Google Cloud

実践 生成 AI

～ Vertex AI で始める Google の 大規模言語モデル PaLM の活用～

Google Cloud
AI/ML スペシャリスト
牧 允皓

Table of Contents

スピーカー紹介

Google Cloud の AI/ML スペシャリスト。

これまで構造化データの分析や統計解析、画像・自然言語の機械学習システムのビジネス実装に携わる。現在は Vertex AI 上の MLOps や 生成 AI 実装を支援。

牧 允皓
Google Cloud
AI/ML スペシャリスト

01

Google Cloud における
生成 AI

Google の 生成 AI におけるイノベーション

膨大なデータを学習させる技術、文脈理解や会話生成までをリード

2015

機械学習をより
高速に、大規模に行える
専用チップである
Google TPU を発表

2017

Transformer：
大規模なデータセットで
の学習が可能に
（Large Language
Model）

2018

BERT：
Transformerを利用した
大規模言語モデル
要約、応答、翻訳など
複数タスクを
汎用モデルで高性能を発揮

2020

会話データで学習された
LaMDA：
Google が公開したほぼす
べてのものについて会話す
ることが可能

2023

PaLM 2

PaLM 2 を搭載した
会話型 AI サービス
Bard を発表

PaLM 2 - 100 以上の言語にわたる多言語対応の言語生成モデル

Google Cloud で API として利用可能なほか、Duet AI として Google Workspace、Google Cloud に組み込みお客様の生産性を向上

- **多言語**

  100 以上の言語の学習により、
  慣用句、詩、なぞなぞなど
  ニュアンスも理解、生成

- **推論**

  数式を含む科学論文や Web を
  学習。ロジックや常識に基づく推論や数
  学のハンドリングが可能

- **コーディング**

  Python や JavaScript など

https://japan.googleblog.com/2023/05/palm-2.html

生成 AI のエンタープライズにおける活用シーン

会話、検索、クリエイティビティの 3 大ユースケースで生産性を向上

顧客接点

オンラインのやりとりを自然な会話で自動化・効率化

- 顧客サポートの自動化
- イントラのナレッジのQ&A
- ウェブサイトのナビゲーション

など

ビジネスユーザー / 分析者

複雑なデータに簡単にアクセス

- 製品・コンテンツカタログの探索
- ビジネスプロセスの自動化
- ドキュメント検索

など

クリエイティブ / エンジニア

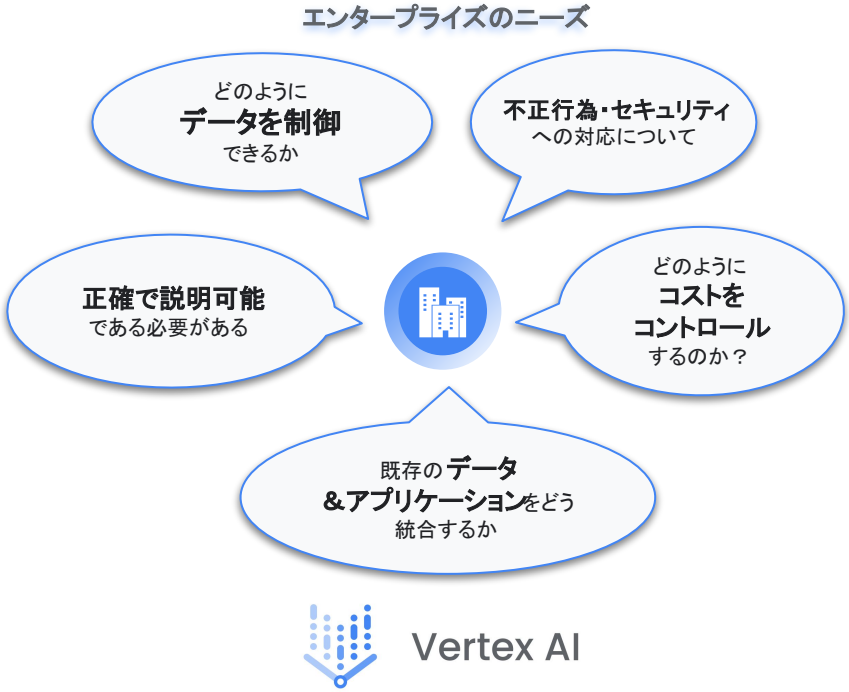
ワンクリックでコンテンツを生成、生産性を向上

- コード自動生成と提案
- チャットによるアプリ開発の半自動化
- 画像生成による壁打ち

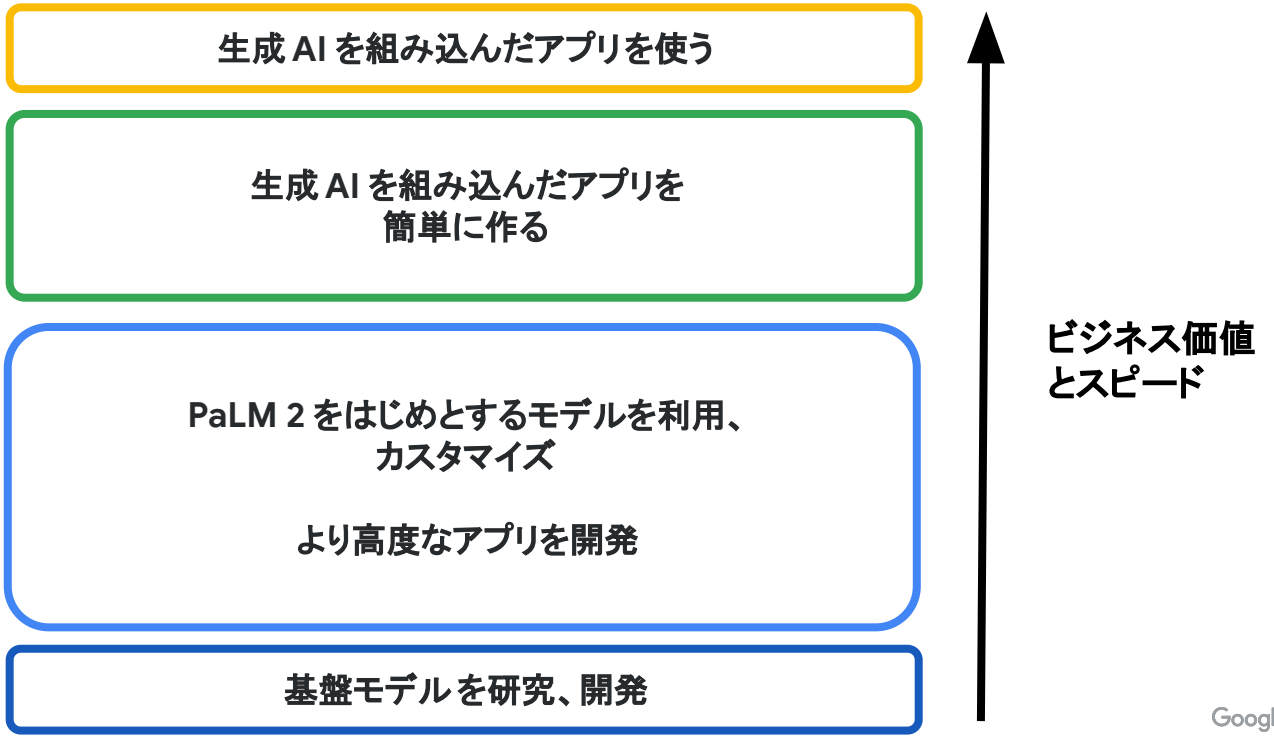
など

エンタープライズデータで活用

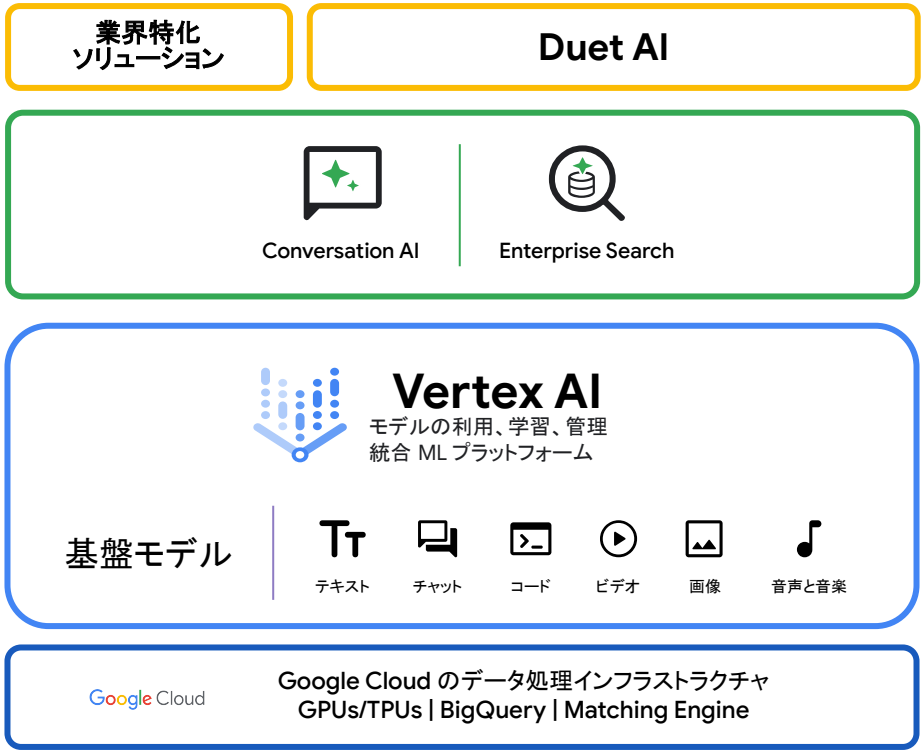
コンシューマーとエンタープライズの生成 AI に対する異なるニーズにお応え



企業における生成 AI との付き合い方選択肢



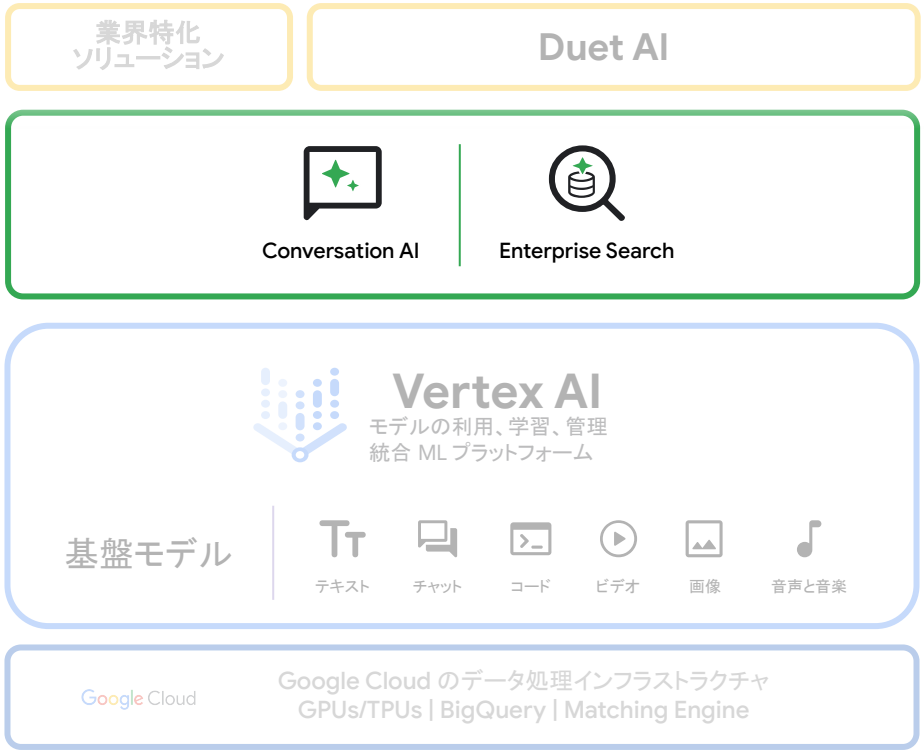
企業のあらゆるユーザーの 生産性を向上 させ
全く新しい顧客体験の変革を実現



02

生成 AI アプリを開発する際の
プロダクトの選び方

生成 AI アプリケーションを開発する



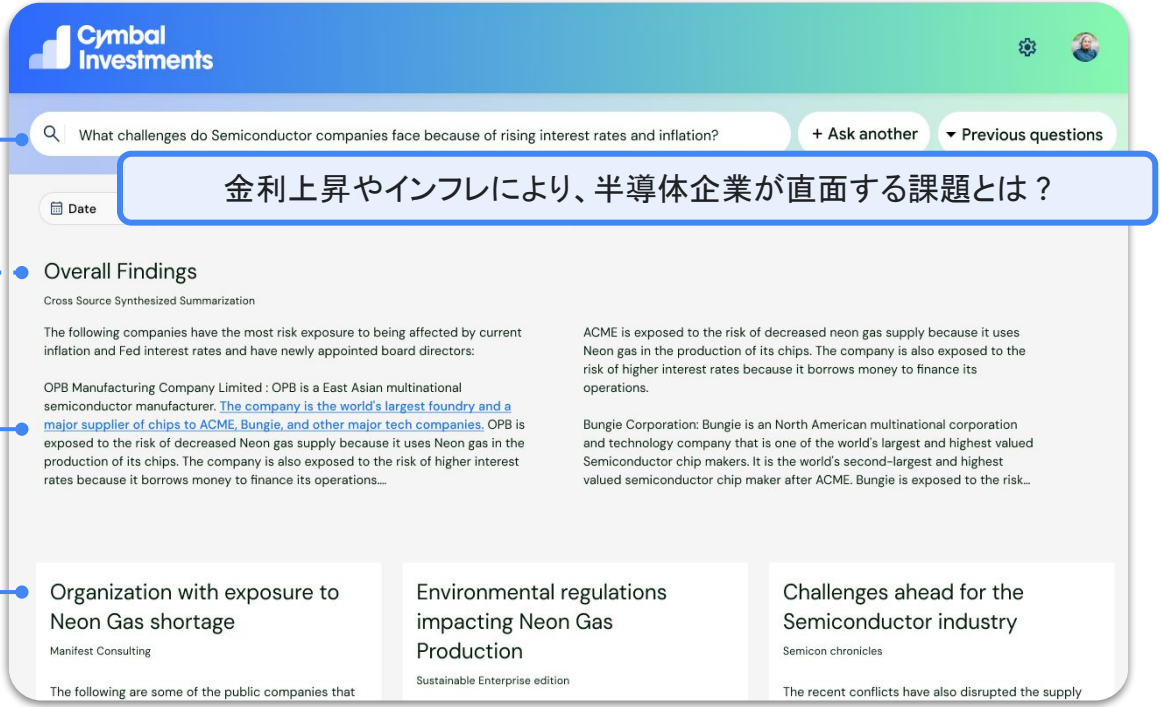
「より深く意図を理解」「より的確な情報へアクセス」するGoogle 検索を企業内に実現

会話型の検索

検索結果の情報統合・要約

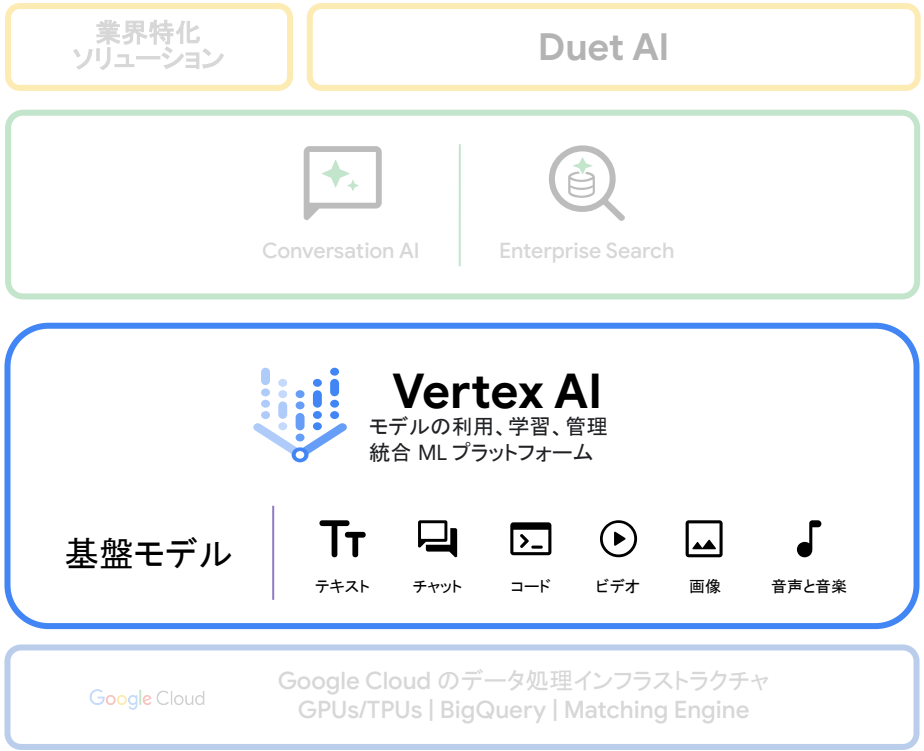
情報ソースからの引用

レコメンドされた検索結果



生成 AI アプリを実装する際に必要な構成要素

Vertex AI でサポートされる生成 AI のコア機能



DIY で生成 AI を実装するためのコア機能

LangChain などの OSS 技術によって、自然言語インターフェースのアプリケーション実装がより身近に。Vertex AI に実装されたコア機能を組み合わせることでより自由な設計が可能。

コア機能を組み合わせて
高度な生成 AI のアプリを作る際の注意点

ハルシネーションの問題

LLM における生成 AI は、基本的に自然言語の単語分布などを学習したモデル。**生成されたテキストが事実に即している保証はなく**、アプリケーションを設計する際には細心の注意が必要。

グラウンディング

ハルシネーションの問題の解決策として、回答する情報を信頼できるソースから参照する技術が必要。

生成より検索や抽出に近い振る舞いを指す。**検索対象のデータと検索クエリの近さを定義する必要**がある。

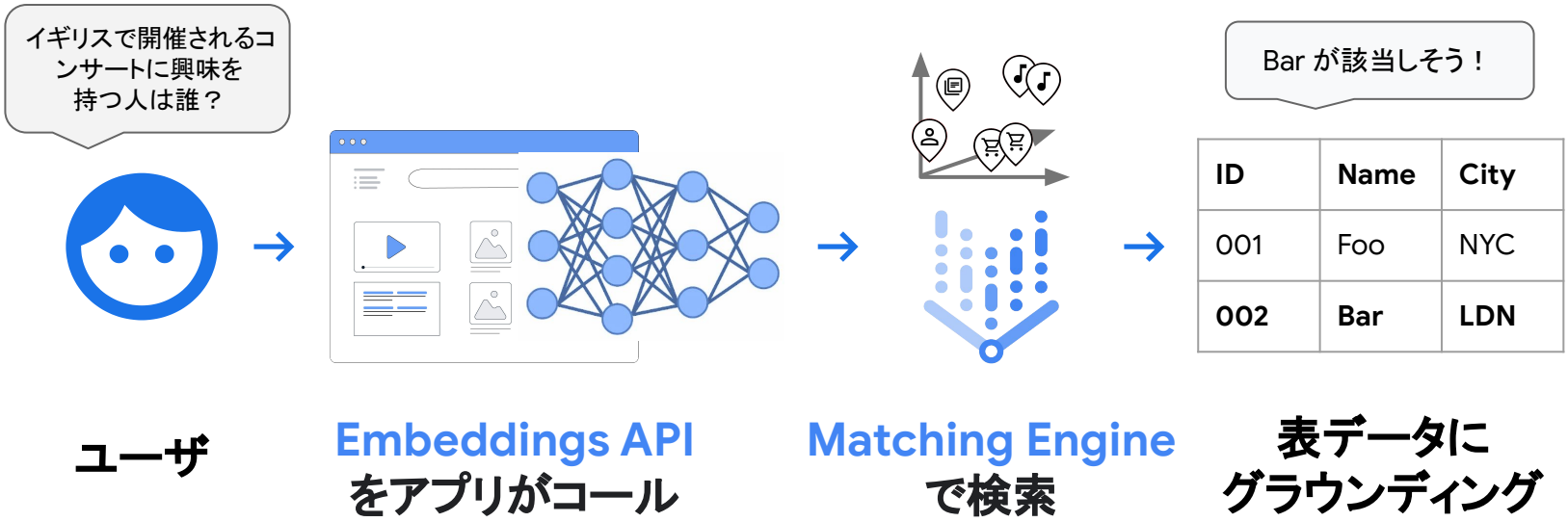
エンベディング

非構造化データ（画像、テキスト、音声など）を、数理モデルで扱えるように、数値に変換すること。データの意味を適切に表現できる変換。

高度なアプリを開発する場合は、**既存データのエンベディングとベクトル検索の実装が重要**。

LLM は汎用的なタスクを解く際に有効だが、
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
グラウンディングを始め、タスクに適したアーキテクチャが重要

エンベディングとベクトル検索を用いたグラウンディングの一例

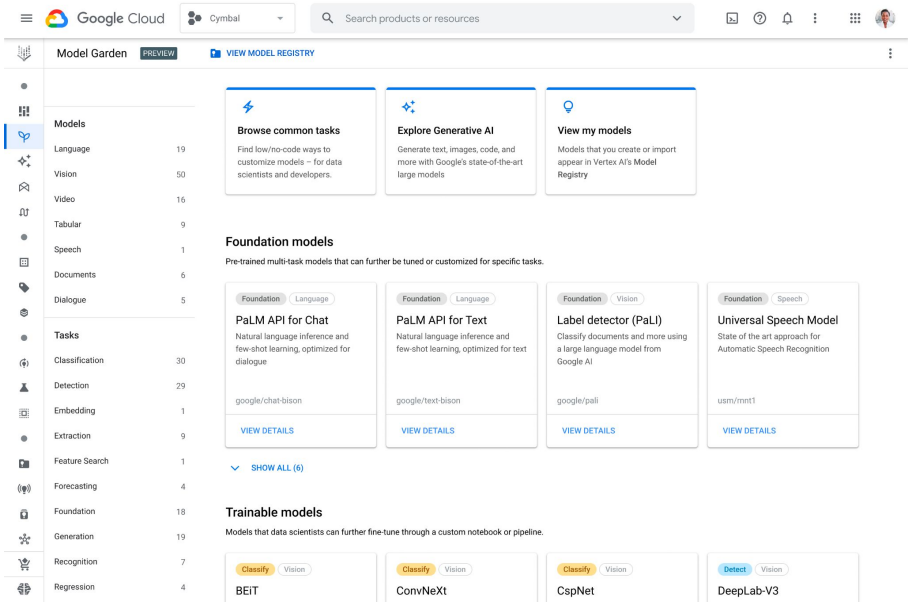


03

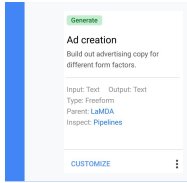
Vertex AI で始める
PaLM の活用

Vertex AI Model Garden

ML のユーザー ジャーニーを一箇所で管理

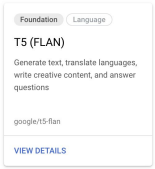


PaLM など Google の Foundation Model、タスクに特化したソリューション、オープンソースモデルを発見し、テストドライブするためのワンストップショップ

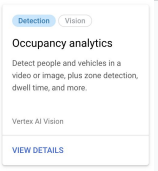


あらかじめ用意されたテンプレートで **基盤モデル を直接使用**

業界やユースケースに合わせたデータ＆プロンプトによる
モデルのチューニング



データサイエンスノートブックとVertex AI Pipelinesでオープンソース
モデルを**カスタマイズ**



言語、視覚、音声、などさまざまな
タスク特化ソリューションへの
API アクセス

Generative AI Studio

生成 AI のワークフローを実現する
統合インターフェース

シンプルでわかりやすい画面

プログラミングや AI/ML の知識がなくても利用できる直感的なインターフェース

独自のデータを使用してモデルを調整

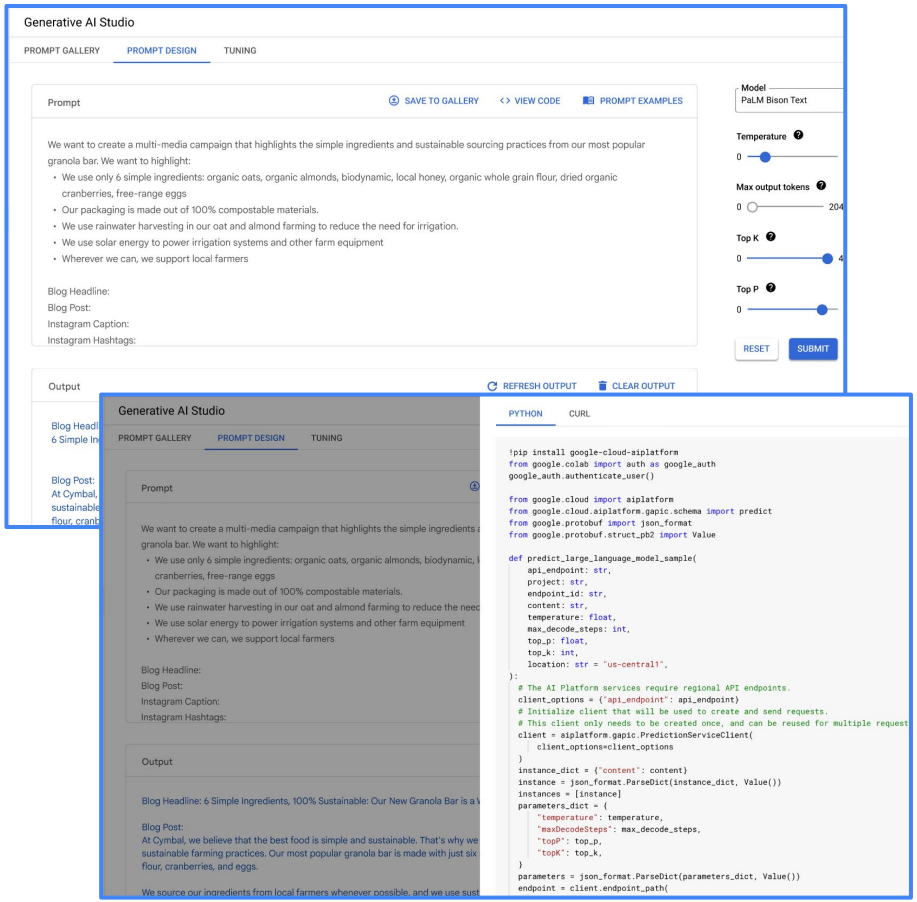
プロンプト エンジニアリング、ファインチューン、RLHF など、さまざまなチューニングに対応

本番環境で迅速にモデルを使用

API コードを迅速に生成およびカスタマイズしてアプリケーションに組み込み

複数のデータ形式に対応

テキスト、画像、コード、音声に対応



PaLM for Text and Chat

Google の大規模言語モデルをビジネスに活用

Google が開発した基盤モデルにアクセス

エンタープライズ用途への PaLM 2 モデルの適用

多様なユースケースへの対応

質問、要約、分類、アイディア作成など
さまざまな用途に活用可能

カスタム言語タスクを実言

事前に用意されたプロンプト ギャラリーで
Zero-shot / Few-shot prompting を簡単に入力

複数ターンに対応したチャット機能

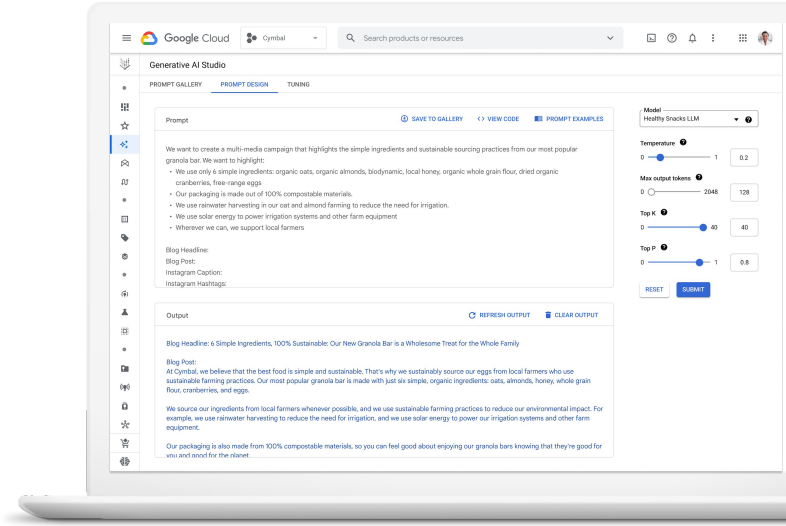
コンテキストを維持しながら長い会話も可能

モデルのカスタマイズ

自社のデータを使ったタスク固有のチューニング

文字単位での課金体系

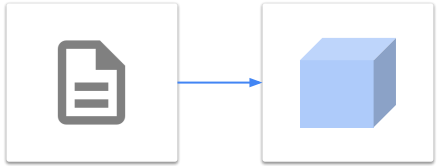
入出力に対して、文字ベースでの課金



Use Cases: センチメント分析、要約、テキスト書き換え、広告コピーの生成、会話

Vertex AI が提供するファインチューニング

Vertex AI で実行可能なプロンプトエンジニアリング



プロンプト

事前学習
言語モデル

※プロンプトを通じて与えられる
データには限界がある

Vertex AI が提供するファインチューニング

企業のデータ

事前学習
言語モデル

事前学習言語モデルをタスクに特
化してチューニング

※独自データを用いたモデルのカスタマイズをコスト効率に優れた方法で提供

- プロンプトエンジニアリングだけでは実現が困難であったタスクを独自データを用いたモデルのカスタマイズで実現
- コストとカスタマイズ性のバランスの取れたチューニング
方法を提供

Private Preview

Reinforcement learning from human feedback (RLHF)

人間のフィードバックを利用してモデルの有用性を高める ー 人間中心の生成AI へのアプローチ

ニーズに合わせてモデルのパフォーマンスを向上

人間の「どう感じたか」というフィードバックを生成 AI に反映することで、より人間にとって自然な対応へと改善を行う手法（RLHF）をサポート

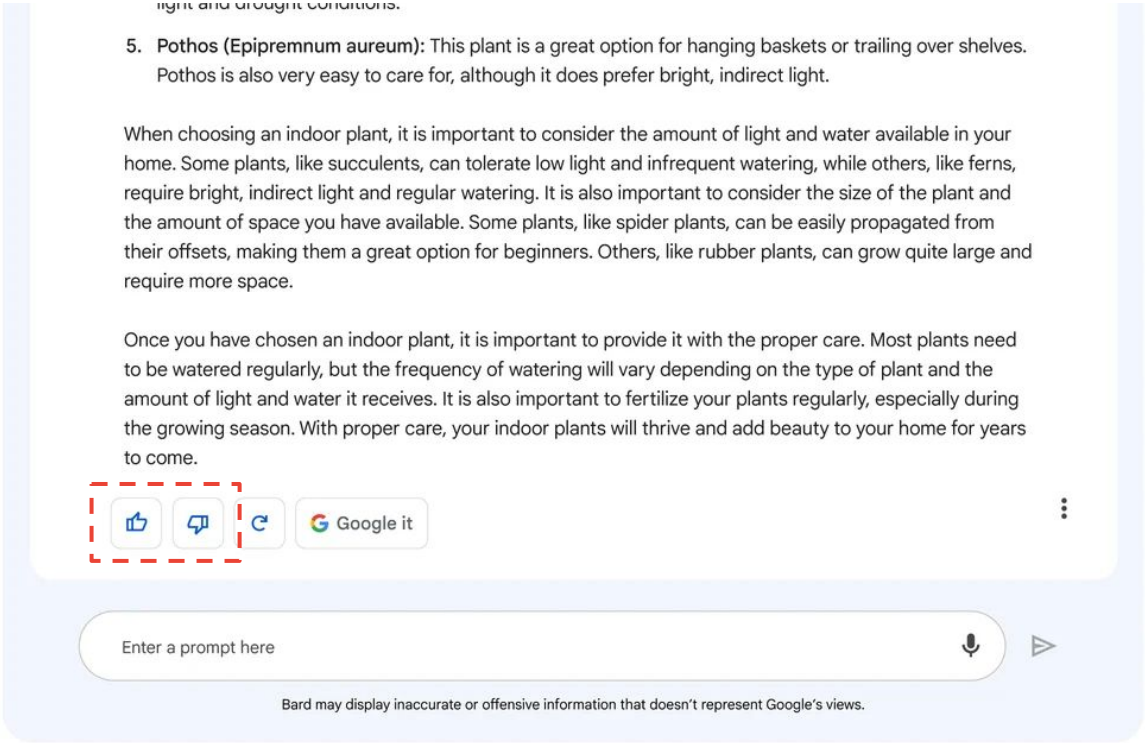
モデルのサイズを増やすことなく、事前トレーニングされた LLM を改善

RLHF 調整モデルは通常サイズを大きくすることなく、他のトレーニング方法と比較してより正確で信頼性が高くすることが可能

顧客満足度の重要な業界で効果を発揮

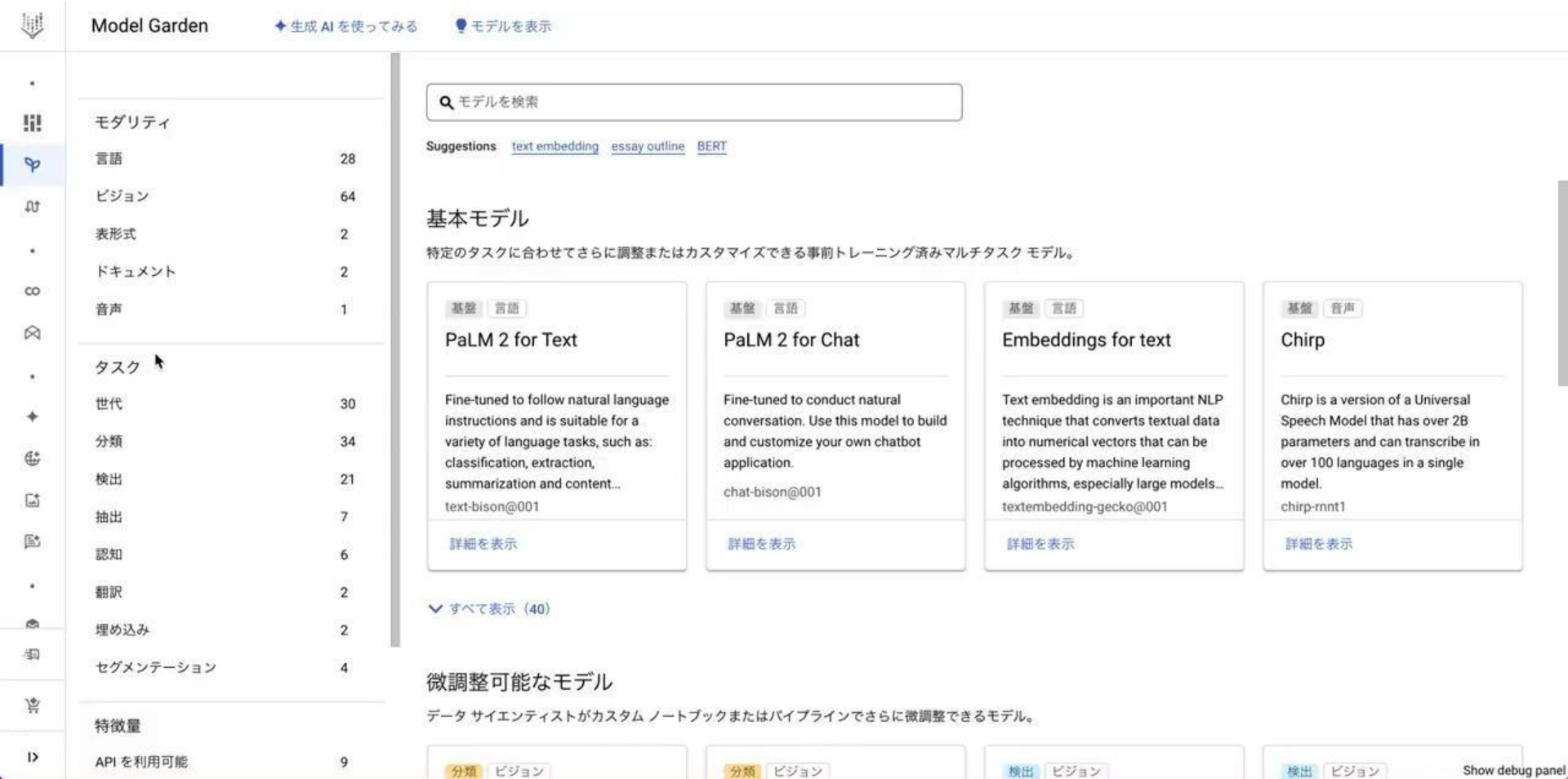
医療、金融、EC など。
RLHF により、ユーザーのニーズをよりよく理解して対応できるようにモデルをトレーニングでき、最終的には顧客満足度やエンゲージメントの向上につながります。

人間のフィードバックのイメージ

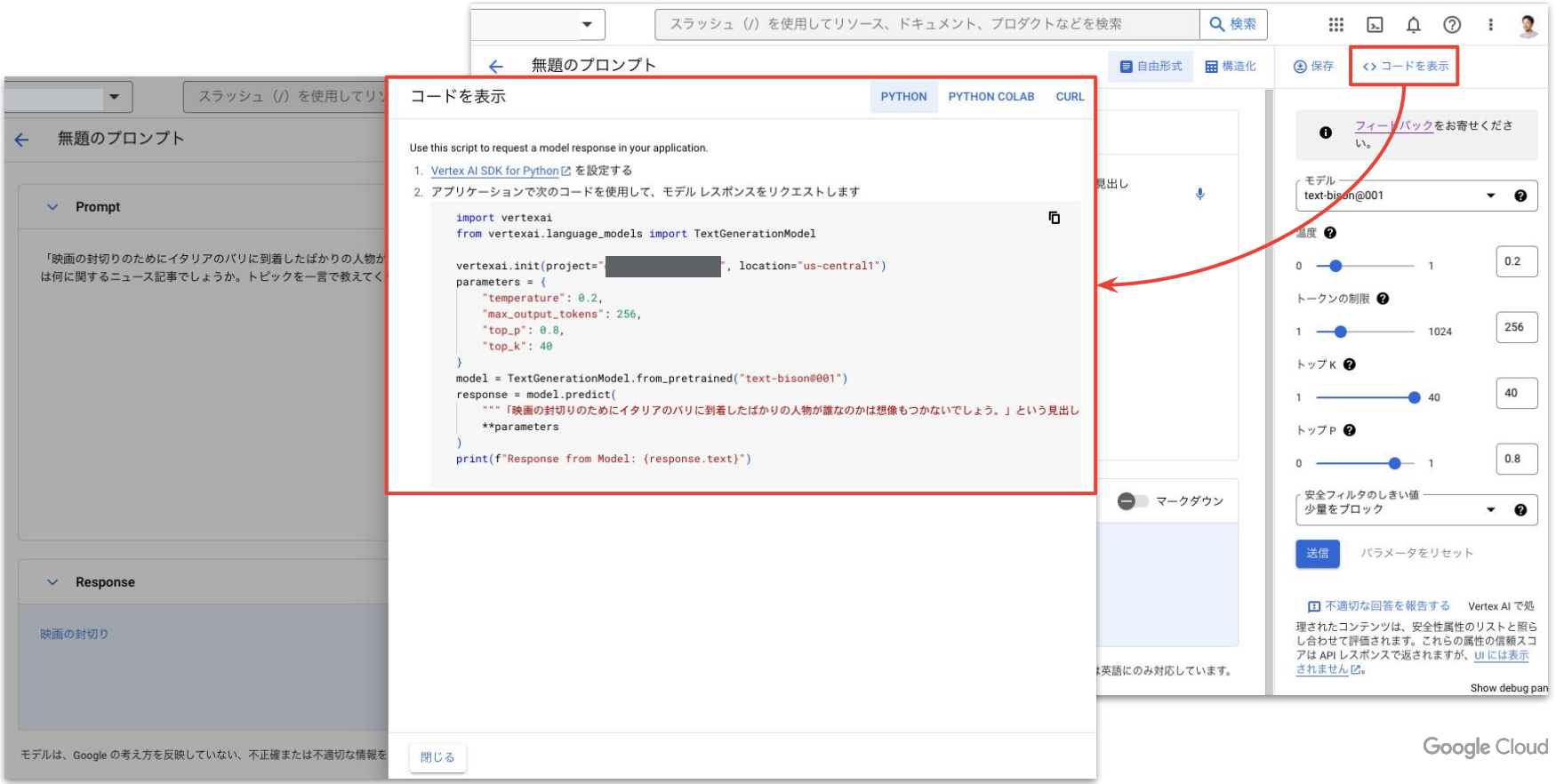


参照: [Try Bard and share your feedback]

Generative AI Studio デモ



Generative AI Studio コード取得



まとめ

Google Cloud が提供する生成 AI のプロダクトカテゴリ

Vertex AI でサポートされる生成 AI

生成 AI の API、基盤モデルの提供から調整までのエンド ツー エンドのサポート

- 複数データに対する Generative AI API
- Model Garden と Generative AI Studio
- プロンプトエンジニアリング
- ファインチューニング
- 基盤モデルに対する ML Ops

Generative AI App Builder

エンタープライズ要件を満たした、基盤モデルによるチャットと検索による新しいユーザー体験を提供するアプリケーション開発を加速

Enterprise Search

- Google 品質ですぐに利用可能な検索
- チューニング可能
- マルチモーダル
- 正確、事実、新鮮さ

Conversation AI

- 単一、複数ターン
- 情報検索
- トランザクション
- フローの制御

Thank you.

Google Cloud