Session Report

AI へのニーズが増大する中で
継続的な監視と評価で精度を維持する

生成 AI 時代の MLOps 実現方法とは？

Google Cloud
カスタマー エンジニアリング データ アナリティクス スペシャリスト
高村 哲貴

Google Cloud

Google Cloud

セッション レポート概要

生成 AI モデルのカスタマイズは、AI/ML モデルをゼロから構築するより簡単かもしれませんが、その運用においては共通の課題が多く見られます。ここでは MLOps パイプラインの最適化や AI モデルの管理、評価、セキュリティ フィルターの活用などを通じて、生成 AI 時代に MLOps をどのように適応させるかを解説します。

プレゼンター紹介

Google Cloud
カスタマー エンジニアリング
データ アナリティクス スペシャリスト
高村 哲貴

日系 SIer 企業にて、データベース エンジニア、クラウド アーキテクトを経て、2021 年より現職。現在は、Google Cloud のデータ アナリティクス領域のスペシャリストとして、お客様のデータ活用を技術観点から支援。

目次

生成 AI によって変わる MLOps

Google Cloud では、MLOps のことを「ML（機械学習）システムを迅速かつ確実に構築、導入運用するための標準化されたプロセスと機能のセット」と表現しています。MLOps には、一般的なシステム運用とは異なる特有の難しさがあります。それが AI の精度の維持です。

例えば、何かを予測する AI モデルを活用している場合、事前に準備したデータで学習モデルを作り、精度を評価したうえで問題がないものをリリースする手順が一般的です。しかし、データは日々刻々と特徴を変えるため、常に高い精度が維持できるとは限りません。MLOps では、こうした変化に対応しながら一定の精度を保ちながらビジネス価値につなげる必要があります。

では、これらの課題に対して生成 AI はどのような変化をもたらすでしょうか。重要な 2 つのポイントに、「既存の MLOps への投資を捨てる必要はない」「生成 AI 特有のニーズを理解する」が挙げられます。後者の生成 AI 特有のニーズについて、具体的には以下の通りです。

1.マルチタスク モデルやプロンプト、増大する AI インフラストラクチャ ニーズへの対応

生成 AI のモデルは、テキスト生成や画像生成など用途に応じて非常に多くのモデルが存在します。このようなマルチタスクなモデルに対してどれを選ぶか、そしてどのようなプロンプトを与えていくかが大切です。

2.チューニングと厳選されたデータによるカスタマイズ

既存の大規模言語モデル（LLM）の活用だけではなくて自社のデータや、特定のドメインのデータをつなぎ合わせてどのようにチューニングやカスタマイズするかが問われます。

3.新しいアーティファクトの管理

プロンプト、エンベディング、アダプター層など、生成 AI によって誕生した新しいデータや概念が加わります（ここではアーティファクトと呼びます）。これらを適切に管理しなければなりません。

4.生成された出力の評価と監視

生成 AI によって生成されたアウトプットに対して、どのように評価・監視していくかを考えなければなりません。

5.エンタープライズデータとの接続

ハルシネーション（事実に基づかない誤った情報をAIが生成してしまうこと）を回避する手段として、自社のエンタープライズ データに接続したいというニーズがあります。自社データと LLM をつなぐことで、より特定のドメイン知識を持ったり、最新の情報を踏まえた回答やアクションを AI が行えます。こうした特徴が、今までの MLOps と異なる点です。

このようなニーズに対して、Google Cloud の AI ソリューション「Vertex AI」は予測 AI（Predictive AI）だけではなく、生成 AI も含めた統合プラットフォームとして、生成 AI 向けの固有のニーズに対応するためにアップグレードを続けています。その具体的な機能は次項から解説していきます。

従来の MLOps と生成 AI 時代の MLOps の違い

生成 AI 時代の MLOpsは、従来の MLOps と具体的にどのように変わってくるのでしょうか。

下図は従来の MLOps のフレームワークです。簡易的ですが、実験・トレーニングを行い、次にデプロイして予測します。それらを支えるガバナンスのレイヤーと、それを評価・モニタリングするレイヤーが存在します。

従来の MLOps フレームワーク

この既存の MLOps のフレームワークに対して、生成 AI によって変化したものを示したのが下図です。生成 AI における MLOps を考える上で必要な概念として、まずディスカバリ、プロンプトの 2 つがあることが分かります。

「ディスカバリ」と「プロンプト」に従来との違いが見られる

「ディスカバリ」とは、利用可能な LLM の中から、それぞれのユースケースに合うモデルを探し出すことを示します。「プロンプト」については、生成 AI が返すデータの精度を向上させるためにインプットをどのように扱い管理すればよいかというということです。これに対して Google Cloud では、それぞれに対応する以下の2 つのソリューションを提供しています。

1．Vertex Model Garden

Vertex Model Garden の中には、ファースト パーティのモデルだけではなく、サードパーティも含めた非常に多くのモデルが登録されています。例えば、Llama2 や Stable Diffusionなどのモデルも含めてそれぞれのユースケースに合うモデルを探すことができます。

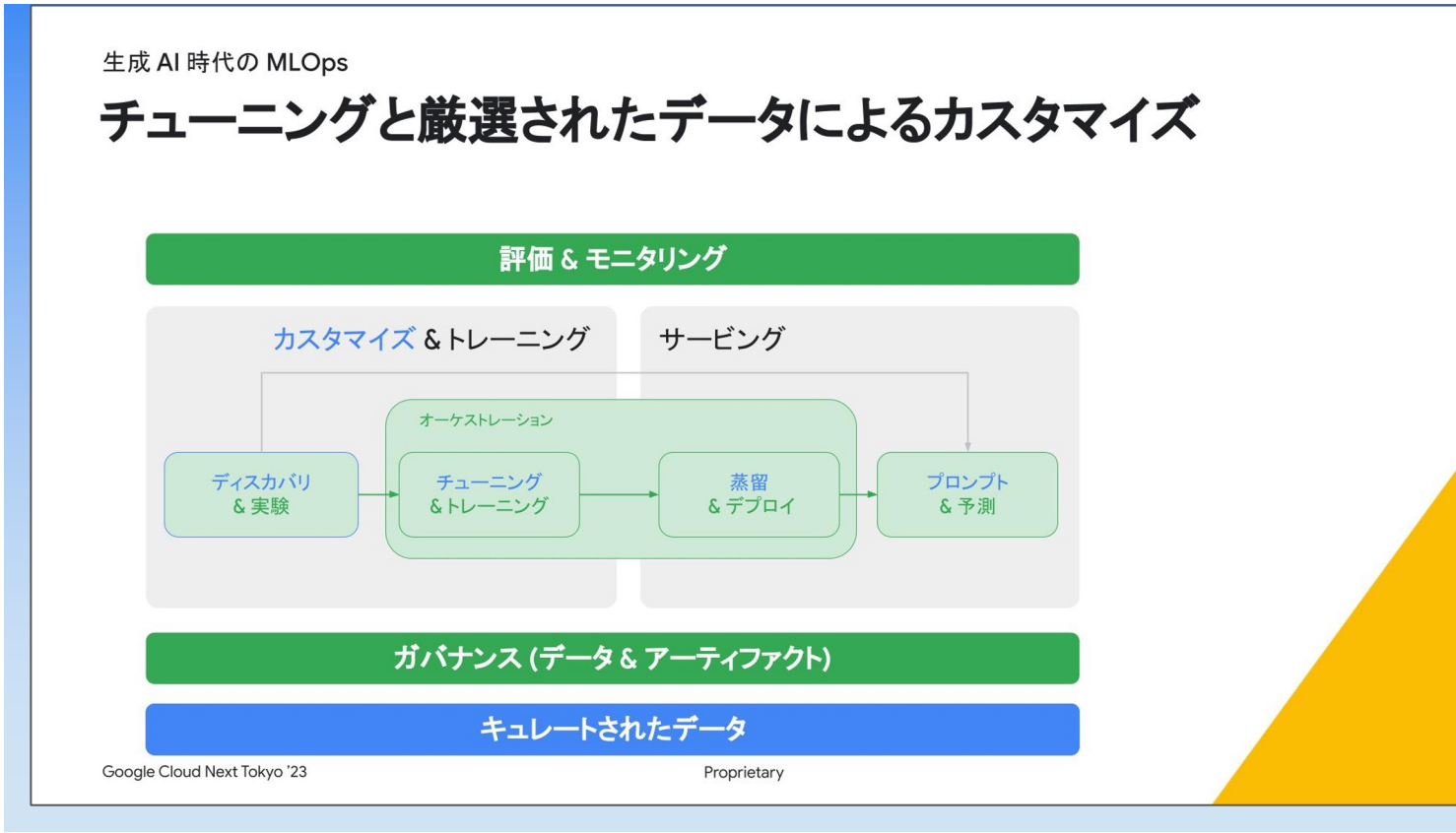
2．Vertex Generative AI Studio

自身が選んだモデルに対して、どのようなプロンプトを与えたことでどのようなアウトプットがされるのかを確認することも、スムーズなモデル選定に欠かせません。Vertex Generative AI Studio は、非常にシンプルな操作と GUI の画面によって、選定したモデルに対して、好きなプロンプトを入力し、そのアウトプットを確認するという行為を簡単に行えます。

LLM の選定とプロンプトの管理を簡便化する2つの機能

AI モデルのチューニングの重要性

もちろん、生成 AI によって MLOps に生じる変化は上記だけではありません。このほか注目したいのが、下図に示す「チューニング」「蒸留」「キュレートされたデータ」というキーワードです。蒸留とは、大規模なモデルを教師モデルとしてその知識を小さいモデルに学習させることを意味し、「モデル圧縮」とも呼ばれます。



生成 AI ならではの MLOps で押さえておきたいチューニング、蒸留、キュレートされたデータ

この中のチューニングに関して、Google Cloud では、先述の Vertex Generative AI Studio を用いて 2 種類の方法で実現することができます。

1 つ目は教師あり学習です。ファイン チューニングと呼ぶ場合もあります。各々が持っている特定のタスクに特化させるために必要なデータを学習させることで、より特定の知識を持ったり、出力の振る舞いを変えたりすることができます。

もう 1 つが人間のフィードバックからの強化学習（Reinforcement Learning Human Feedback：RLHF）です。生成 AI 系のアプリを使う際によく目にすると思いますが、回答に対してサムズアップ・サムズダウンを選択したり、または 2 種類の回答が出てきた際に、どちらが好みかを選択したりする行為です。そのモデルをより人間の好みや価値観に近づけるチューニング手法です。

Vertex Generative AI Studio を用いた 2 つのチューニング方法

データ キュレーションの側面では、例えば、特定のドメインに特化した学習をさせたければ、そのドメインのコーパスとなるデータや専門的なタスクのデータ、RLHF のための人間のフィードバックのデータなどを準備して学習させることでチューニング可能です。

新しいアーティファクトの管理が必要に

生成 AI における MLOps では、新しいアーティファクトを適切に管理していかなければなりません。その具体的な対象は、チューニングのジョブや、エンベディング、チューニング済みモデルなどです。

Google Cloud では、これらを管理・実行するためのソリューションを提供しています。例えばチューニング ジョブに関しては、「Vertex Pipelines」を活用します。これは従来の MLOps から使用されているソリューションですが、生成 AI のチューニングに対しても有効です。

次に、チューニングされたモデルのラインアップやバージョン管理には「Vertex Model Registry」が有効です。モデルの管理のほか評価までを手軽に実行できます。

最後に、エンベディングに関しては、「Vertex Feature Store」を活用できます。エンベディングされたデータの保存・管理・展開を容易に行えます。

新しいアーティファクトと対応する Google Cloud のソリューション

生成されたアウトプットの評価 & モニタリング

次に、従来の MLOps から存在する「評価 & モニタリング」のプロセスの中で、生成 AI ならではの観点として「メトリックス」の重要性に言及したいと思います。主に以下の 3 つが挙げられます。

(1) 評価メトリックス

生成 AI がどれだけ精度が高いか、信頼できるか、正確であるかといったさまざまな観点から、評価を定量的に行う必要があります。精度や正確性、好みといったさまざまな評価の概念があるので、あるタスクをどのメトリックスで評価するべきか考える必要があります。

(2) 安全性スコア

Google Cloud は責任ある AI（Responsible AI）というものを差別化要素として考えており、それに代表される機能がこの安全性スコアです。例えば危険性の高いものが評価されたり、アウトプットされたりすることをブロックします。例えば、爆弾の作り方に対して生成 AI から正確な回答が返ってしまえば、犯罪の温床になってしまうでしょう。

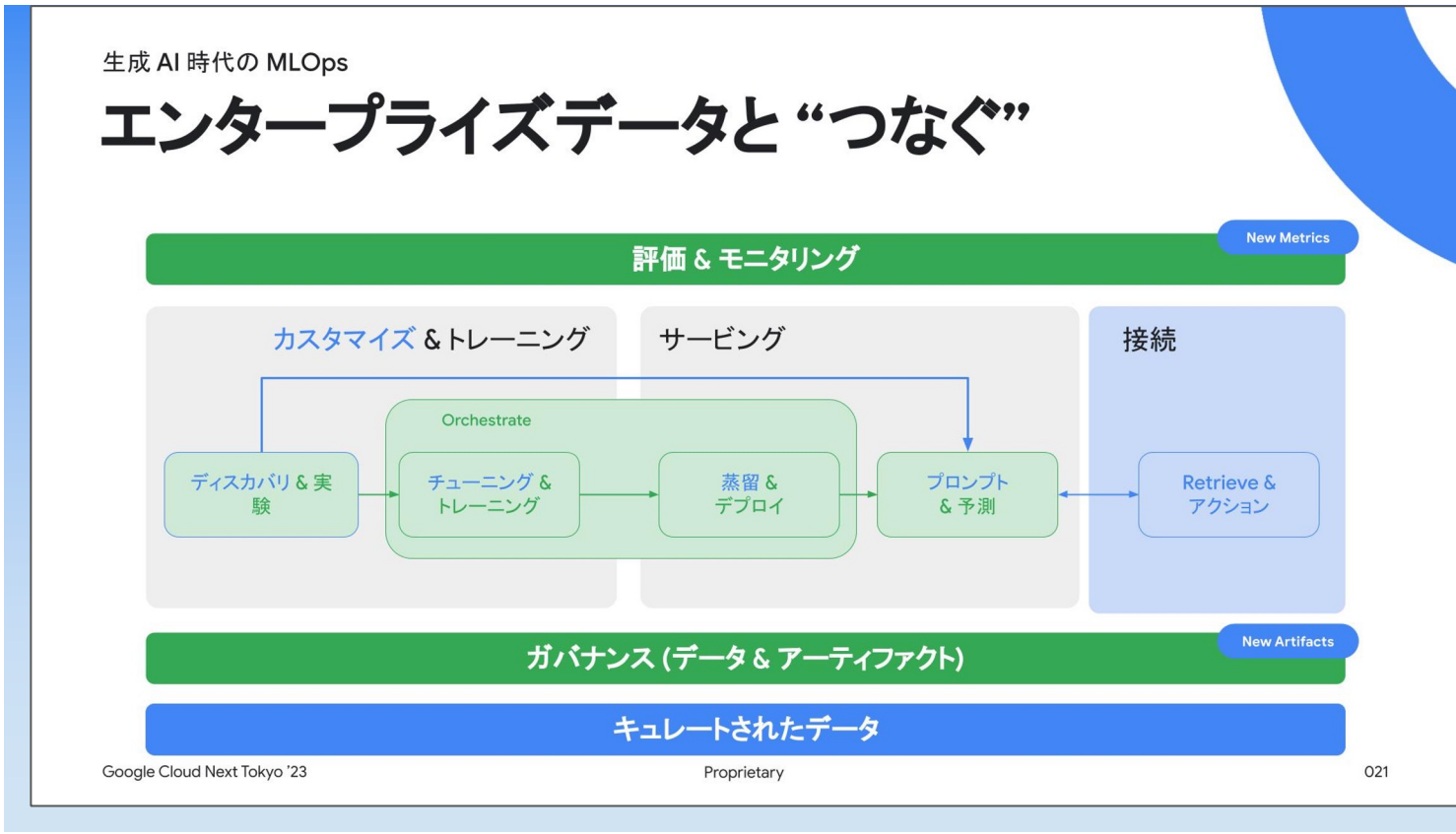
(3) 引用（Recitation）チェック

生成されたコンテンツが引用されていないかどうかをチェックするものです。Google Cloud が提供する AI は、著作権を保護するために、基本的に既存のコンテンツを複製しないような形で動作します。その機能を担うのがこの引用チェックです。

生成されたアウトプットの安全性判定や評価を実行する 3 つの機能

エンタープライズ データとの接続

生成 AI ならではの MLOps のフレームワークとして、最後に言及したいのが、下図の「接続」のプロセスにある「Retrieve &アクション」です。



エンタープライズ データとの接続を考慮する必要がある

先述したように、社内の情報検索などで生成 AI を使えるように、自社固有のデータと LLM を接続したいという強いニーズが多くの企業から寄せられています。これを実現するにあたり Google Cloud では 3 つの観点が必要だと考えています。

(1) エンベディングとベクトル検索

テキストや画像などマルチモーダルに対応したエンベディングの仕組みをもつことで、より高度化されたセマンティック検索が可能になります。さらにベクトル検索の技術によって低レイテンシーかつスケーラブルな検索を実現できます。Google Cloud では、具体的にこのためのソリューションとして「Vertex Embeddings APIs」を提供しています。

(2) グラウンディング

グラウンディング（Grounding）とは、生成 AI が外部の情報を文脈として理解し、その情報をもとに正確な結果を返す手法です。Google Cloud でも、ファースト パーティのデータだけでなく、サードパーティのデータと接続してグラウンディングに関する技術のケイパビリティを高めようとしています。

(3) エクステンション

これは、例えば LLM の回答やチャット、外部の予約システムなど複数のシステムを組み合わせることでシームレスなユーザー体験を提供するというように、現実世界のアクションとAIを接続しようとする考え方です。

Google Cloud の生成 AI 評価サービス

最後に、Google Cloud の生成 AI 評価サービスについて紹介します。生成 AI の正しさや正確さ、安全性が評価できなければ、お客様のビジネスに直接大きなインパクトを与えることになります。それゆえに「評価は MLOps ライフサイクル全体のコア」といっても過言ではありません。正しく評価やモニタリングを行いながら、生成 AI が本番環境で稼働する状態を作ることが、MLOps を考える上で、最も重要なポイントになります。

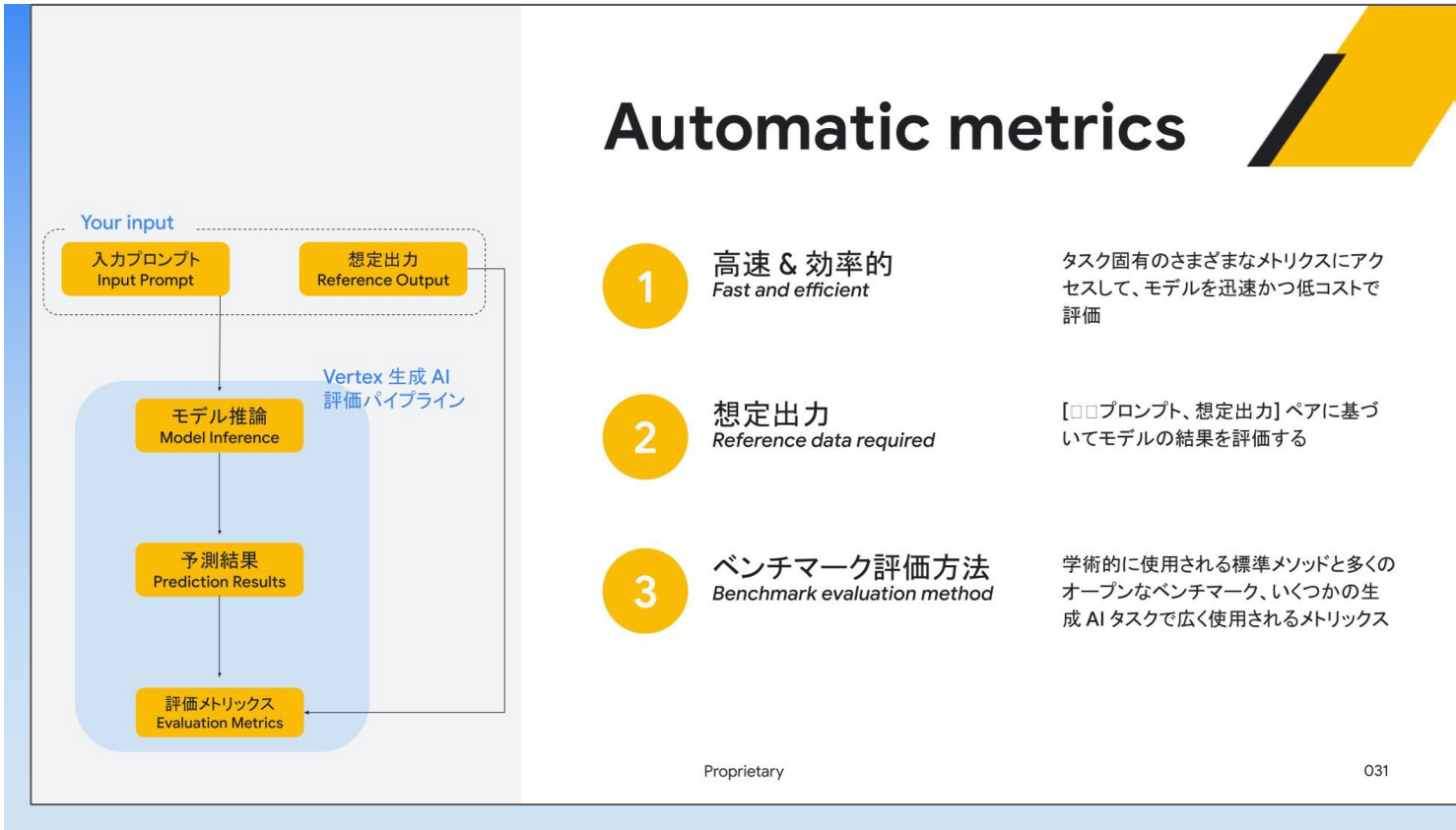
Google Cloud では、生成 AI の評価に関して、「Automatic Metrics」「AutoSxS」「Safety Bias」の 3 つのサービスを発表しています。それぞれを以下にて解説します。

生成 AI を評価する 3 つのサービス

(1) Automatic Metrics

さまざまな生成 AI のタスクやその固有のタスクに最適な評価のメトリックスを、Google Cloud が自動で選定します。これにより AI モデルを定量的に評価することが可能です。例えば、要約なら「ROUGE-LSum」、またはテキスト生成なら「BLEU」など、適切なメトリックスの選定とそれに対する定量評価を自動で実行します。

具体的には、入力にはプロンプトとグラウンド トゥルース（正解データ、想定している出力のデータ）を与えます。そうすると、Google Cloud の中でモデルの推論と予測結果、またそれに対する評価メトリックスのスコアリングが提示されます。



複数存在するメトリックスから適切なものを自動で選定してくれる

(2) AutoSxS （オートサイドバイサイド）

これは、AB テストのような形でモデル A・モデル B のうちどちらの精度が高いかを提示するサービスです。

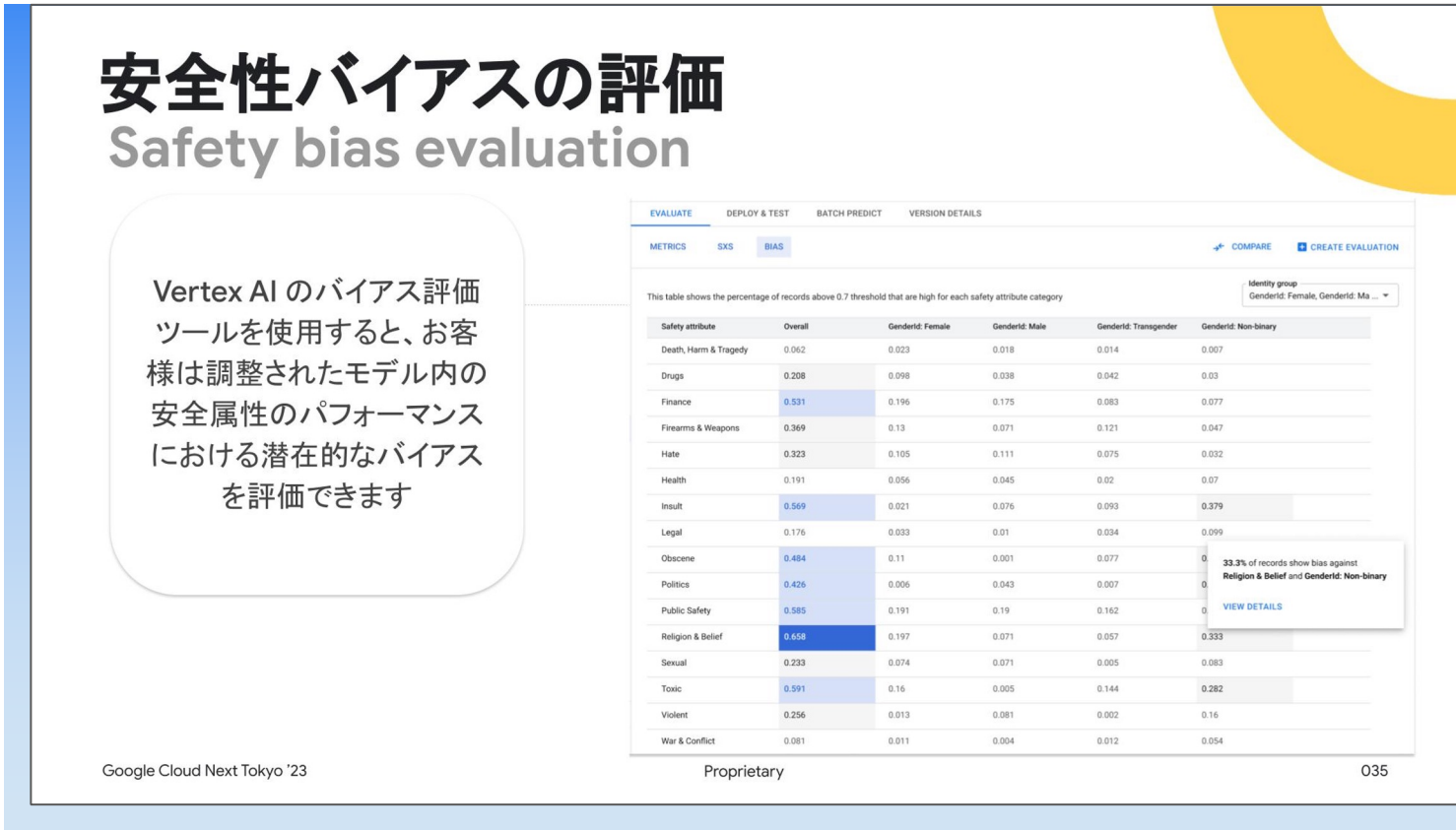
こちらも、まずは入力としてプロンプトを与え、さらに評価したいモデルを 2 つ選びます。各モデルに対して推論と予測結果が自動的に算出されるので、次に Arbiter モデルという当社が持つ第 3 の評価モデルを使って評価リクエストを送ることで Win レートが算出されます。
Arbiter モデルによる評価は、Google が実際に社内でも使っている評価の手法です。

LLM の評価に特化した AB テスト

(3) Safety Bias

これは今後ローンチされる予定の機能であり、安全性バイアスを評価するものです。さまざまなアイデンティティ（ID）グループに分けながら、グループの中でどのスコアリングが良いか悪いか、つまり AI の安全性を事前に測定します。

下図では横軸、つまり列で ID グループが分かれており、それぞれのグループごとに AI モデルがどのような評価であるか（特定のグループで AI が提示する回答にバイアスや偏りがないかなど）を把握できます。皆様が作成されたチューニング済みモデルが本当に安全かどうか、特定の ID グループで危険なスコアが出ていないかを確認できます。



安心して利用できる AI かどうかを事前に確認できる

今回は、生成 AI 時代こそ留意しておくべき MLOps のニーズについて触れながら、それに対応する Google Cloud の Vertex AI のソリューションも紹介してきました。Vertex AI は既存の MLOps のケイパビリティを拡張しながら、生成 AI のニーズに対応するように進化を遂げているので、お使いの既存の MLOps 投資を捨てる必要はありません。ぜひ、皆様の MLOps または生成 AI 活用の取り組みにご活用いただければと思います。

参照リンク

1. Google Cloud のジェネレーティブ AI の概要
2. 生成 AI 時代の MLOps 実現方法とは？ 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