Google Cloud

実践 生成 AI

～ Vertex AI で始める Google の大規模言語モデルの活用～

Google Cloud
カスタマー エンジニア
遠山 雄二

スピーカー紹介

Google Cloud カスタマー エンジニア

SIer で大規模基幹系システムの開発案件に従事した後、2019 年より現職。Google では業種 / 技術問わず、フルスタックで案件を支援。

最近は生成 AI のエンタープライズ利用について、多くのお客様の　課題解決に勤めている。主な書籍は「エンタープライズのための　Google Cloud クラウドを活用したシステムの構築と運用」。

遠山 雄二
Google Cloud
カスタマー エンジニア

Table of Contents

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

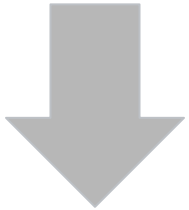
会話データで学習された
LaMDA：
Google が公開したほぼ
すべてのものについて会
話することが可能

2023

PaLM 2 & Gemini

2 月 実験的プレイグラウンド、
Bard を発表
5月 PaLM 2 を発表
12 月 Gemini を発表
Bard (英語版のみ) で利用可能に
Vertex AI を通じて本日より提供開始

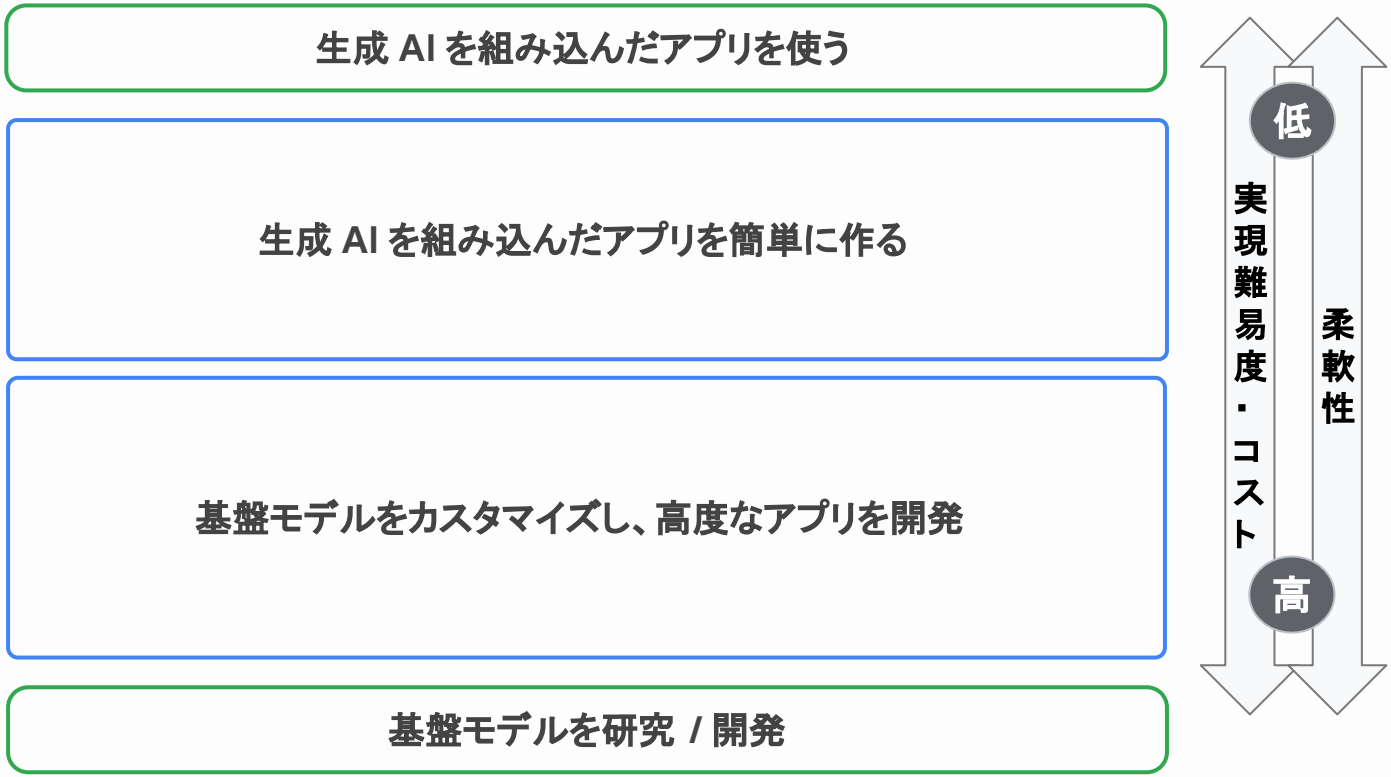
生成 AI を触ってみる



生成 AI を活用した

ソリューションを作る

企業における生成 AI の利用パターン



Vertex AI - Google Cloud の AI プラットフォーム

Google AI Studio

- 無料の Web ベースの個人開発者向けツール
- API キーを利用し、プロンプトを開発
- ユーザーはいつでも Vertex AI へ移行可能

Vertex AI

- Google Cloud のエンタープライズレディな AI プラットフォーム
- Gemini をはじめとするモデルをより高度なカスタマイズ（チューニング）、データ保護、セキュリティ、他 Google Cloud サービスとの統合を完備
- 生成 AI アプリケーションを開発するための、より拡張された機能（モデル グラウンディング、Extentions 等）
- **お客様のデータを Google Cloud が利用することは決してありません**

Vertex AI - Google Cloud の AI プラットフォーム

Duet AI

AI パワード なアプリケーションを開発

AI ソリューション
Contact Center AI | Document AI | Risk AI | ...

Search & Conversation
典型的生成AI アプリケーションである検索、
対話型アプリケーションをより素早く開発

AI Platform
生成 AI アプリケーションを実現する技術要素をトータルで提供
Extensions | Connectors | Grounding | Prompt | Serve | Tune | Distill | Eval | MLOps

Model Garden
Google のモデルや、OSS、Partner モデルをすぐに利用可能
(Gemini, PaLM 2, Imagen, LLama 等)

AI Hyper Computer | Cloud TPU/GPU | BigQuery による AI レイクハウス

ビジネスユーザー

開発者

AI 実践者

**本セッションの
フォーカス**

02

生成 AI ソリューションの開発

実用的な生成 AI のソリューションを作る際の注意点

ハルシネーションの問題

LLM 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 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、アプリケーションを設計する際には細心の注意が必要。

目的に応じて LLM のカスタマイズが必要

LLM の代表的なカスタマイズ手法

1

プロンプト デザイン

入力プロンプトを工夫することで、モデルの再学習をさせることなく、LLM に期待する振る舞いをさせる手法

2

ファイン チューニング

プロンプトと出力のデータセットを用意することで、LLM のパラメータを効率的に更新する手法

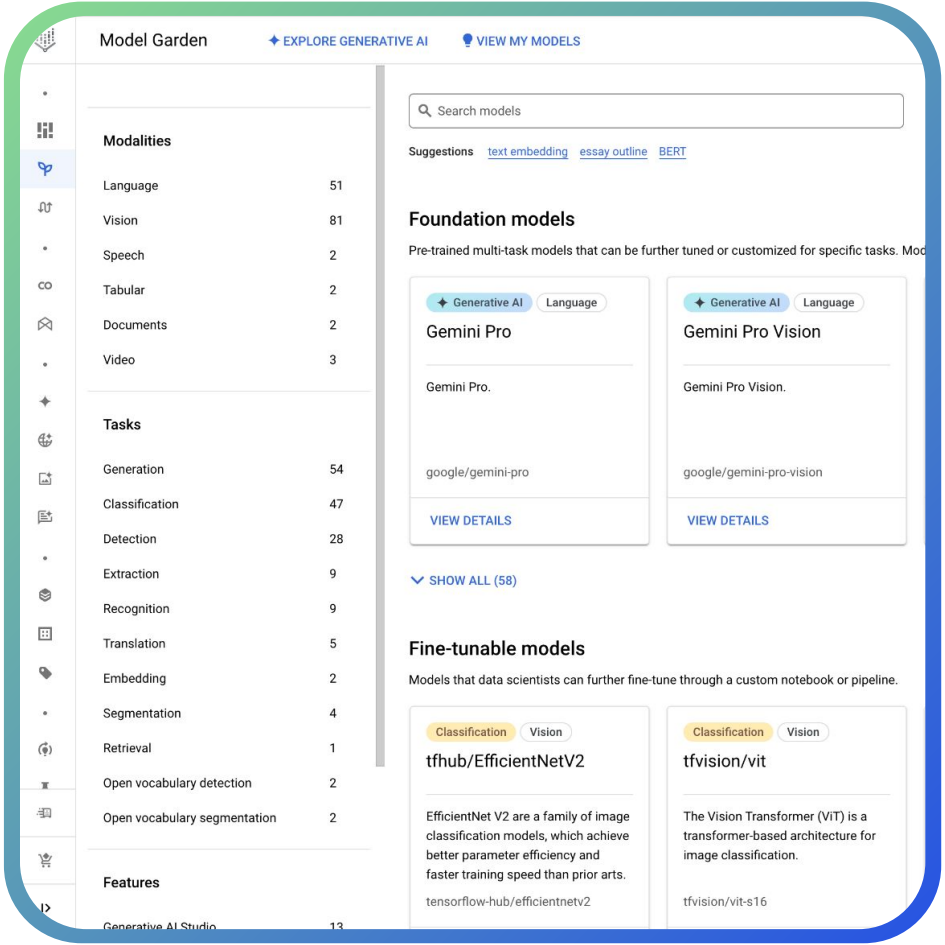
3

グラウン ディング

回答する情報を信頼できるソースから参照する手法。

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

Gemini Pro が本日より
Vertex AI にて利用可能に



Vertex AI 上の AI モデル

Model Garden にて、 Google 製モデル、 OSS モデル、パートナーモデルを
お客様の用途に合わせすぐにご利用可能

新たな Google 製モデル

Gemini Pro を追加 （テキスト、動画、画像からテキストを生成）プレビュー

PaLM Unicorn の追加

Google 製モデルのアップデート

PaLM がよりコストパフォーマンス向上（バージョンアップとともに価格の値下げ）、Imagen 2、MedLM

新たな OSS モデル

Mistral, ImageBind, DITO を追加

Vertex 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

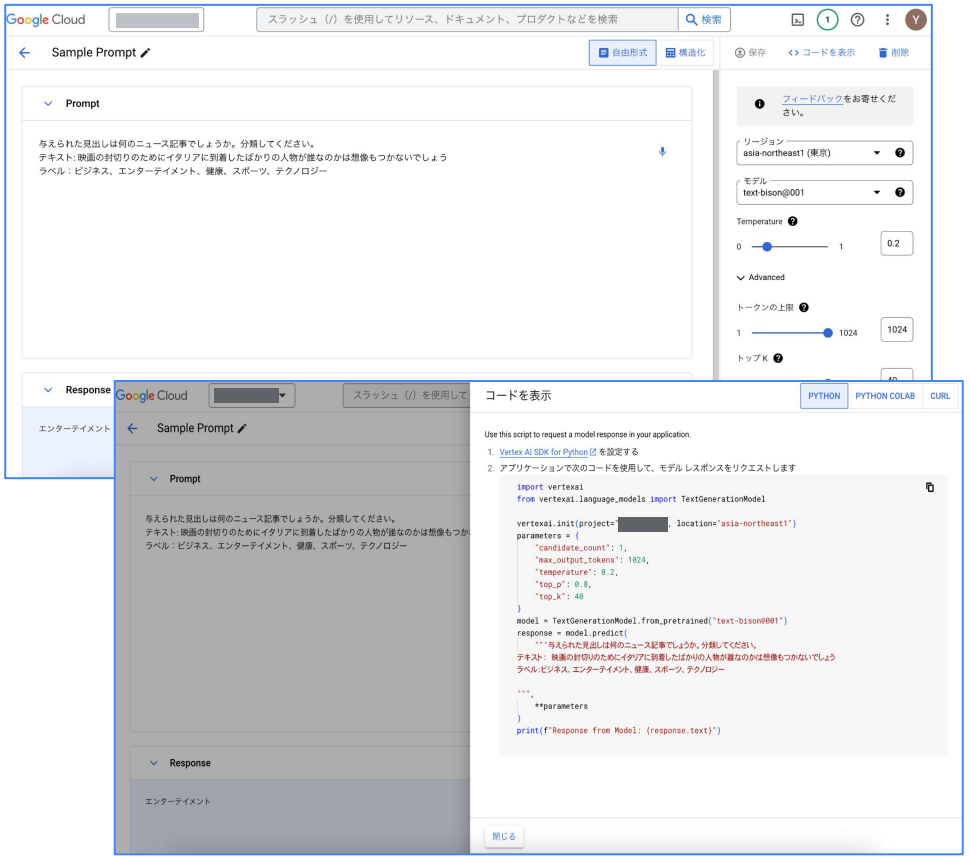
プロンプト デザイン、ファインチューン、RLHF
など、さまざまなチューニングに対応

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



LLM の代表的なカスタマイズ手法

1

プロンプト デザイン

入力プロンプトを工夫することで、モデルの再学習をさせることなく、LLM に期待する振る舞いをさせる手法

2

ファイン チューニング
(Supervised Tuning)

プロンプトと出力のデータセットを用意することで、LLM のパラメータを効率的に更新する手法

3

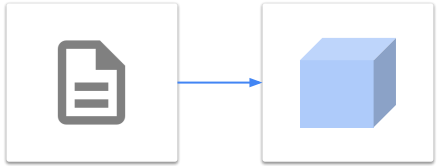
グラウン ディング

回答する情報を信頼できるソースから参照する主要。

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

プロンプト デザイン vs ファイン チューニング

プロンプト デザイン



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

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 デザイン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

基盤モデルに提供される、モデルをタスクに適合させる手法

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
LLM の出力を少量のデータでカスタマイズできる機能
Supervised Tuning が Text-bison, Chat-bison, Codey, Text embeddings で利用可能

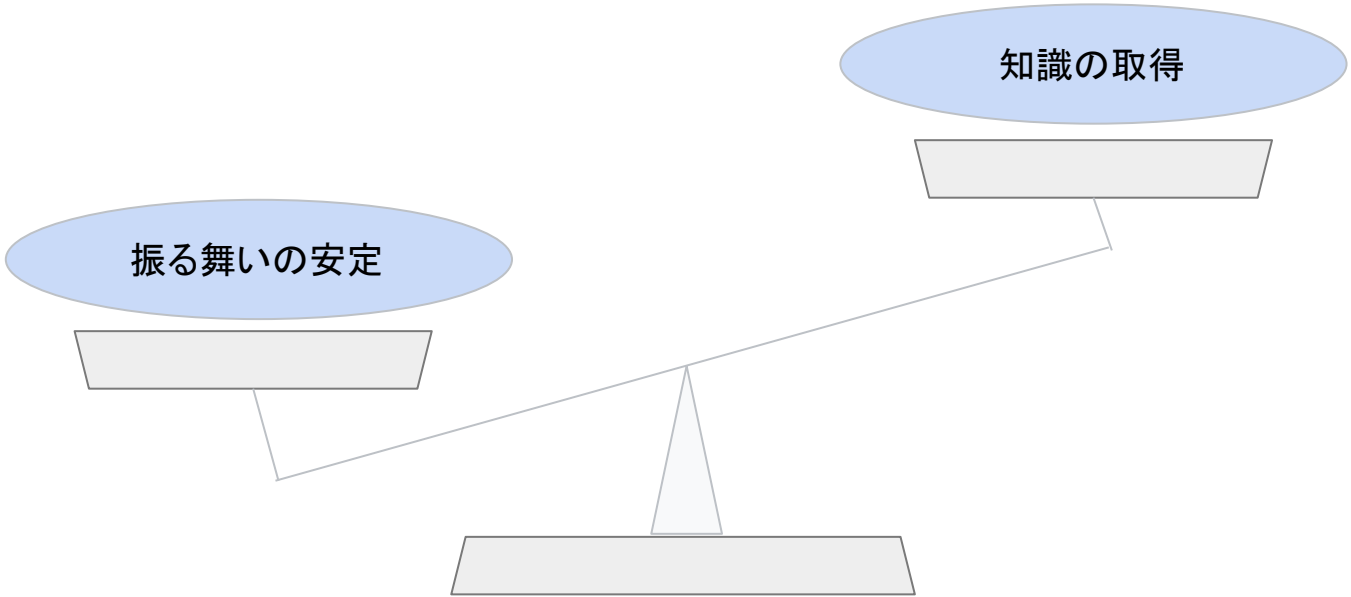
モデルをより良くするために人間のフィードバックを使用
人間のフィードバックを用いた強化学習（RLHF）を利用して
フィードバックに基づいてモデル パフォーマンスを最適化

それぞれのビジネスにカスタマイズされた画像を生成
Imagen のカスタマイズ機能、オブジェクト チューニングで
製品やロゴに基づいて画像を生成

スタイル チューニングで独自データのスタイルに沿って画像を生成

Supervised Tuning のユースケース

振る舞いを安定させるものであり、知識を習得させる要素にはあまり適さない



参考: Tune text models by using supervised tuning

LLM の代表的なカスタマイズ手法

1

プロンプト デザイン

入力プロンプトを工夫することで、モデルの再学習をさせることなく、LLM に期待する振る舞いをさせる手法

2

ファイン チューニング

プロンプトと出力のデータセットを用意することで、LLM のパラメータを効率的に更新する手法

3

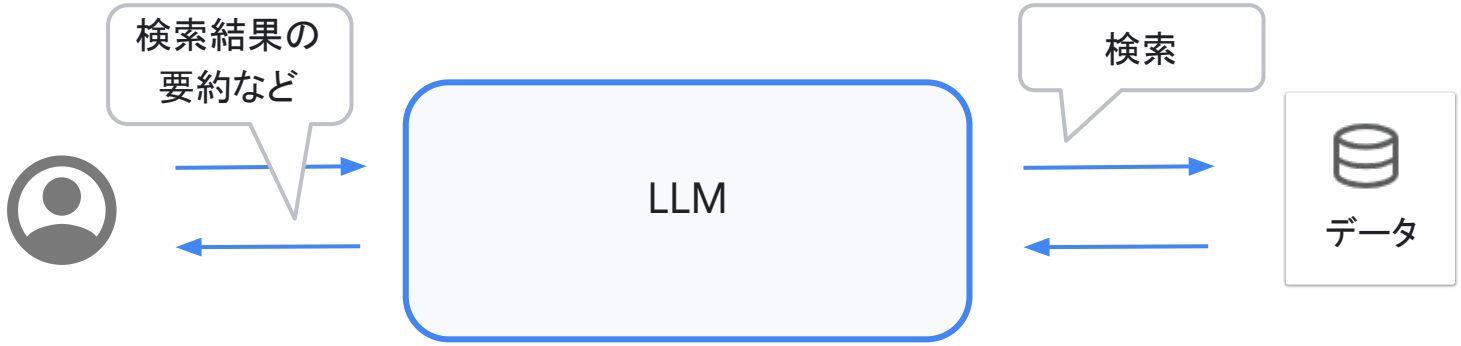
グラウン ディング

回答する情報を信頼できるソースから参照する主要。

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

グラウン ディングとは

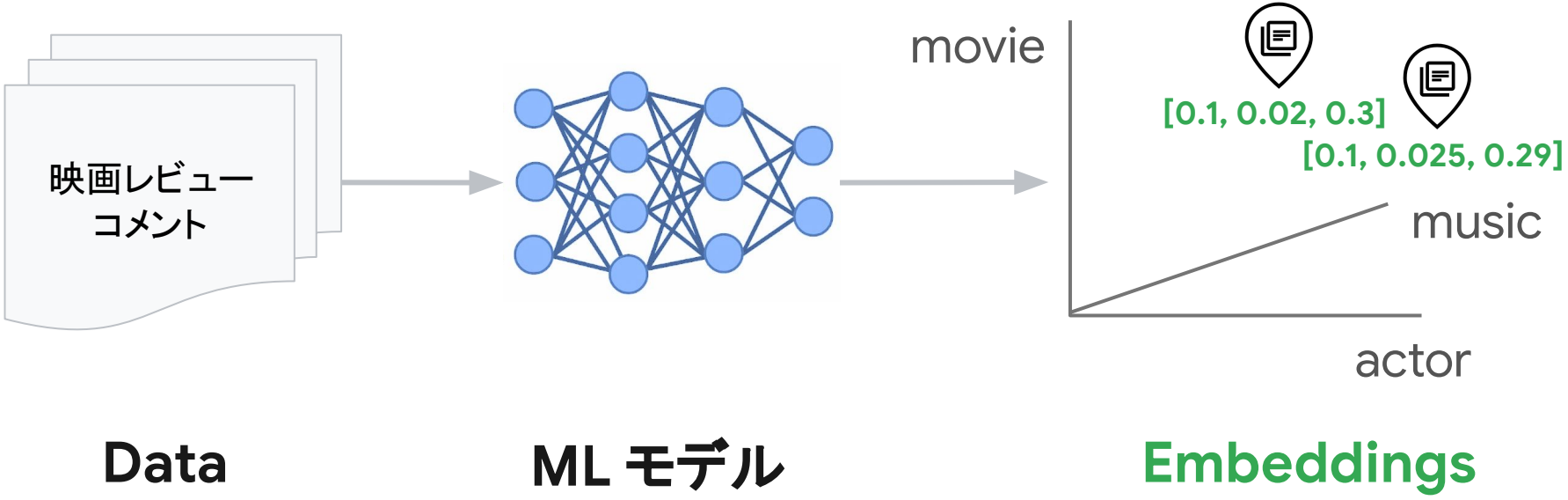
利用者が指定した情報だけに基づいて、LLM に回答を生成させる手法



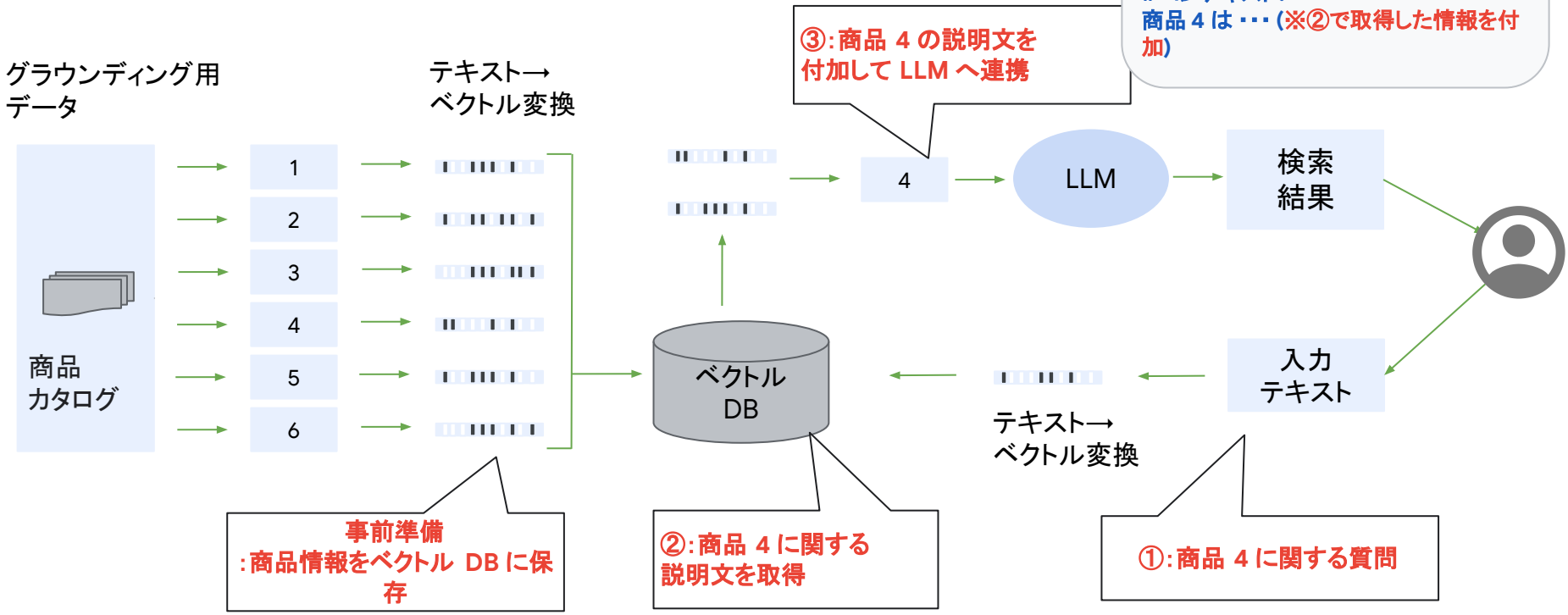
LLM が持っている知識で回答させるのではなく **検索と生成を組み合わせる** アプローチ

- 質問に対してもっともらしい答えを外部データベースから取得（検索）
- 単なる検索結果を返すだけでなく、検索で得られた情報を LLM のインプットとして、検索結果の要約などを行ってユーザーに返す（生成）

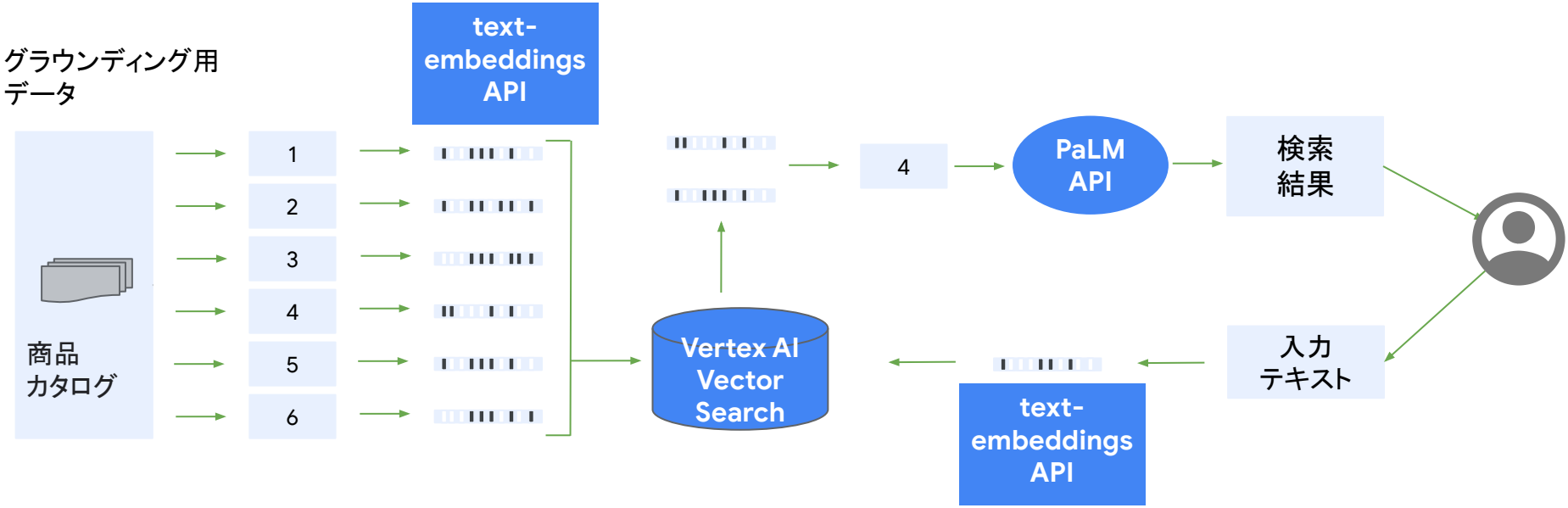
データのベクトル化 について



グラウン ディングの代表的なアーキテクチャ



グラウン ディングの代表的なアーキテクチャ - Google Cloud



※ 日本語対応版 text- embeddings API: textembedding-gecko-multilingual
※ Vertex AI Vector Search: https://cloud.google.com/vertex-ai/docs/vector-search/overview

グラウン ディングの実装手法

LangChain

LLM を利用してアプリケーションを効率的に開発するためのフレームワーク
自前でコードを実装しアプリケーションをカスタマイズ

Vertex AI Search & Conversation

生成 AI を用いた検索、対話型アプリケーションの構築機能
UI 付きで、より素早くリッチに開発

開発ガイド

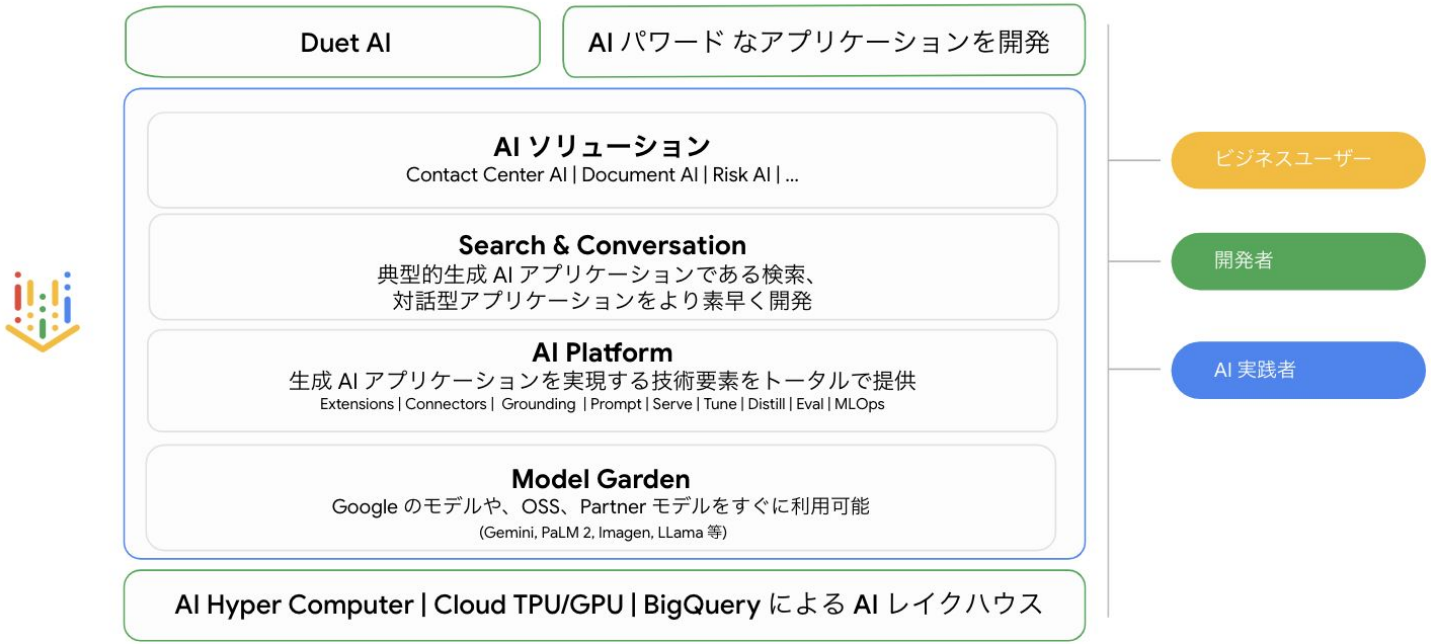
- Github repository - LangChain / Vertex AI Search & Conversation 含むサンプルコード
  - https://github.com/GoogleCloudPlatform/generative-ai
- Zenn blog - 生成 AI アプリケーション作成例
  - https://zenn.dev/google_cloud_jp/articles/google-cloud-generative-ai

03

まとめ

まとめ

- 「生成 AI を触ってみる」→「生成 AI を活用したソリューションを作る」フェーズへ
- Vertex AI を利用して、用途ごとに効果的なソリューションを構築する



Thank you.

Google Cloud