

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

Google Cloud における生成 AI	01
-------------------------------	-----------

生成 AI アプリを開発する際のプロダクトの選び方	02
----------------------------------	-----------

Vertex AI で始める PaLM の活用	03
--------------------------------	-----------

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

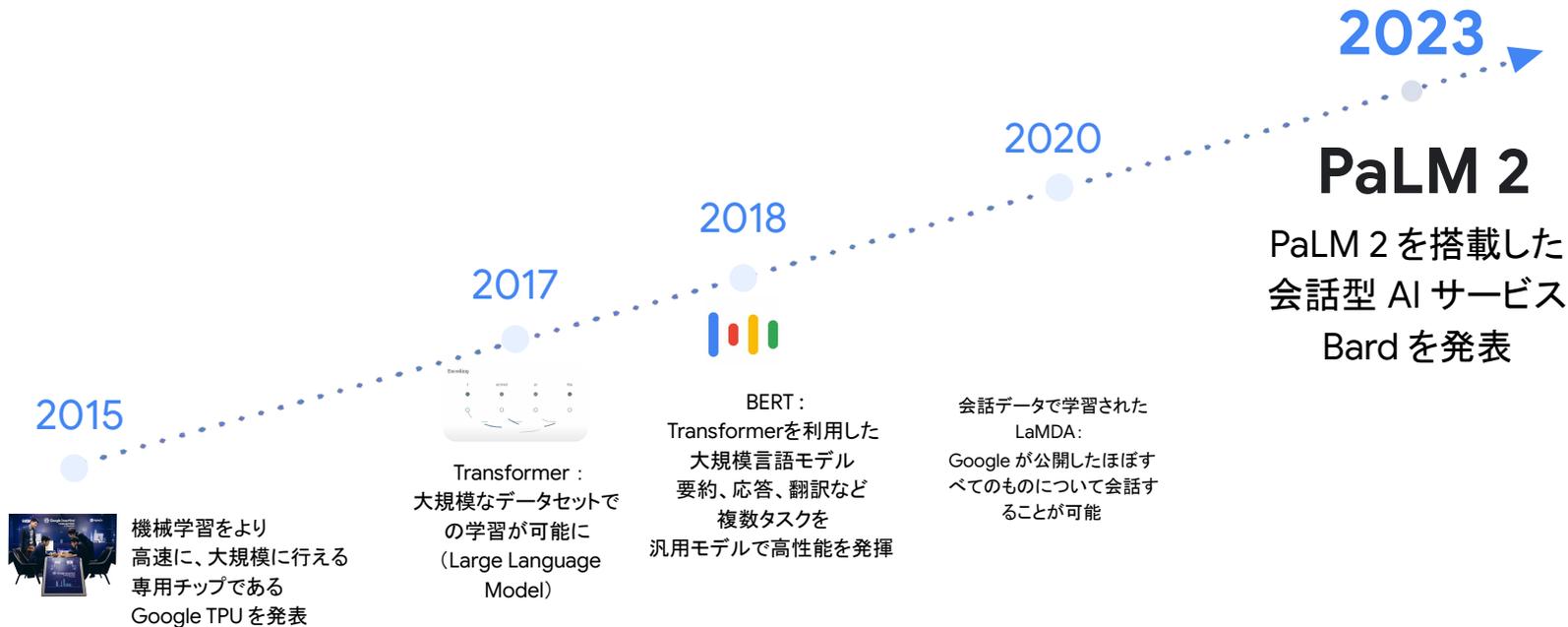
Google Cloud
AI/ML スペシャリスト

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



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

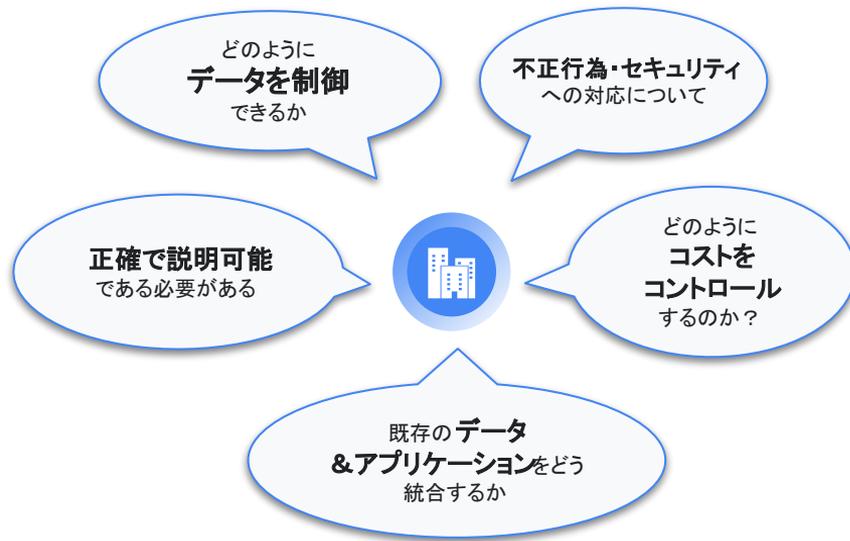
コンシューマーとエンタープライズの 生成 AI に対する異なるニーズにお応え

コンシューマーのニーズ



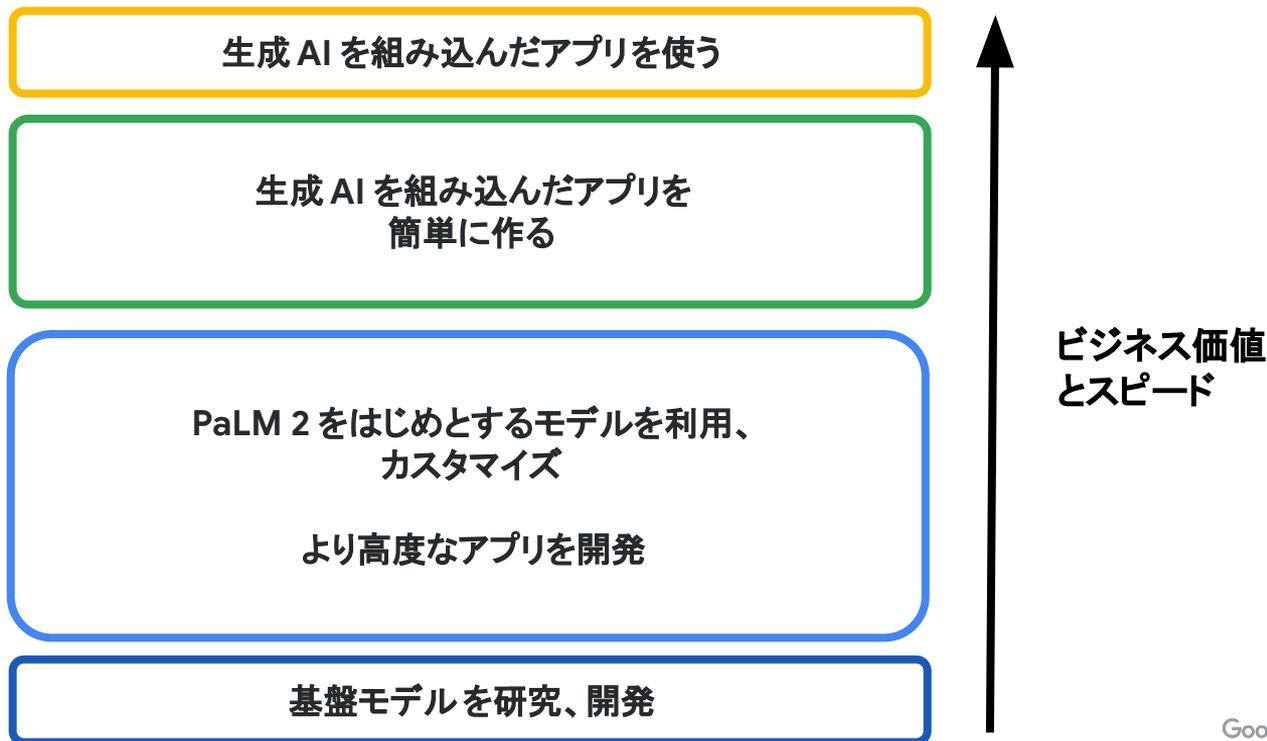
Bard

エンタープライズのニーズ

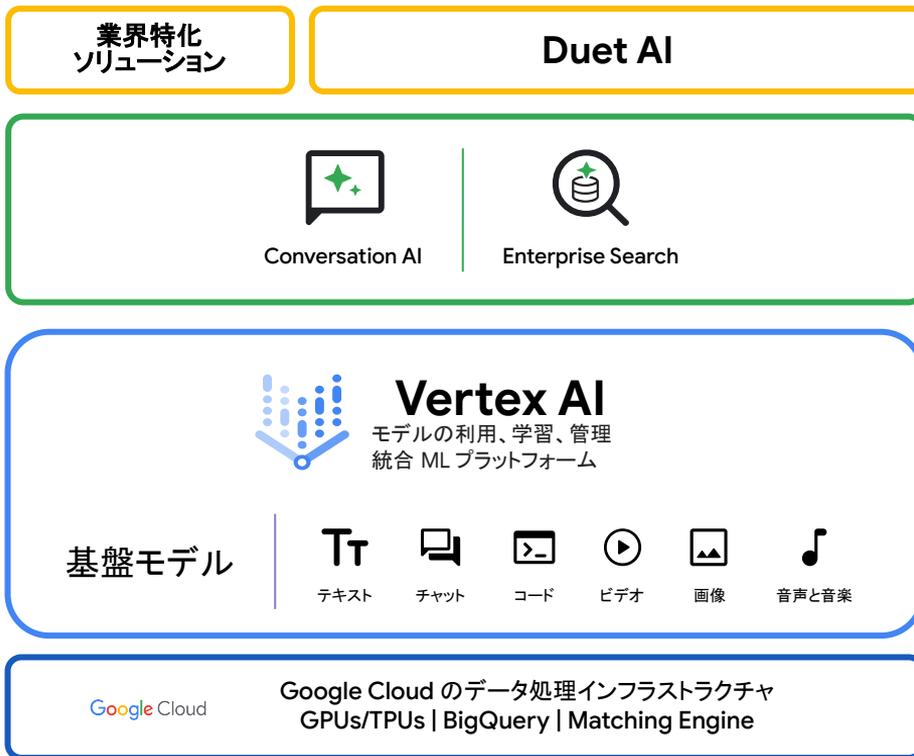


Vertex AI

企業における生成 AI との付き合い方選択肢



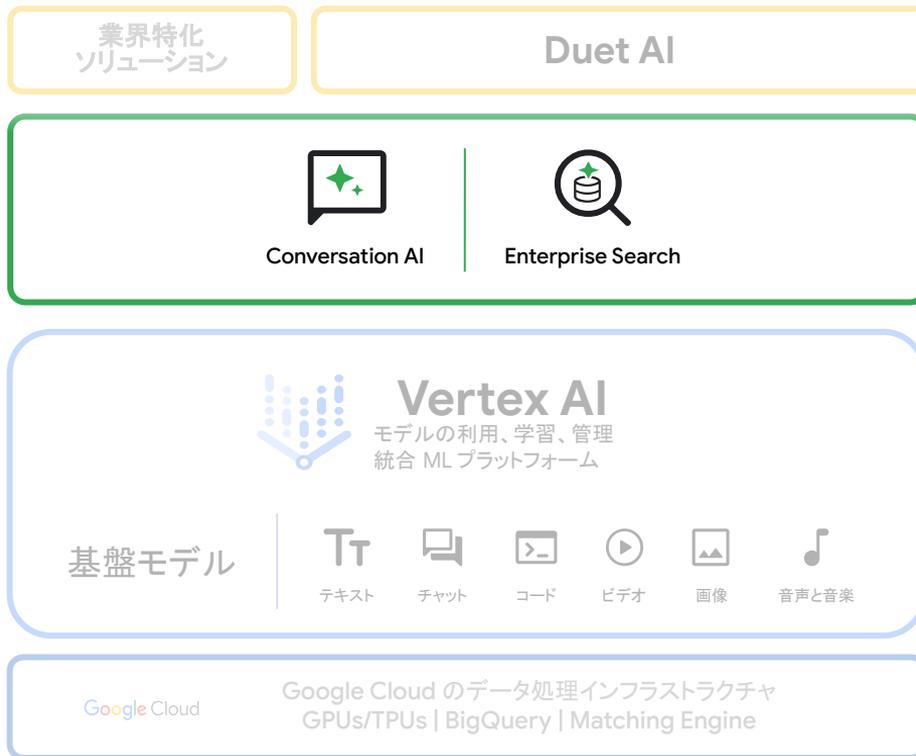
企業のあらゆるユーザーの生産性を向上させ 全く新しい顧客体験の変革を実現



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」「よりの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

Cymbal Investments

What challenges do Semiconductor companies face because of rising interest rates and inflation? + Ask another Previous questions

金利上昇やインフレにより、半導体企業が直面する課題とは？

Overall Findings
Cross Source Synthesized Summarization

The following companies have the most risk exposure to being affected by current inflation and Fed interest rates and have newly appointed board directors:

OPB Manufacturing Company Limited : OPB is a East Asian multinational semiconductor manufacturer. [The company is the world's largest foundry and a major supplier of chips to ACME, Bungie, and other major tech companies.](#) OPB is exposed to the risk of decreased Neon gas supply because it uses Neon gas in the production of its chips. The company is also exposed to the risk of higher interest rates because it borrows money to finance its operations...

ACME is exposed to the risk of decreased neon gas supply because it uses Neon gas in the production of its chips. The company is also exposed to the risk of higher interest rates because it borrows money to finance its operations.

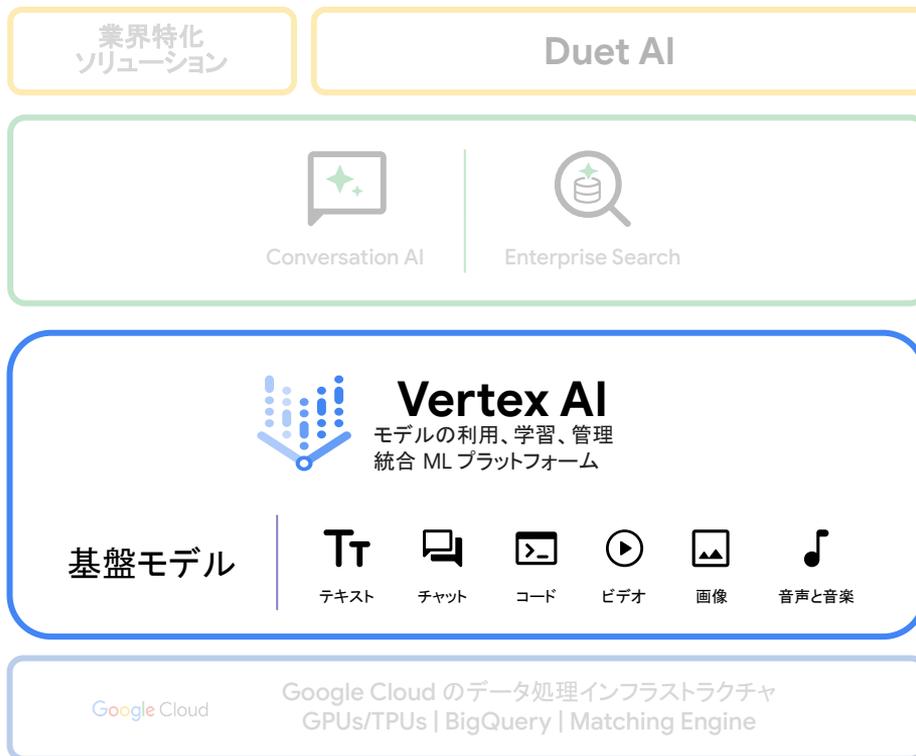
Bungie Corporation: Bungie is an North American multinational corporation and technology company that is one of the world's largest and highest valued Semiconductor chip makers. It is the world's second-largest and highest valued semiconductor chip maker after ACME. Bungie is exposed to the risk...

Organization with exposure to Neon Gas shortage Manifest Consulting The following are some of the public companies that	Environmental regulations impacting Neon Gas Production Sustainable Enterprise edition	Challenges ahead for the Semiconductor industry Semicon chronicles The recent conflicts have also disrupted the supply
--	--	---

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。**

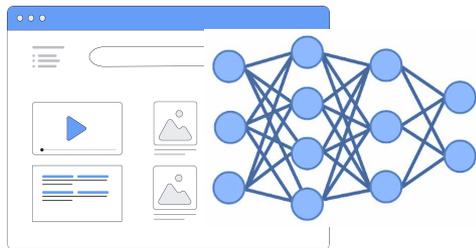
LLM は汎用的なタスクを解く際に有効だが、個別のタスクを解く場合や、事実に基づいた応答を返したい場合グラウンディングを始め、タスクに適したアーキテクチャが重要

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

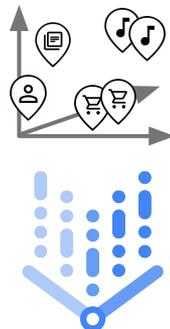
イギリスで開催されるコンサートに興味を持つ人は誰？



ユーザ



Embeddings API
をアプリがコール



Matching Engine
で検索



Bar が該当しそう！

ID	Name	City
001	Foo	NYC
002	Bar	LDN

表データに
グラウンディング

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

The screenshot shows the Vertex AI Model Garden interface. On the left is a navigation menu with categories: Models (Language, Vision, Video, Tabular, Speech, Documents, Dialogue), Tasks (Classification, Detection, Embedding, Extraction, Feature Search, Forecasting, Foundation, Generation), and Tractable models (Recognition, Regression). The main content area is divided into sections: 'Browse common tasks' (Find low/no-code ways to customize models), 'Explore Generative AI' (Generate text, images, code, and more with Google's state-of-the-art large models), 'View my models' (Models that you create or import appear in Vertex AI's Model Registry), 'Foundation models' (Pre-trained multi-task models that can further be tuned or customized for specific tasks), and 'Trainable models' (Models that data scientists can further fine-tune through a custom notebook or pipeline). Each section contains cards for various models with descriptions and 'VIEW DETAILS' links.

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

The vertical stack of panels describes different model capabilities:

- Ad creation** (blue header): Build out advertising copy for different form factors. Includes a small screenshot of the interface.
- PaLM API for Text** (yellow header): Natural language inference and few-shot learning, optimized for text. Includes a small screenshot of the interface.
- T5 (FLAN)** (green header): Generates text, translate languages, write creative content, and answer questions. Includes a small screenshot of the interface.
- Occupancy analytics** (grey header): Detect people and vehicles in a video or image, plus zone detection, dwell time, and more. Includes a small screenshot of the interface.

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

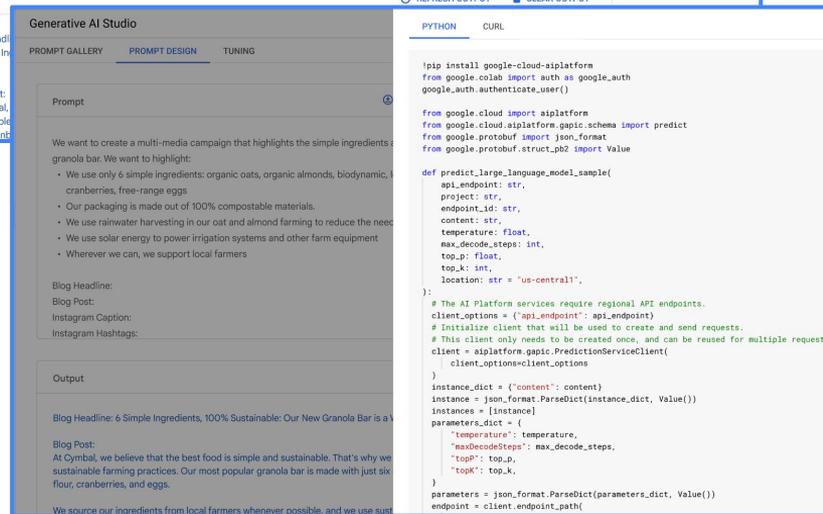
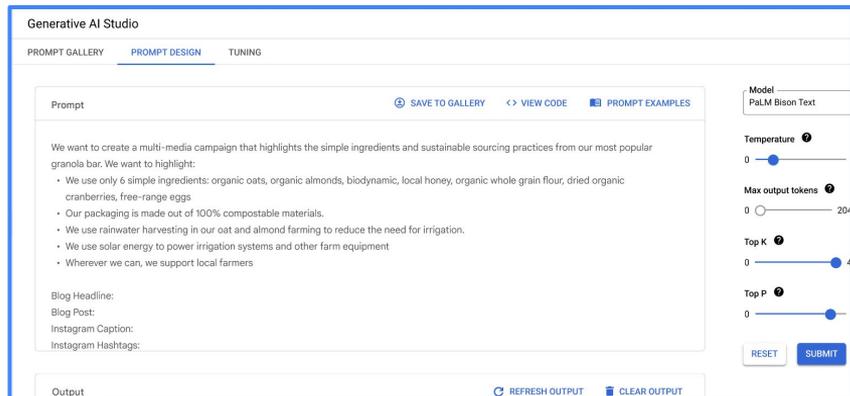
プロンプト エンジニアリング、ファインチューン、RLHF など、さま
ざまなチューニングに対応

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行

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

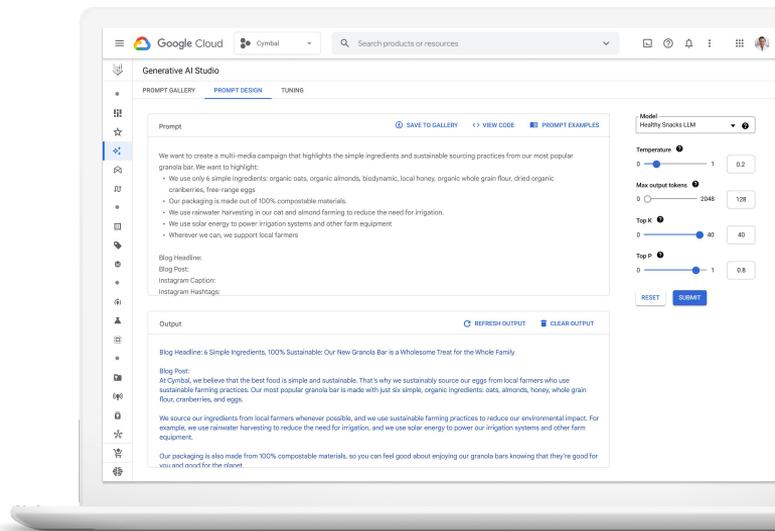
コンテキストを維持しながら長い会話も可能

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



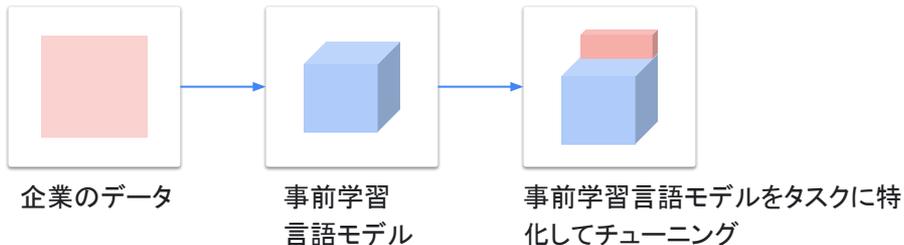
Use Cases: センチメント分析、要約、テキスト書き換え、広告コピーの生成、会話

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



Vertex AI が提供するファインチューニング



※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

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 — 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

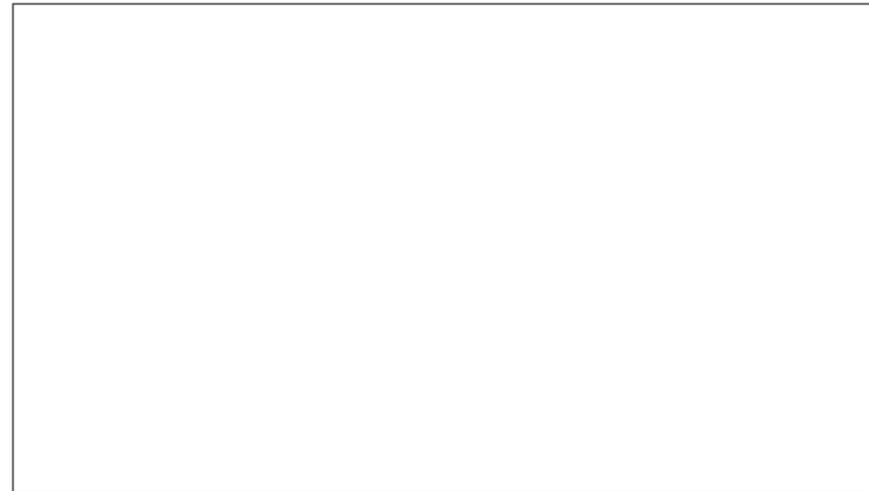
モデルのサイズを増やすことなく、事前トレーニングされたLLMを改善

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なることが可能

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

right and drought conditions.

5. **Pothos (Epipremnum aureum):** This plant is a great option for hanging baskets or trailing over shelves. Pothos is also very easy to care for, although it does prefer bright, indirect light.

When choosing an indoor plant, it is important to consider the amount of light and water available in your home. Some plants, like succulents, can tolerate low light and infrequent watering, while others, like ferns, require bright, indirect light and regular watering. It is also important to consider the size of the plant and the amount of space you have available. Some plants, like spider plants, can be easily propagated from their offsets, making them a great option for beginners. Others, like rubber plants, can grow quite large and require more space.

Once you have chosen an indoor plant, it is important to provide it with the proper care. Most plants need to be watered regularly, but the frequency of watering will vary depending on the type of plant and the amount of light and water it receives. It is also important to fertilize your plants regularly, especially during the growing season. With proper care, your indoor plants will thrive and add beauty to your home for years to come.



Enter a prompt here



Bard may display inaccurate or offensive information that doesn't represent Google's views.

Generative AI Studio デモ

Model Garden

生成 AI を使ってみる

モデルを表示

モダリティ

言語 28

ビジョン 64

表形式 2

ドキュメント 2

音声 1

タスク

世代 30

分類 34

検出 21

抽出 7

認知 6

翻訳 2

埋め込み 2

セグメンテーション 4

特徴量

API を利用可能 9

モデルを検索

Suggestions [text embedding](#) [essay outline](#) [BERT](#)

基本モデル

特定のタスクに合わせてさらに調整またはカスタマイズできる事前トレーニング済みマルチタスク モデル。

基盤 言語

PaLM 2 for Text

Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: classification, extraction, summarization and content...

text-bison@001

[詳細を表示](#)

基盤 言語

PaLM 2 for Chat

Fine-tuned to conduct natural conversation. Use this model to build and customize your own chatbot application.

chat-bison@001

[詳細を表示](#)

基盤 言語

Embeddings for text

Text embedding is an important NLP technique that converts textual data into numerical vectors that can be processed by machine learning algorithms, especially large models...

textembedding-gecko@001

[詳細を表示](#)

基盤 音声

Chirp

Chirp is a version of a Universal Speech Model that has over 2B parameters and can transcribe in over 100 languages in a single model.

chirp-rnnt1

[詳細を表示](#)

すべて表示 (40)

微調整可能なモデル

データ サイエンティストがカスタム ノートブックまたはパイプラインでさらに微調整できるモデル。

分類 ビジョン

分類 ビジョン

検出 ビジョン

検出 ビジョン

Show debug pane

Generative AI Studio コード取得

The screenshot displays the Generative AI Studio interface. A red box highlights the 'コードを表示' (Show Code) button in the top right corner of the prompt area. A red arrow points from this button to the code editor window, which is also titled 'コードを表示'. The code editor shows Python code for using the Vertex AI SDK to request a model response. The code includes imports for 'vertexai' and 'TextGenerationModel', initialization of the Vertex AI client, and a call to the 'predict' method with a prompt and parameters. The output of the model is printed to the console.

```
import vertexai
from vertexai.language_models import TextGenerationModel

vertexai.init(project="██████████", location="us-central1")
parameters = {
    "temperature": 0.2,
    "max_output_tokens": 256,
    "top_p": 0.8,
    "top_k": 40
}

model = TextGenerationModel.from_pretrained("text-bison@001")
response = model.predict(
    """「映画の封切りのためにイタリアのバリエーションに到着したばかりの人物が誰なのかは想像もつかないでしょう。」という見出し
    **parameters
)

print(f"Response from Model: {response.text}")
```

The interface also shows a 'Response' section with the text '映画の封切り' (Movie Opening). The settings panel on the right includes a 'フィードバックをお寄せください' (Provide feedback) button, a model selector set to 'text-bison@001', and various sliders for temperature (0.2), max output tokens (256), top-k (40), and top-p (0.8). A '送信' (Send) button is also visible.

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.