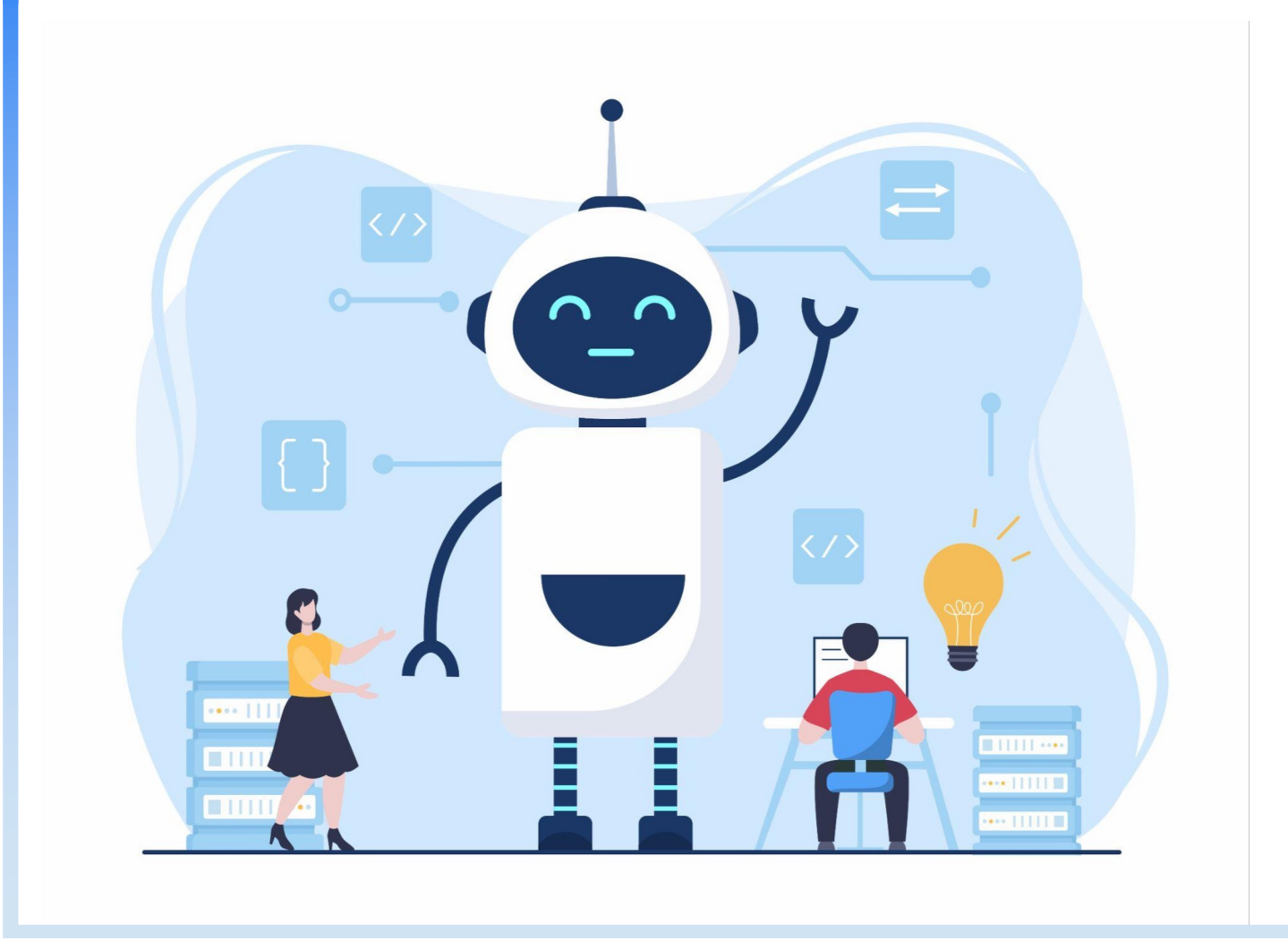


Session Report

開発に特化した生成 AI サービスの活用で
エンジニアの業務生産性を向上させる

これからはじめる Vertex AI Codey と Duet AI のアプリ開発

Google Cloud
S&T Application Modernization Specialist
諏訪 悠紀

Google Cloud

Google Cloud

セッション レポート概要

生成 AI が広がりを見せる一方で、その波に乗り遅れている開発者も少ないないかもしれません。そうした方に向けて、本稿では Vertex AI Codey や Duet AI for Workstations といった開発者向けサービスに触れ、その特徴や活用シーン、活用方法を解説します。

プレゼンター紹介

Google Cloud
S&T Application Modernization Specialist
諏訪 悠紀

Google Cloud の アプリケーション モダナイゼーション スペシャリスト。スマホアプリ開発やサーバーサイド アプリケーション開発を経験したのち、現在は Tech Acceleration Program などを通してサーバーレスやコンテナなどを活用したアプリケーション開発の技術支援をしている。

目次

開発を支援する生成 AI サービス

まず、開発を支援する生成 AI サービスの概要について解説します。Google では、生成AIのサービスをコンシューマー向けとエンタープライズ向けに分けて提供しています。

例えばコンシューマー向けの場合「旅行の計画を立ててほしい」や「パンケーキの作り方を知りたい」といったプライベートなユースケースが考えられます。一方でエンタープライズ向けの場合は、「データをどのように管理したら良いか」や「コストをどのように管理したら良いか」といった、より企業の実務向けのニーズが想定されます。

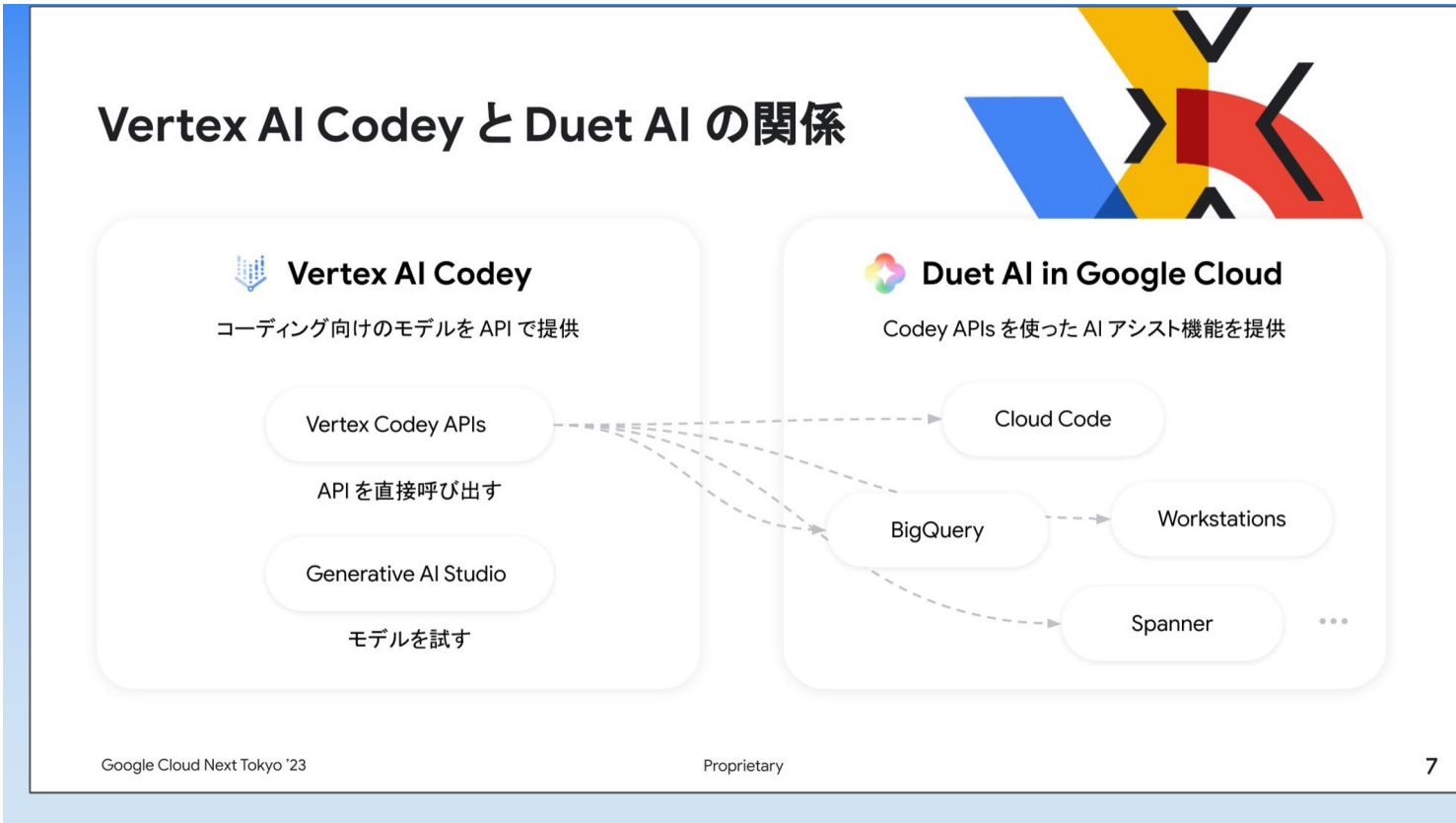
後者の場合、社内データの公開範囲やポリシー、生成されたデータの信憑性など、エンタープライズとしての要件が求められるようになります。

具体的に、コンシューマー向けでは、「Bard」のほか、モデルをカスタマイズ検証する環境として「MakerSuite」を、エンタープライズ向けでは「Vertex AI」、「Duet AI」といったサービスを提供しています。

Google が提供する生成 AI 関連に関連するサービス

Vertex AI Codey と Duet AI in Google Cloud の関係

エンタープライズ向けのうち、Vertex AI と Duet AIはそれぞれどのような役割があるのでしょうか。下図は Vertex AI Codey と Duet AI in Google Cloud の関係性を示したものです。



2つのサービスは別のものだが、関連性もある

Vertex AI Codey は、コーディングに特化したモデルを API で提供するサービスで、API を直接呼び出すことによってコードの生成やコードの補完といった機能を利用できます。この中に含まれる「Generative AI Studio」は、Google Cloud コンソールからアクセスできる GUI から Codey の機能を試せるものです。

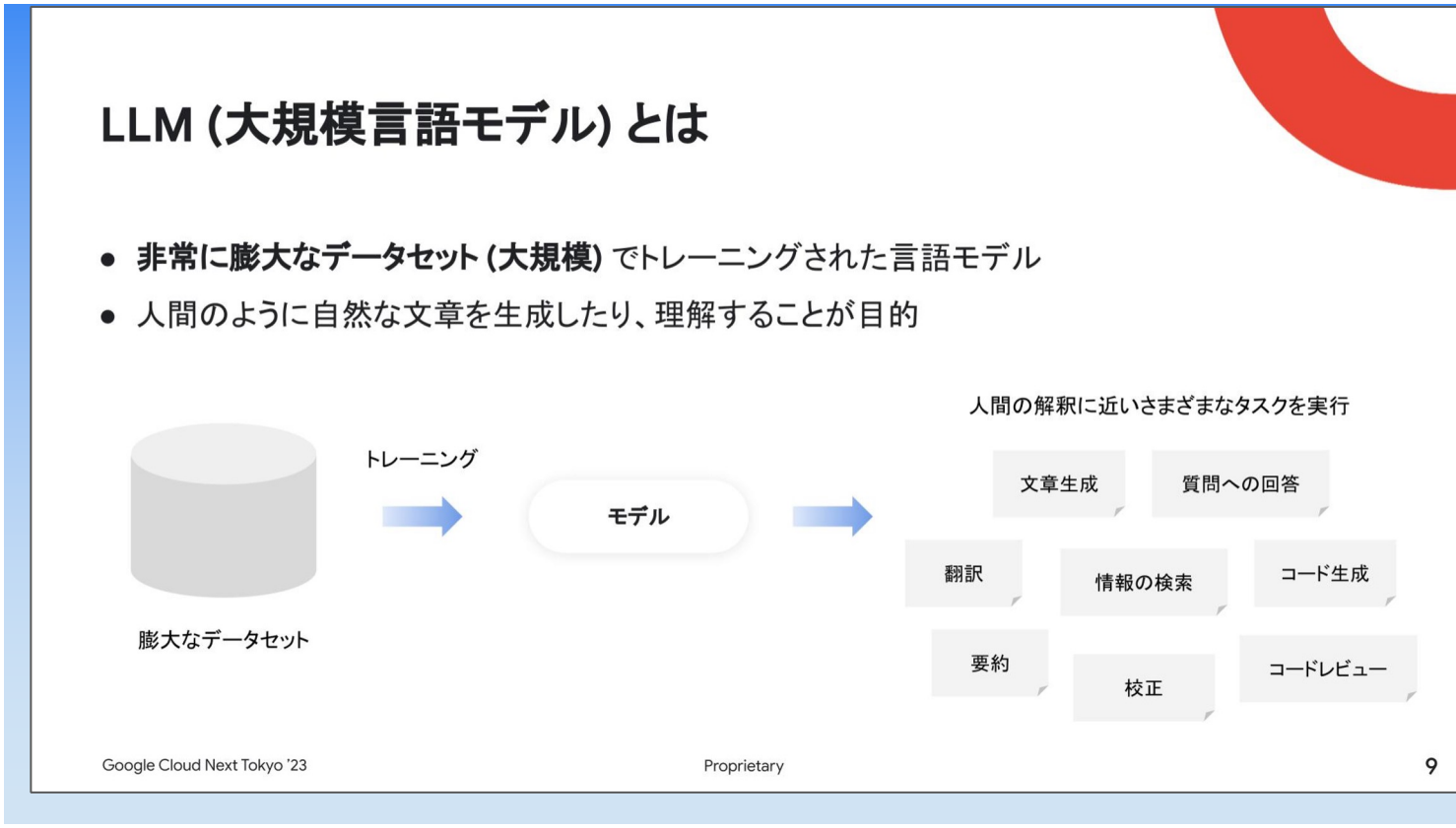
一方、Duet AI in Google Cloud に関しては、Google Cloud の AI アシスト機能の総称とイメージするとわかりやすいでしょう。具体的には、BigQuery のクエリ エディター内でのクエリの自動生成や IDE 経由でのコード生成、「Cloud Logging」のログの解説など、Google Cloud を使う上でのアシスト機能を提供しています。そして、これらの AI アシスト機能の一部に Codey が利用されているという関係性があります。

これらの使い分けについては、例えば、生成 AI を使う際に普段の業務を改善したいという場合には、Google Cloud 上で生成 AI 系の AI アシストの機能を使用できます。一方で Codey の場合は、API で提供しているため、生成 AI を活用した開発支援ツールを自分で作りたいといったような場合や、業務の開発のプロセスの中に直接 API を叩いて、生成 AI の機能を取り入れたいといった場合に使用するイメージとなります。

Google が提供する LLM である PaLM

ここで、生成 AI の基礎知識についておさらいしたいと思います。LLM（Large Language Modelは、直訳すると大規模の言語モデルであり、読んで字のごとく、非常に膨大で大規模なデータセットによってトレーニングされた自然言語処理モデルです。この大規模なデータセットを使ってトレーニングしたモデルが、人間のように自然な文章を理解し、文章を生成する AI を実現します。それが生成 AI です。

このように、あたかも人間のような理解力を持つようになったことで、文章の生成や要約、コードの生成や質問の回答、情報の検索などこれまで人間が行ってきたタスクの一部を任せられるようになりました。



LLM の概要

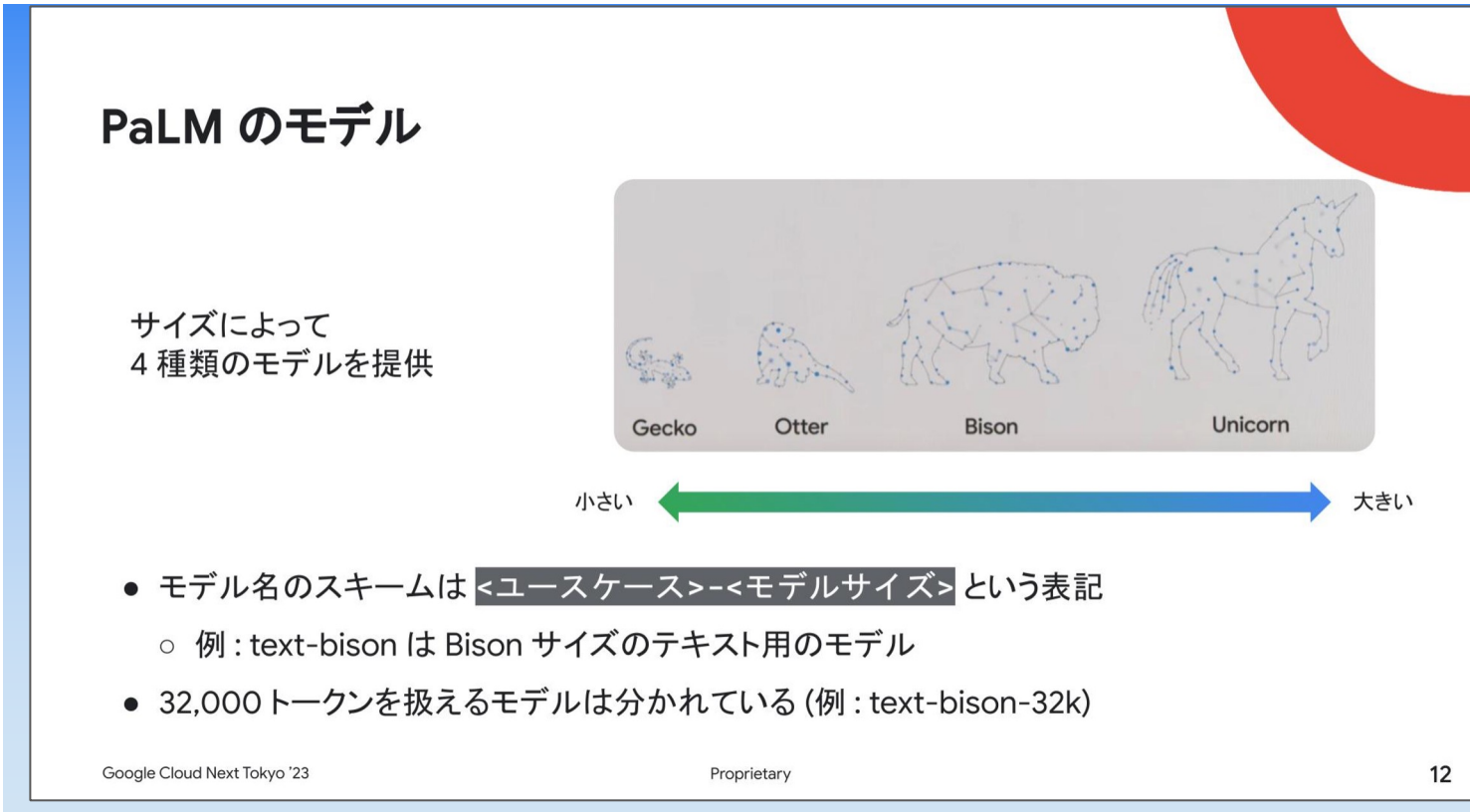
この LLM の中でも、PaLM（Pathways Language Model）は Google が開発したものです。Pathways とは、モデルをトレーニングするためのインフラ環境のことを指します。PaLM を使えば、モデルを効率的にトレーニングできるインフラ環境を Google のサービス上で構築できます。

現在の最新版は PaLM 2 であり、日本語を含む多言語を理解し、ロジックや常識に基づく回答ができるほか、より多くのプログラミング言語をサポートしています。

PaLM は新しく登場したモデルではなく、これまで Google が開発した機械学習モデルの最新バージョンという位置付けになっています。PaLM に限らず生成 AI モデルは、Google が開発した革新的な大規模言語モデル「Transformer」に大きな影響を与えています。

PaLM は、モデル名のスキームが全て「ユースケース」と「モデルサイズ」の 2 つを組み合わせて表されています。小さいモデルから順に、ゲッコー、オッター、バイソン、ユニコーンと 4 種類のモデルがあり、最も小さいモデルのゲッコーは、非常に軽量なためモバイルデバイス上で使用できるモデルとなっています。

その上のバイソンは、機能とコストのバランスに優れている位置付けになっており、Codey のコード生成機能にもバイソンが活用されています。

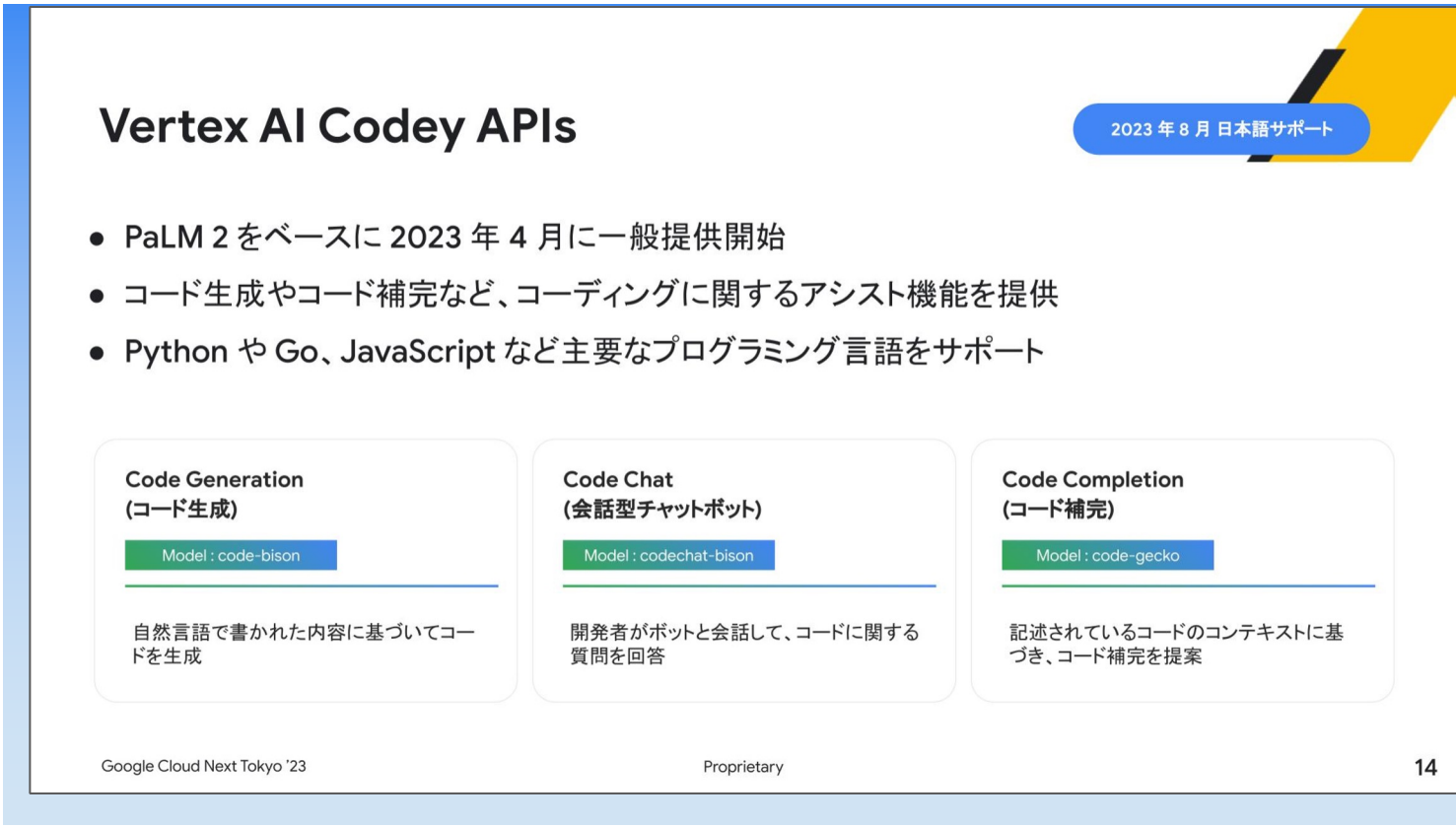


PaLM には 4 種類のモデルがある

コーディングに特化した生成 AI「Vertex AI Codey」

Vertex AI Codey は、PaLM をコーディング向けにカスタマイズしたモデルを API 経由で利用できるサービスです。PaLM 2 ではより多くのプログラミング言語をサポートしています。2023 年 8 月には、日本語のサポートも始まったので日本語での応答が可能になりました。モデルは、ユースケースごとに「コード生成」、「会話形チャットボット」、「コード補完」の 3 つを提供しています。

コード生成は、プロンプトを元にコードブロックを生成するモデルであり、会話型チャットボットはボットを通じてコーディングに関するマルチターンの会話ができるようなモデルです。コード補完は、コーディングの中の行動の内容を理解して次の数行のコードを低レイテンシーで補完するモデルです。

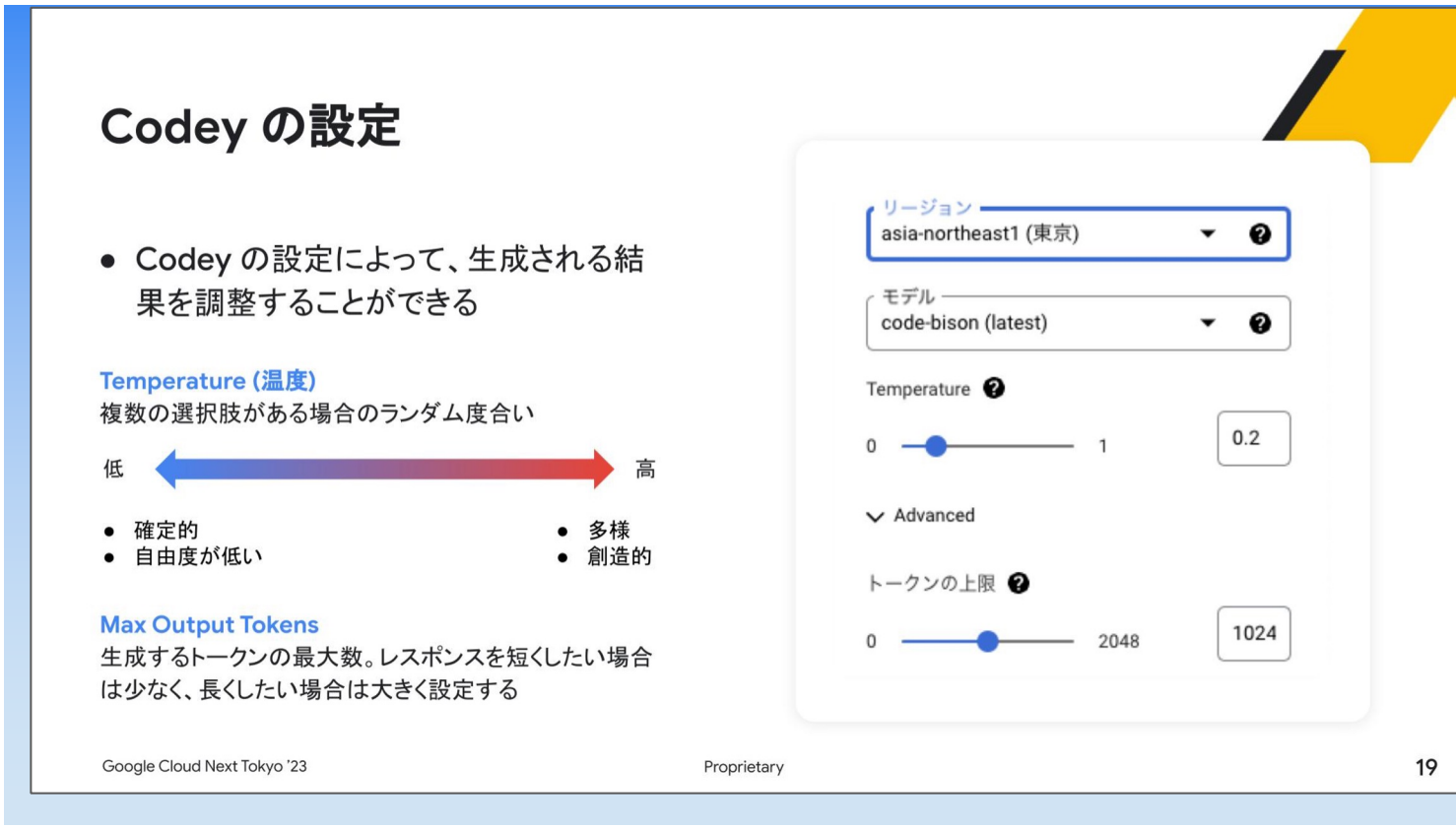


3 つの AI アシスト機能を提供する

Codeyでは、多くの主要なプログラミング言語をサポートしています。中でも特徴的なのが、GoogleSQL という BigQuery や「Cloud Spanner」で採用されている SQL の拡張言語をサポートしているところです。そのため、それらのクエリを生成することも可能です。

下図は Codey の設定画面です。最も重要な設定は「温度（Temperature）」です。生成AI の場合、1つのリクエストに対して生成されるコンテンツが複数候補に挙がる場合があります。

その際に、温度の設定によってどのような結果を返すべきかを決めて、最終的にその結果をレスポンスとして返します。温度が低い場合はより確定的な結果が出力され、温度が高い場合は多様でより創造的な出力結果になります。



回答のランダム性を左右する Temperature は特に重要な設定となる

また、出力された結果のトークンの大小を調整することも可能です。トークンの出力を抑えたい場合は、トークンの設定を見直すことで、より少ないトークンで返すといった調整が可能です。

Codey のモデルのファインチューニングも可能

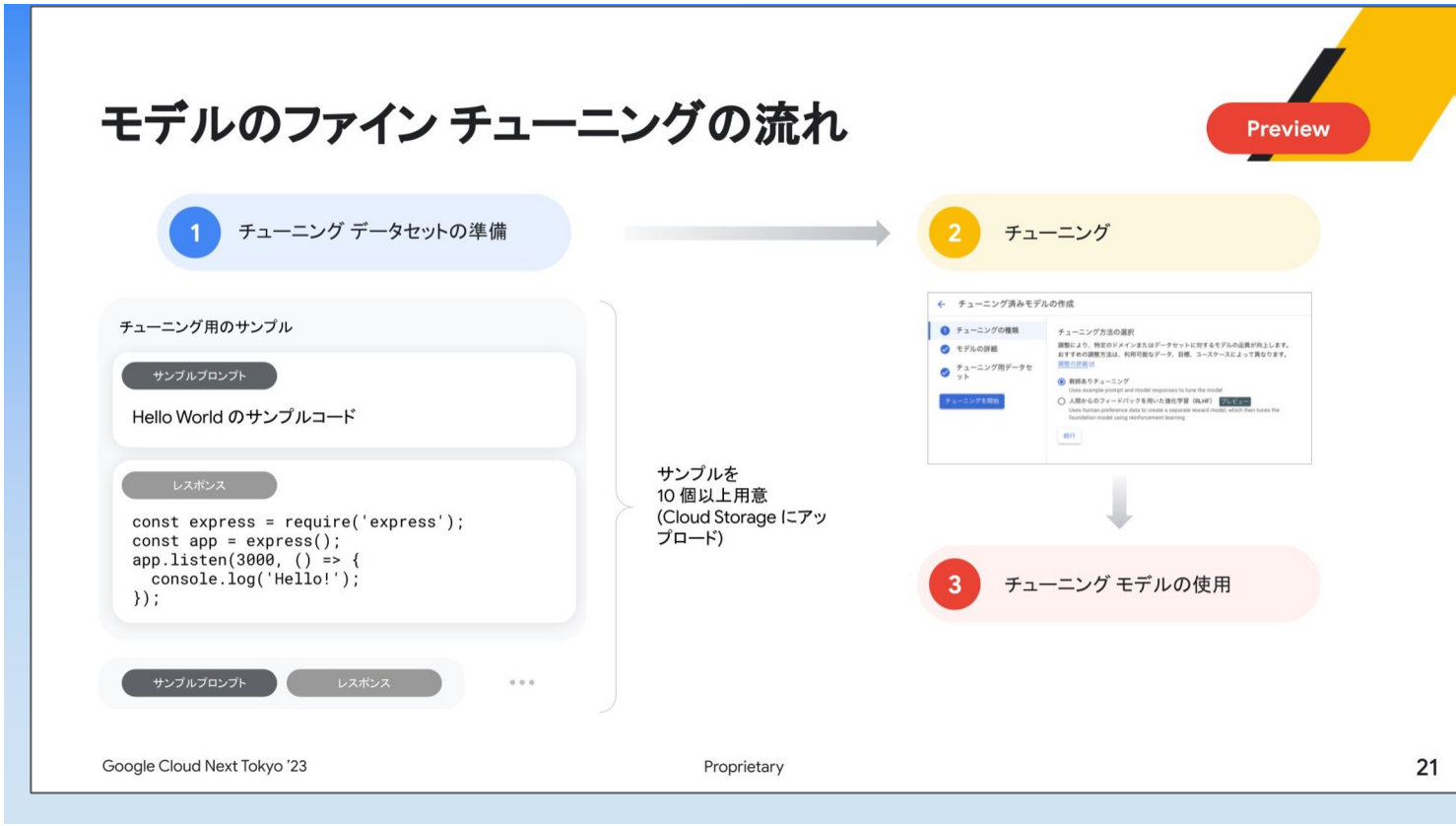
続いて、モデルのファイン チューニングについて解説します。

ファイン チューニングとは微調整を意味し、自社のデータやオリジナルのトレーニング データを学習させて、独自の要件に対応したモデルにチューニングすることを指します。

例えば、内製で開発をしていて、外部公開していない SDK やライブラリの使い方をモデルに教えてそれを踏まえた上でのコードを出力したいといった場合や、外部公開していない情報をモデルにトレーニングさせて、それを前提にコードを生成したいといった場合に、このような追加学習が必要です。

ほかには、社内のコーディング規約が決まっていて、それに沿ったコードを生成してほしいといった場合にもファイン チューニングが役立ちます。

ファイン チューニングは、実際に入力したいプロンプトと求める結果のサンプルを用意します。Codey の場合は、例えば、下図のようなプロンプトを投げ、コード生成のサンプルを数パターン用意し、それを行区切りの JSON ファイルで 1 ファイルにして「Cloud Storage」にアップロードします。プロンプトと実際のコードのサンプルの用意が必要です。



Codey でもモデルのファイン チューニングが可能

IDE に統合された形で利用できる Duet AI in Google Cloud

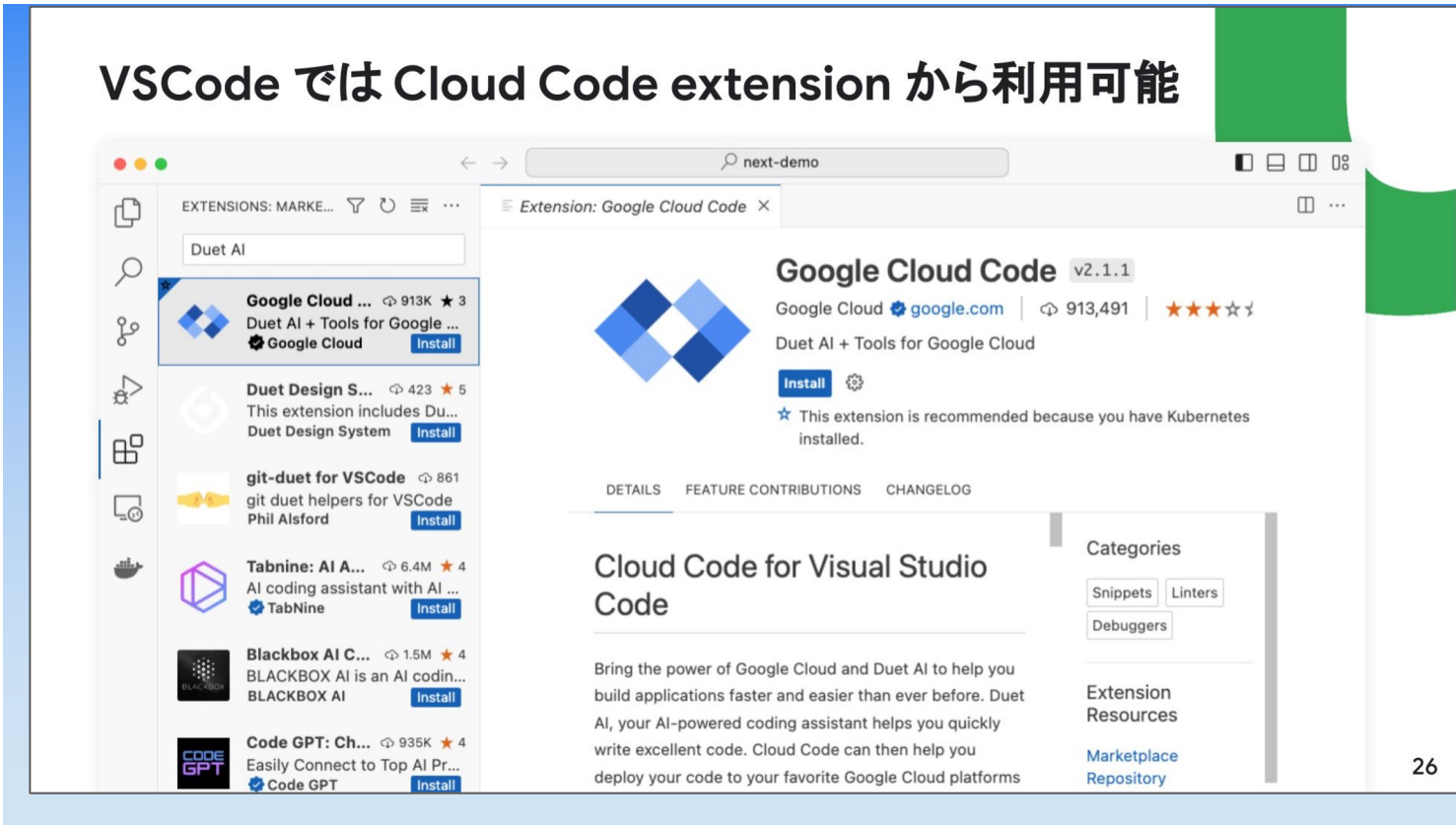
続いて、Duet AI in Google Cloudについて説明します。Duet AI は、Google Cloudの AI アシスト機能の総称です。BigQuery のエディターの機能や Cloud Logging の中にも機能として付属しており、サービスごとの追加機能として AI アシストを利用できます。

この AI アシスト機能は、ここではコーディング支援にフォーカスしていますが、ほかにも運用支援やデータ活用支援、ノーコード開発の支援、セキュリティの支援など多岐に渡っています。

Duet AI はさまざまな支援が可能

Duet AI のコーディングの支援に関しては、IDE に統合された形で利用することができます。VSCode のエクステンションとして利用でき、IntelliJ IDEA などの JetBrains 系の各種 IDE もサポートしているため、普段の開発環境に導入しやすいでしょう。

簡単な導入方法としては、VSCode のエクステンションから利用することをおすすめします。通常の VSCode のエクステンションの利用の仕方と同様に、「Google Cloud Code」のエクステンションをインストールし、Codey を利用する Google Cloud のプロジェクトを指定するだけで環境設定は完了です。



VSCodeからの利用がおすすめだ

プロンプト デザインの工夫が必要

次にプロンプト デザインについて説明します。プロンプト デザインとは、LLM に指示をするテキスト、つまりプロンプトの設計の仕方です。生成 AI を使う上では、ラフに質問して回答を得る場合に、質問の仕方によっては期待した応答を返してくれないことがよくあります。

これは、「こういう状況だからこういうものを出してほしい」という質問のコンテキストがモデルに正しく伝わっていないためです。このような場合、最初からプロンプトをどのようにデザインすればよいのかを理解しておくと、求めている回答を一発で得ることができたり、求めている回答をより効率的に得られたりできます。

プロンプト デザインの違いで LLM 活用の効果は変わる

注意点として、LLM にはトークンの制限事項があります。LLM では、自然言語の文章をトークン化した上で扱うというプロセスが存在します。一度に扱えるトークンの数に上限があり、PaLM 2 では 3 万 2000 トークンという制限があります。PaLM 2 の課金形態は文字数による従量課金をとっているので、トークンの数が少なくなるプロンプトをデザインすると、効率的に活用できます。

プロンプトに含める 4 つの要素

一般的にプロンプトは、大きく分けて以下の「命令・コンテキスト・入力データ・出力形式」の 4 つの要素を含めることが望ましいとされます。

(1) 命令

モデルに実行してほしいタスク、例えば「Hello World のコードを生成してください」といった内容です。

(2) コンテキスト

タスクに関する追加の情報や追加の文脈を書きます。例えば、「あなたは JavaScript のエンジニアです」というように表現します。先ほどの命令では「Hello World のコードを生成してください」としか書かれていないので、何の言語で書けば良いのかの指示がありません。ここで、コンテキストに「あなたは JavaScript のエンジニアです」と書くことで JavaScript のエンジニアとして回答してくれます。

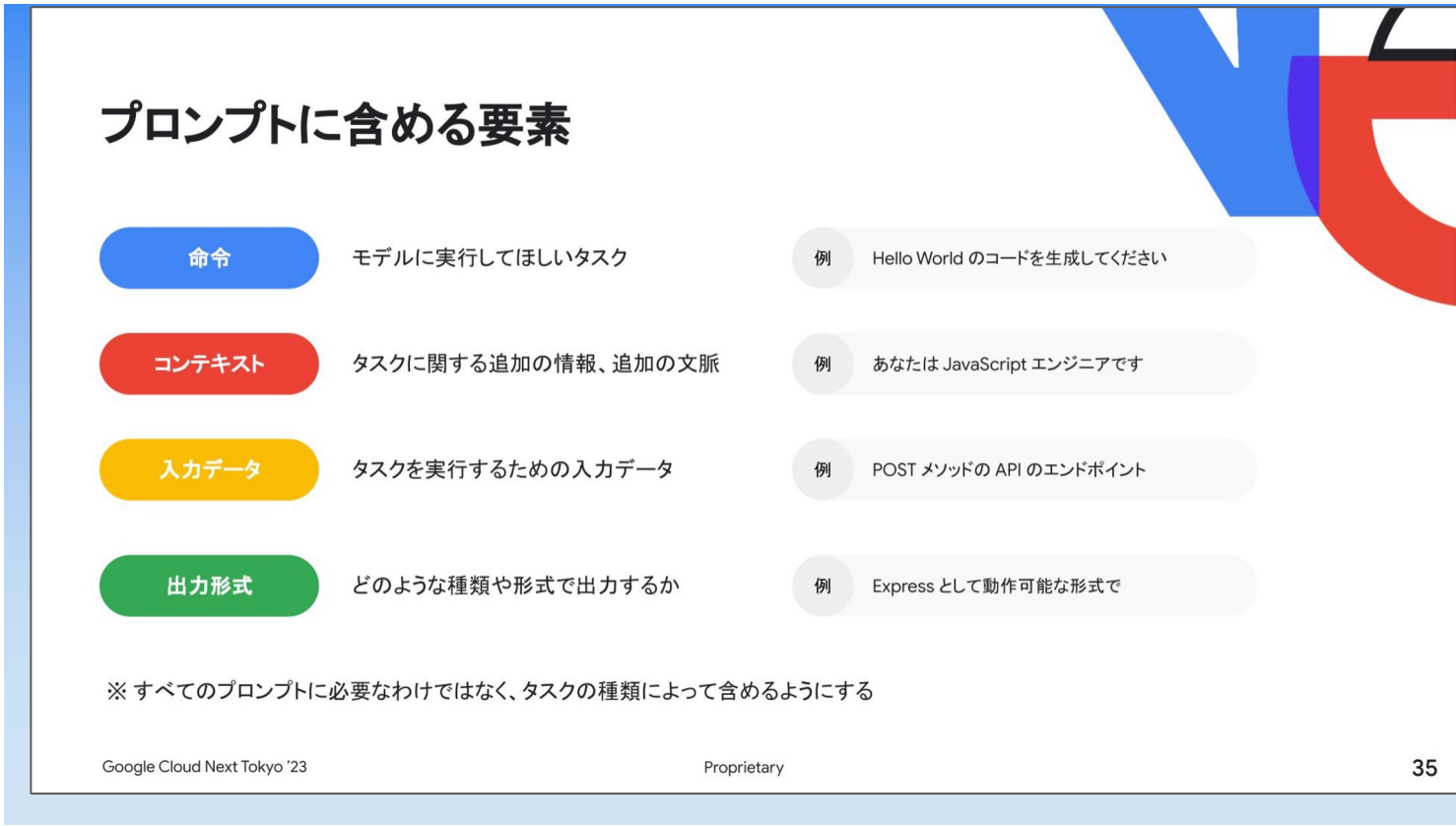
(3) 入力データ

タスクを実行するための入力データとして、例えば「POST メソッドの API エンドポイントを出力してください」といった指示をします。

(4) 出力形式

どのような種類や形式で出力するのか、例えば「Express として動作可能な形式で出力してください」といったものです。

このような形で命令・コンテキスト・入力データ・出力形式が適切にプロンプトに含まれているのかを確認しながらリクエストすると、より効率的に生成 AI を活用していけるはずです。



プロンプトに含める要素

適切なプロンプトで利用するという観点からも、Duet AI in Google Cloud は非常に有用です。Duet AI in Google Cloud では、Cloud Code のコンソール上の機能として AI にリクエストする形を取ります。そのため、本来プロンプト デザインとしてイチから全部書かなければならないようなことも、ユーザーが実際に見ている GUI から取れるコンテキストを一部理解した上で結果を返してくれます。

開発への生成 AI 活用は今後も進化する

Duet AI in Google Cloud は、プログラムに申請していただくと、そのプロジェクトの中で有効化することができるような形を取っています。皆さんが普段お使いの Google Cloud プロジェクトで有効化し、積極的にご利用いただいてフィードバックをいただければと思います。

手っ取り早く使い始めるにはDuet AI in Google Cloudが有効ですが、より自社の状況に合わせて使うには、Codeyを使ってインテグレーションする必要がありますので、こちらもぜひ発展形として検討していただければと思います。

開発プロセスへの生成 AI 活用はまだ発展途上であり、Google Cloud としてもまだこれからさらなる進化を遂げていく予定です。多くの方に使っていただきつつ、どのような形でビジネス価値を生み出せるかを、引き続きお客様と一緒に考えていきたいと思います。

参照リンク

1. Google Cloud のジェネレーティブ AI の概要
2. これからはじめる Vertex AI Codey と Duet AI のアプリ開発 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