて、特にテキストのタスク、例えばスパム検知あるいは機械翻訳などは、インターネットが普及した当初から長年研究が続けられてきました。

特に翻訳は、深層学習が流行った頃から多様な研究が行われてきた領域です。現在は LLM を使ったより新しいアプローチとして「Adaptive Translation」という機能を提供しています。

現在はまだプライベート プレビューの段階ですが、LLM の強みを生かしたデータセットを使うと、Translation を誘導できることが分かっています。例えば、機械翻訳を内製化しようと思うと大量の日本語と英語のデータセットで対になる文章を集めなければなりません。しかし、事前学習モデルを使えば、予測のためのデータ（推論データ）を与えるだけでよいのです。

生成 AI のハルシネーション回避には別の技術が必要

ここまでお話した通り、事前学習モデルは旧来の機械学習では有効です。では生成 AI ではどうでしょうか。もちろん生成 AI でもマネージドな機能として、アプリケーションやソリューション開発を加速させるための機能が提供されています。

生成 AI でアプリケーションを構築する際に注意していただきたいのが、ハルシネーションの問題です。LLM における生成 AI は、基本的に自然言語の単語分布を予測しています。例えば、「2023 年 5 月 16 日の東京の天気」という質問に対して、言語モデルとしては天気を返せば良いというのは理解できているのですが、今日の東京の天気が晴れか雨かは、知識として持っていません。そのため、「文法としては正しいが内容は間違っている」という答えを返してしまう可能性があります。

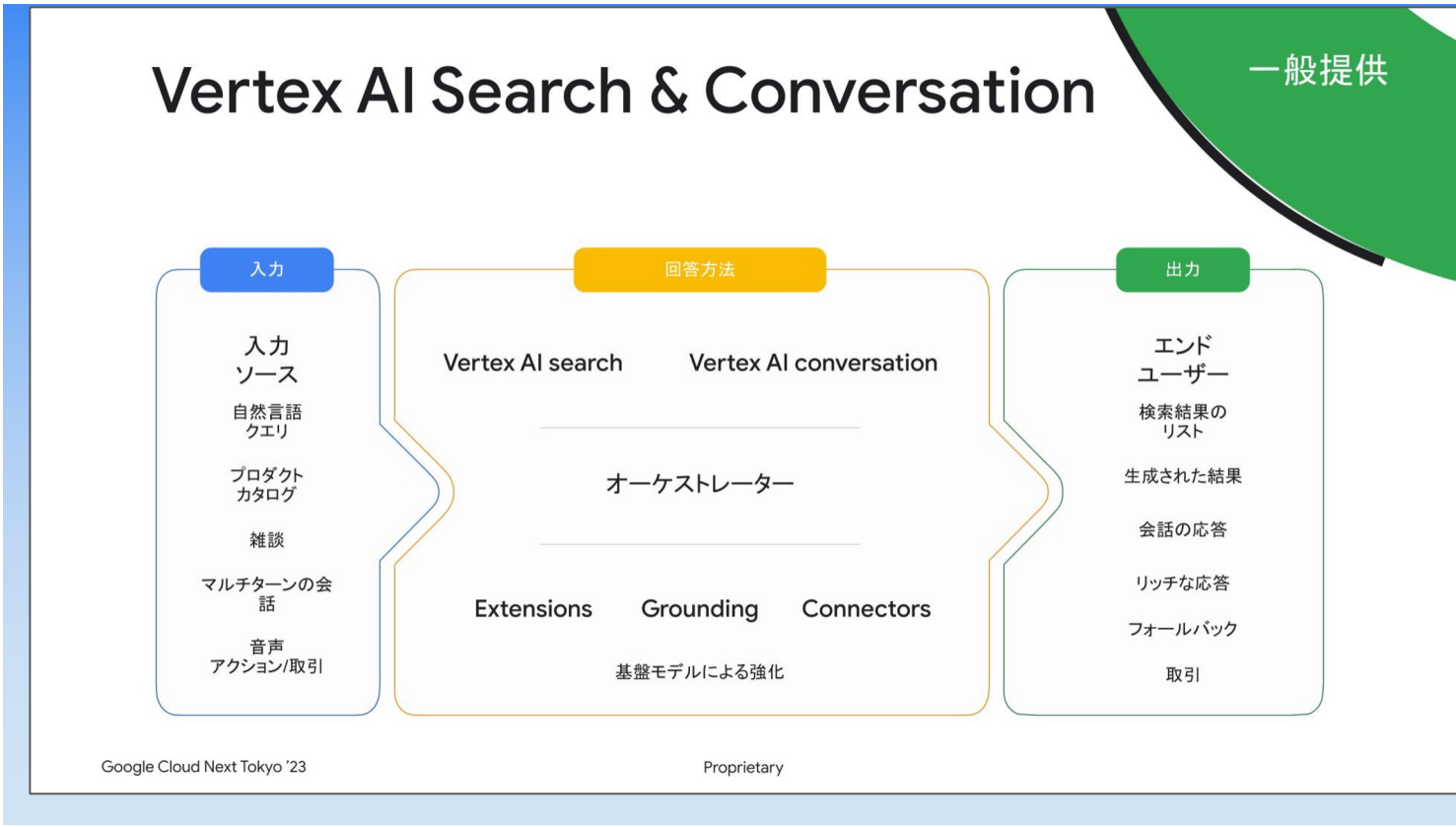
そこで、グラウンディングという技術が必要になります。グラウンディングによって、信頼できるデータソースからデータを参照するという行為ができるようになります。

LLM は、あくまでも大きなパラメータを持った言語モデルです。汎用的なタスクを解く際には有効ですが、個別のタスクを解く場合や事実に基づいた応答を返したい場合には、グラウンディングを始めとする別の技術が必要になってきます。

生成 AI でアプリケーションを開発するときのマネージド サービス

ここからは、生成 AI でアプリケーションを開発するときに利用できるマネージドサービスの具体例を紹介します。

一般提供が開始された「Vertex AI Search」「Vertex AI Conversation」では、入力と出力にあらゆるシナリオが想定されています。入力の場合は、今日の天気、昨日の売上といった自然言語のクエリが想定されています。一方、出力では検索結果のリスト、例えば昨日売れた商品をリストアップしてほしいといったニーズ、あるいは会話の応答などのリッチな応答が想定されています。



一般提供が開始された Vertex AI Search と Vertex AI Conversation

Vertex AI Search と Vertex AI Conversation では、あらゆる機能がオーケストレートされており、それらの機能が拡張できるようになっています。複数の機能を組み合わせてお使いいただくことで、あらゆるシナリオに対応できます。

また、サード パーティーのデータソースに対して、グラウンディングを可能にする「Vertex AI Data Connectors」がプレビューで準備されています。

新しく準備されている Vertex AI Data Connectors

さらに、基盤モデルによるアクションが可能になる「Vertex Extensions」もあります。チャットで会話をする際、自然言語の理解は当然 LLM によってできますが、このVertex Extensions には、LLM によってユーザー体験を改善した後に、実際のビジネス上のしかるべきアクションにつなげられることが重要です。皆さんの環境で独自の API があるという前提はありますが、LangChain との接続も簡単にでき、さまざまなユーザー体験へとつなげられます。

マルチターンのチャット インターフェースを実現するのが「Vertex AI Conversation」です。Vertex AI Search 同様、社内外のデータにグラウンディングして生成できます。

モデルのカスタマイズのステップ

ここまでお話したのは、どちらかというと「AI をどう使うか」、すなわち「use AI」の文脈でした。ここからはモデルのカスタマイズについてお話します。

マネージドな機能では物足りないという場合があると思います。例えば、出力される表現を自分たちの意図した表現にしたい、「丁寧な表現にしたい」、「カジュアルな表現にしたい」といった場合があるでしょう。

もちろん Vertex AI はエンド ツー エンドのプラットフォームなので、あらゆるシナリオに対応できる機能を取り揃えています。そして、実験環境も拡充されています。データやモデルをカスタマイズするための選択肢も、幅広く用意されています。

押さえておきたいのは、汎用的なモデルには限界があることです。汎用的な LLM はデータ収集の工数や学習コストがかからない反面、モデルサイズや推論の計算コストが増大化する傾向にあります。

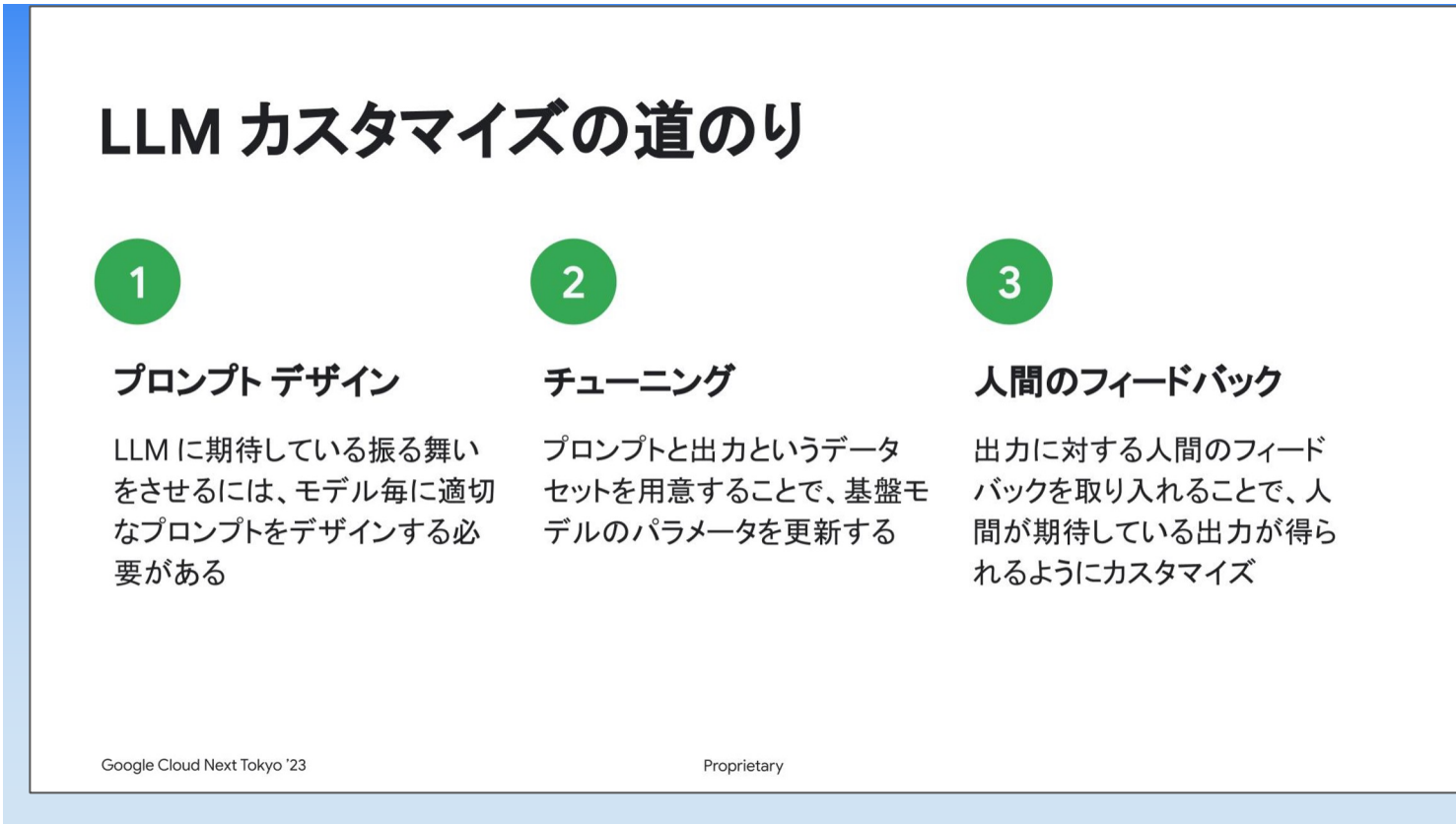
あらゆるタスクを解けるようにする以上、モデルの肥大化は避けられません。ただ、特定のビジネスのタスクを解きたいときには、すべて自力で解かなくてよいタスクに対して、高い精度を出す必要はありません。自身のタスクに特化したモデルにカスタマイズする方が、最終的にコストを下げられます。そのため LLM では、カスタマイズが必要になるケースがあるというわけです。

ここで重要になってくるのは、プロンプト デザインです。

我々が期待している振る舞いを LLM にさせるためには、モデルごとに適切なプロンプトをデザインする必要があります。他社の LLM モデルに使ったプロンプトをそのまま持ってきて、どちらの精度が優れているかが議論されることもありますが、本来は別々の言語モデルですので、同じプロンプトで評価するのはナンセンスです。

次に注目していただきたいのが、チューニングです。プロンプトの出力というデータセットを用意することで、基盤モデルのチューニングが可能になります。チューニングによって、目的の出力フォーマットに誘導できます。さらに、このチューニングは一部のパラメータだけを扱っているため、とても少ないデータセットで成立します。

最後に注目すべきは、人間のフィードバックによるパラメータのアップデートです。出力に対して人間のフィードバックを取り入れることで、人間が期待している出力が得られるようにカスタマイズしていくアプローチが提供されています。



LLM カスタマイズのためのステップ

LLM カスタマイズに役立つツール

これら 3 つの道のりを経て LLM をカスタマイズするときに役立つツールを紹介します。

1 つ目が「Generative AI Studio」です。GUI ベースのツールですが、Vertex AI のコンソール上から利用でき、プロンプトを変えたらどのように結果が変わるのかを、エンジニア以外でも簡単かつ高速に実験できます。アプリケーションに組み込む際には、データソースを生成できます。

2 つ目が、基盤モデルです。カスタマイズするときには、ベースとなる基盤モデルが重要になります。今提供されている PaLM はトークン数が増え、大量の入力にも対応できるようになりました。Imagen という画像系のプロダクトもあります。Codey や Embedding など、テキストのベクトル変換のみを解決するプロダクトもあります。

3 つ目は、チューニング可能なモデル群を扱う「Vertex AI Model Garden」です。冒頭でも触れましたが、Vertex AI Model Garden によって、さまざまなモデルをお使いいただけるようになりました。Google では、これまで開発してきた PaLM でも、あるいは OSS のモデルでも、学習チューニングまたはデプロイまでできます。

チューニングに関して、下図では 100 程度のサンプルデータで、モデル パフォーマンスを向上と記載しましたが、実はすべてのパラメータをアップデートしていません。アダプター チューニングという方法を用いて、一部のパラメータだけをチューニングすることで、短時間かつ、データ量を抑えてチューニングできています。

PaLM、Imagen、Codey で提供されている幅広いチューニング機能

生成された結果 A と生成された結果 B のどちらが良かったかを人間がフィードバックすることで、徐々にそれに準拠するような形でパラメータがアップデートされていきます。

生成 AI を本番環境へ移行するには？

最後に、生成 AI を本番環境へ移行する方法を紹介します。AI/ML は、従来の CI/CD とは少し異なる性質を持っており、継続的に開発あるいは学習を管理しなければならないアセットが多くなります。

データ サイエンティストが、「どのようなアルゴリズムなら期待しているパフォーマンスが達成できるか」という実験をしたときに、開発するのは機械学習のコードのみでしょう。しかし、これを運用するとなると他にもさまざまな疑問が出てきます。

例えば、どういう設定で学習したのか、サービングするときにどのようなインフラにデプロイするのか、マシンタイプは何なのか、データ収集をどこで行っているのか、どういう頻度で実行しているのか、異常値の有無に関するデータ検証をどのようにしているか、デプロイしている機械学習のアプリケーションが適正に機能しているのか、などさまざまです。

このように、あらゆるリスクを想定して本番環境を作らないと、実験では上手くいっても、ビジネスで活用するときには、期待した効果を得られないという結果に終わってしまいます。

ここで重要になるのが、評価とモニタリングです。

デプロイをしたあとのモデルを、そのまま放置するわけにはいきません。去年のデータを使って学習したモデルは、来年には使えなくなっている可能性があるからです。構築したモデルが今も期待しているパフォーマンスを出しているかどうかのモニタリングと評価が必要であり、Vertex AI でもそれらの機能を提供しています。現在では、その機能を生成 AI のシナリオにも拡張しています。下図は、Vertex AI で提供される MLOps の機能をまとめたものです。

MLOps が実現できること

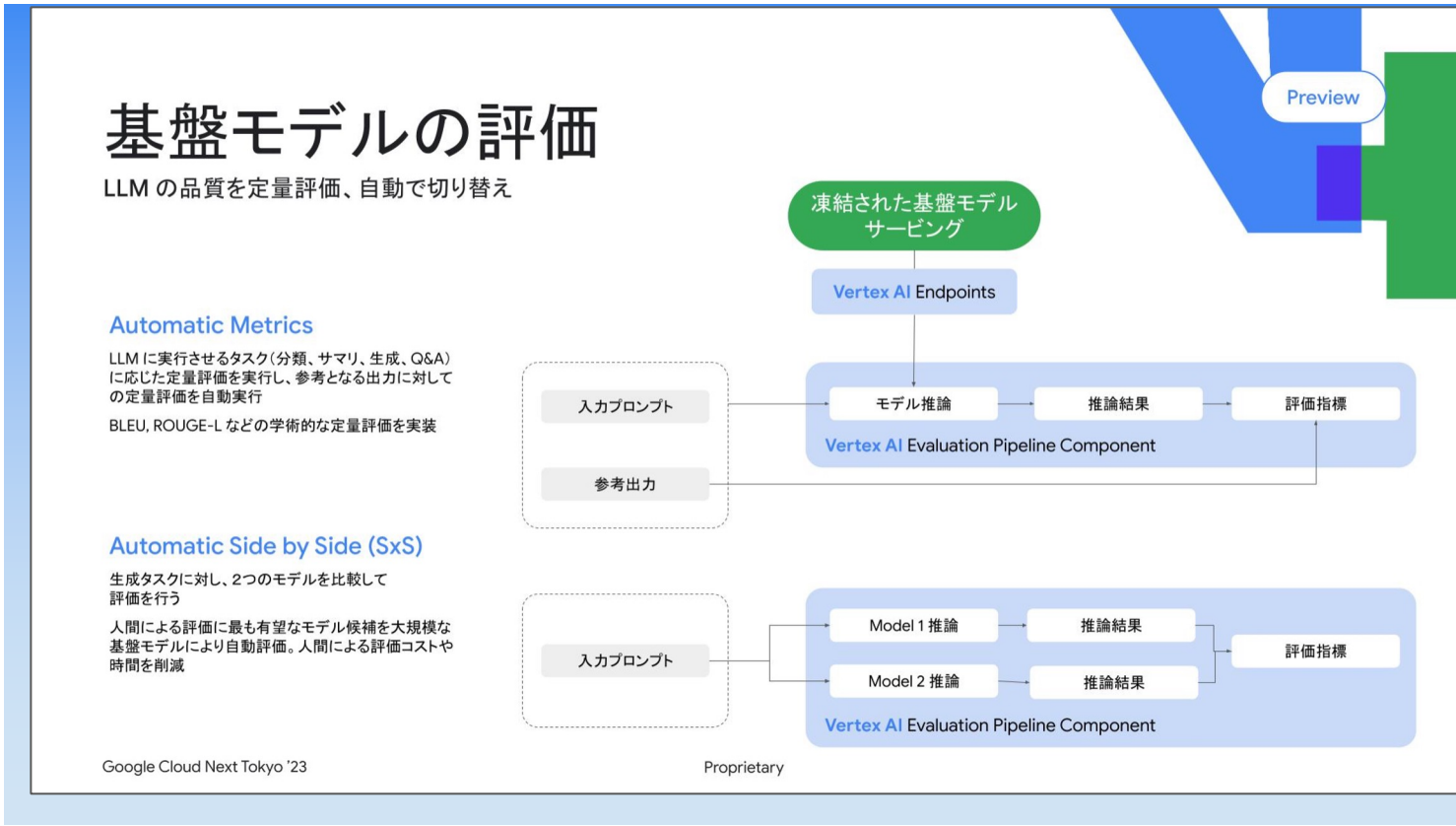
1 つ目の「プロンプト デザイン」では、プロンプトのインポートやエクスポート、共有が簡単にできるようになっています。これによって、コピー & ペーストをしたときに、一部のインストラクションが抜けて精度が落ちるリスクを回避できます。

2 つ目の「チューニングの管理」に関しては、スーパーバイズド チューニングや強化学習を用いたフィードバック、人間のフィードバックを使った強化学習モデルなどが、すべて「Vertex AI Pipelines」で実行されます。Vertex AI Pipelines は、従来の機械学習でも OSS でも、すべてオーケストレーションされているので、実験内容、担当者、実験結果を 1カ所に集約できます。

3 つ目の「チューニングされたモデルの管理」を担うのが「Vertex AI Model Registry」です。パフォーマンスの評価を Model Evaluation で行えば、迅速かつ定量的な評価が可能になります。

基盤モデルの評価について、例えば去年構築したモデルの精度が 90% で、今年新しいデータを使って学習し直したところ精度は 85% になったとします。このとき、単純に 85% と 90% という数値で優劣をつけることはできません。去年のテストデータと今のテストデータが異なっている可能性があるからです。そのようなテストデータの違いを踏まえたうえで、Vertex AI Model Evaluation を使えばモデルを管理できます。

ここで使える機能が、「Automatic Metrics」と「Automatic Side by Side（SxS）」です。これらは、別の関数を Vertex AI で提供することによって、モデル 1 とモデル 2 の優劣を定量的に比較できます。これらの機能の一部は、現段階でプレビュー機能になっていますが、長期的な視点で開発できるよう努めていきます。



基盤モデルの評価に適した機能が一部公開

ここまで、生成 AI ソリューションを実装するためのVertex AI の最新情報をお話してきました。興味のある方は、ぜひ活用していただければと思います。

参照リンク

1. Vertex AI 製品紹介ページ
2. 生成 AI ソリューションを実装するための Vertex AI の最新情報 アーカイブ動画視聴ページ

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