

# 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

Sler 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。

最近では生成 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

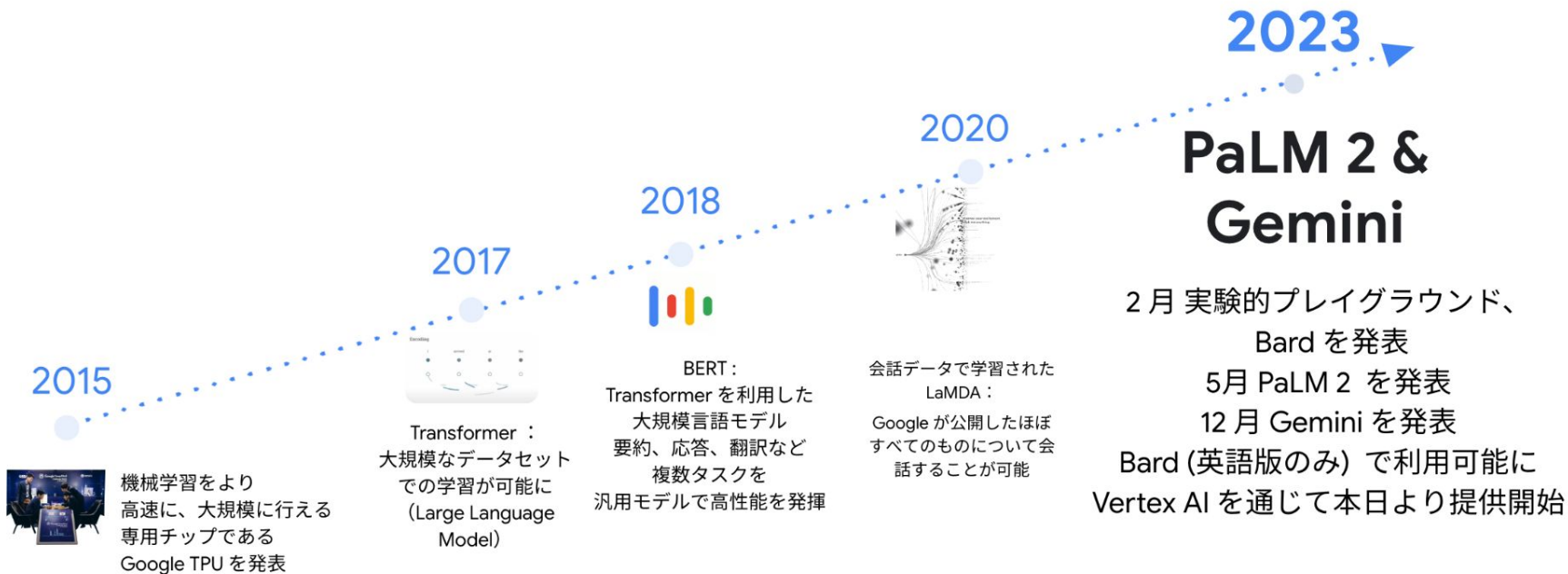
<b>Google Cloud における生成 AI</b>	<b>01</b>
<b>生成 AI ソリューションの開発</b>	<b>02</b>
<b>まとめ</b>	<b>03</b>

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

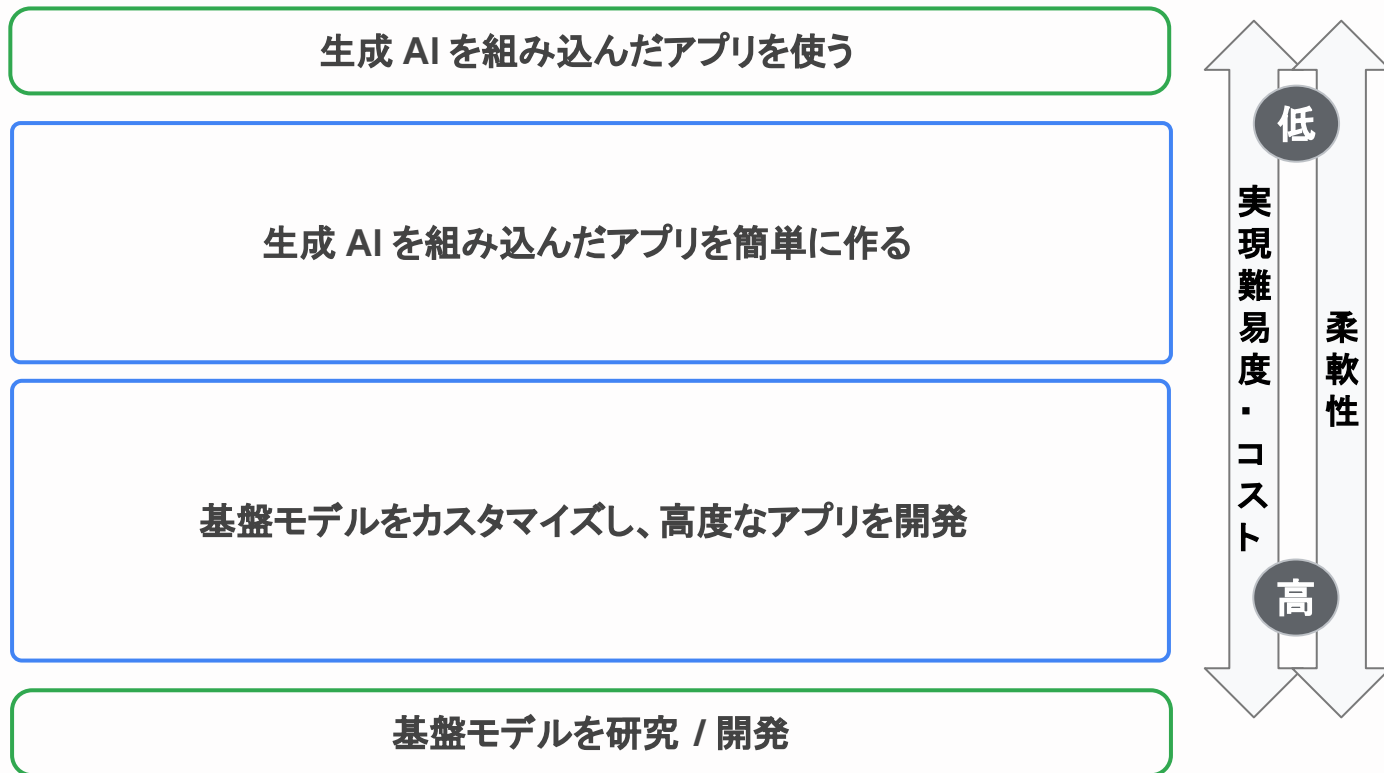


生成 AI を触ってみる



**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、Extensions 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**

Google AI Studio

Vertex AI

Gemini モデル

*Ultra, Pro, Nano*



# 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

The screenshot displays the Google Cloud Model Garden interface. At the top, it says "Model Garden" with links to "EXPLORE GENERATIVE AI" and "VIEW MY MODELS". A search bar is present with the text "Search models". Below the search bar, there are suggestions for "text embedding", "essay outline", and "BERT".

The main content is divided into sections:

- Modalities:** A list of modalities with their respective counts: Language (51), Vision (81), Speech (2), Tabular (2), Documents (2), and Video (3).
- Tasks:** A list of tasks with their respective counts: Generation (54), Classification (47), Detection (28), Extraction (9), Recognition (9), Translation (5), Embedding (2), Segmentation (4), Retrieval (1), Open vocabulary detection (2), and Open vocabulary segmentation (2).
- Features:** A list of features with their respective counts: Generative AI Studio (13).

Under the "Foundation models" section, there are two model cards:

- Gemini Pro:** A Generative AI model for Language. Description: "Gemini Pro." URL: "google/gemini-pro".
- Gemini Pro Vision:** A Generative AI model for Language. Description: "Gemini Pro Vision." URL: "google/gemini-pro-vision".

Below these cards is a link to "SHOW ALL (58)".

Under the "Fine-tunable models" section, there are two model cards:

- tftHub/EfficientNetV2:** A model for Classification and Vision. Description: "EfficientNet V2 are a family of image classification models, which achieve better parameter efficiency and faster training speed than prior arts." URL: "tensorflow-hub/efficientnetv2".
- tfvision/vit:** A model for Classification and Vision. Description: "The Vision Transformer (ViT) is a transformer-based architecture for image classification." URL: "tfvision/vit-s16".

# 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

Google 製の 基盤モデル						
Google 製 タスク専用 モデル	 Speech-to-Text  Text-to-Speech  Natural Language  Translation  Doc AI OCR  Occupancy analytics  Vision  Video Intelligence					
Google 製の ドメイン特化 モデル		<b>MedLM</b> Life Science and Healthcare Preview 様々なタスクに適用可能な大型モデルと ファイン チューニングの可能なさまざまなタスクを スケールできる中規模モデル (数ヶ月中に Gemini ベースも導入)				
パートナーと オープンエコ システム	Llama 2 Code Llama	Falcon	Claude 2 Pre-announce	Mistral ImageBind DITO Announce 		

# 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

The screenshot displays the Vertex AI Studio interface. The top section shows a 'Sample Prompt' with a text input area containing a prompt in Japanese: '与えられた見出しは何のニュース記事でしょうか。分類してください。 テキスト: 映画の封切りのためにイタリア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 ラベル: ビジネス, エンターテインメント, 健康, スポーツ, テクノロジー'. Below the prompt is a 'Response' section with the label 'エンターテインメント'. The bottom section shows the generated Python code for interacting with the model, including imports for 'vertexai' and 'TextGenerationModel', and a function to predict the response based on the prompt and parameters like 'candidate\_count', 'max\_output\_tokens', 'temperature', 'top\_p', and 'top\_k'.

# 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

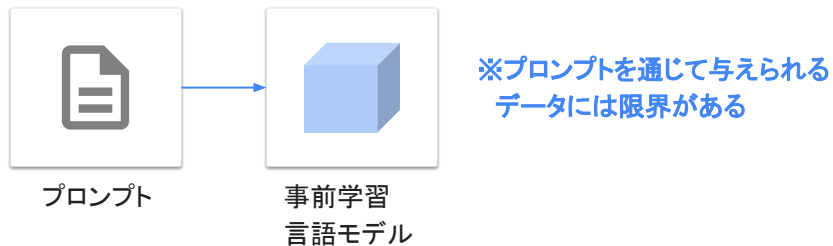
回答する情報を信頼できるソースから参照する主要。

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

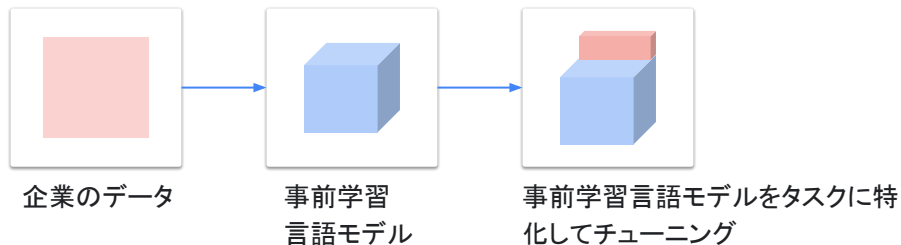


# プロンプト デザイン vs ファイン チューニング

## プロンプト デザイン



## 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 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

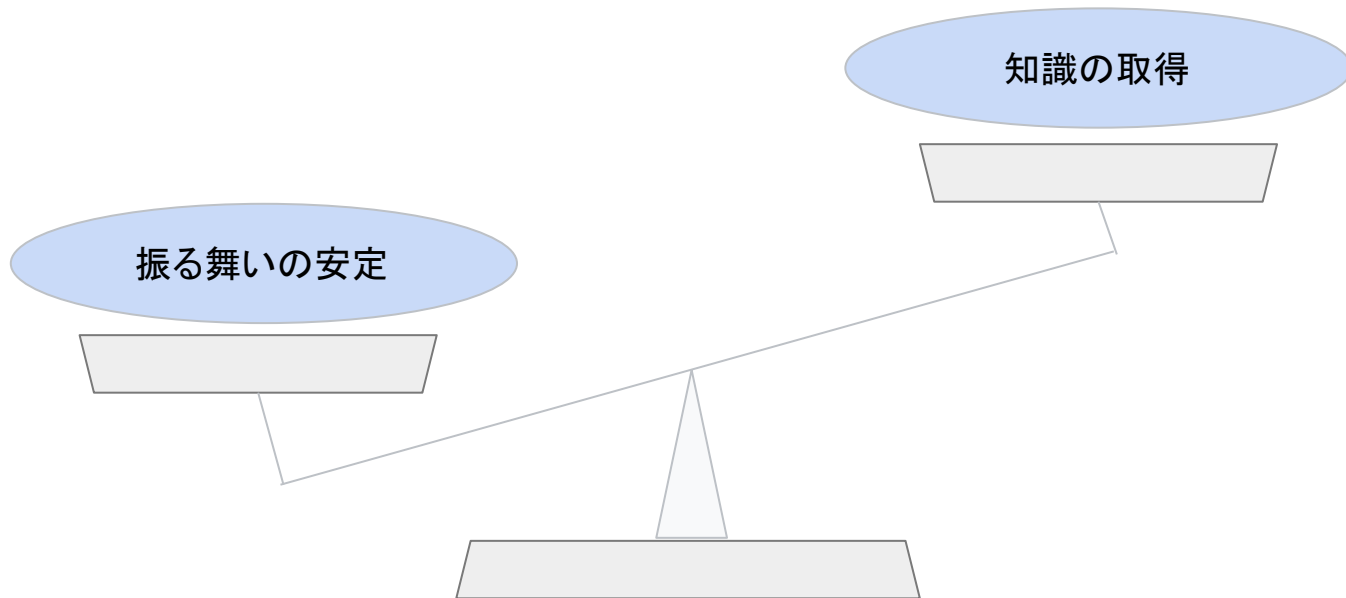
Imagen のカスタマイズ機能、オブジェクト チューニングで製品やロゴに基づいて画像を生成

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



# 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

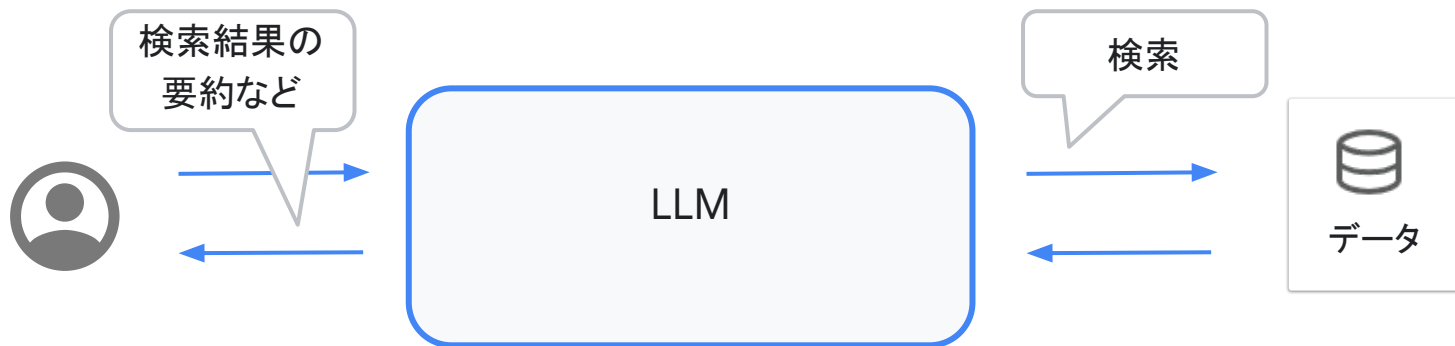
## グラウンディング

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

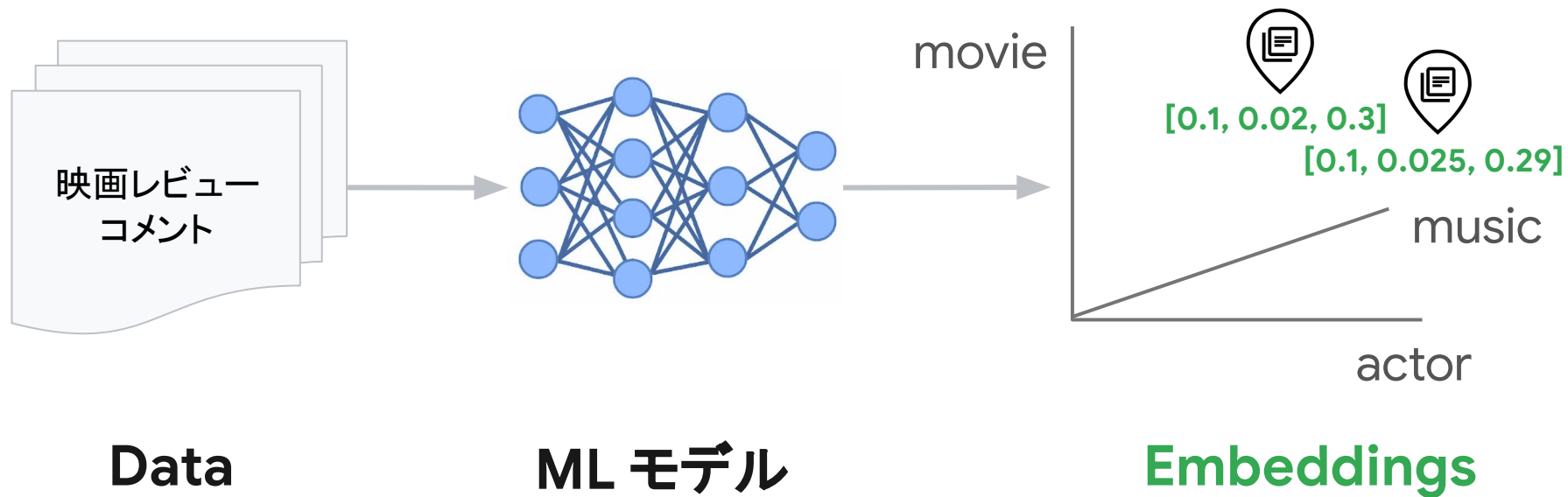
利用者が指定した情報だけに基づいて、LLM に回答を生成させる手法



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 の入力として、検索結果の要約などを行ってユーザーに返す(生成)

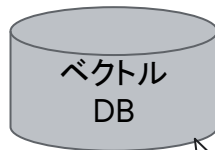
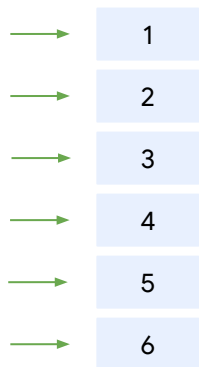
# データのベクトル化 について



# グラウンディングの代表的なアーキテクチャ

グラウンディング用  
データ

テキスト→  
ベクトル変換



4

LLM

検索  
結果

入力  
テキスト

テキスト→  
ベクトル変換

**事前準備**  
: 商品情報をベクトル DB に保  
存

**②: 商品 4 に関する  
説明文を取得**

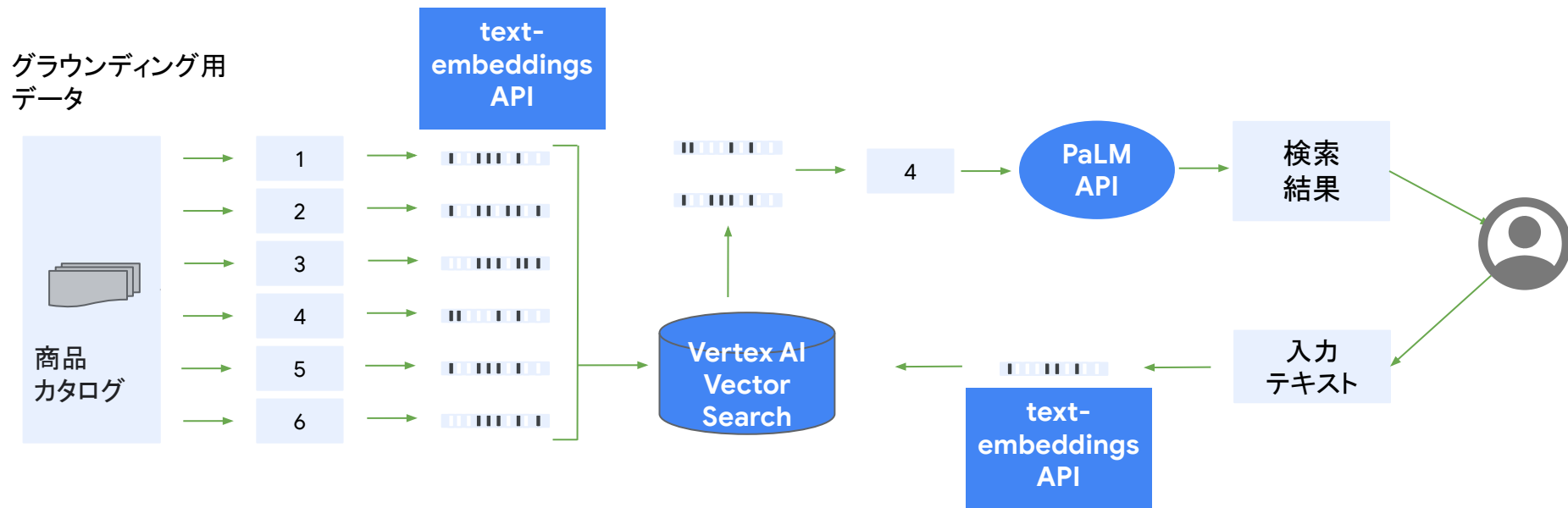
**①: 商品 4 に関する質問**

**③: 商品 4 の説明文を  
付加して LLM へ連携**

# 命令:  
商品の概要について教えてください。  
回答は、以下の情報だけに基づき作成し  
てください。

# コンテキスト:  
商品 4 は ... (※②で取得した情報を付  
加)

# 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](https://cloud.google.com/text-embeddings-api/docs/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



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

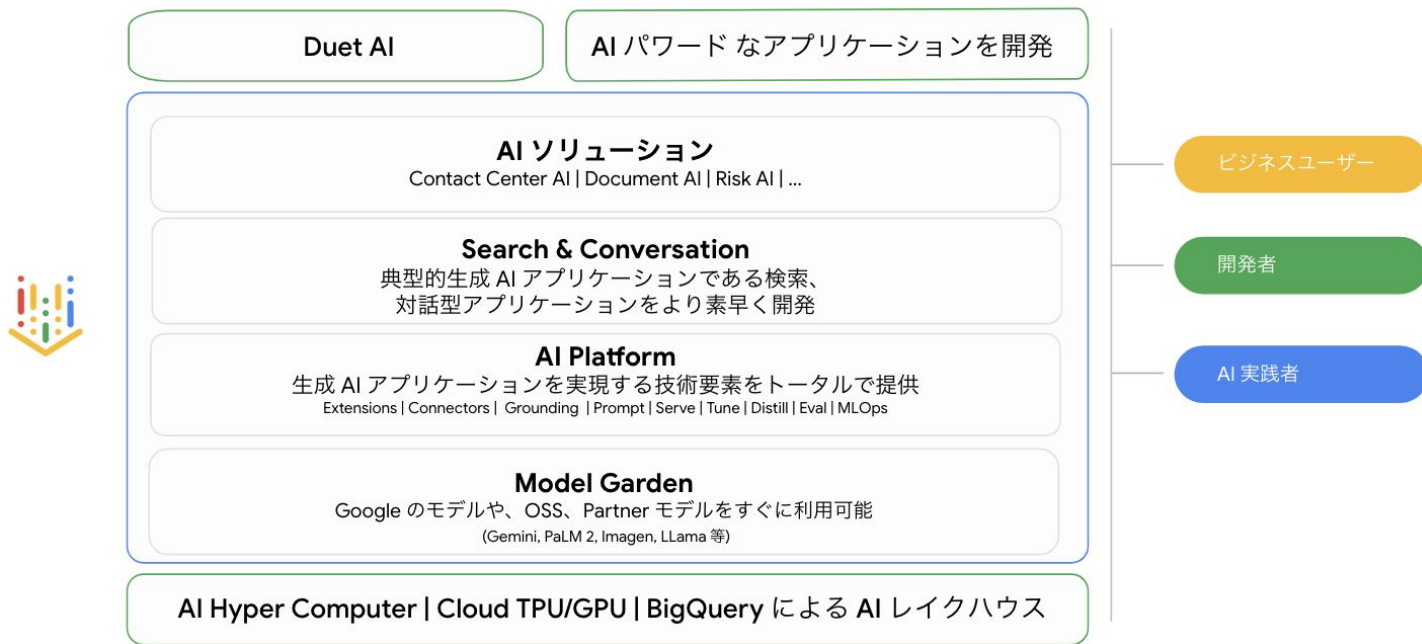
- Github repository - LangChain / Vertex AI Search & Conversation 含むサンプルコード
  - <https://github.com/GoogleCloudPlatform/generative-ai>
- Zenn blog - 生成 AI アプリケーション作成例
  - [https://zenn.dev/google\\_cloud\\_jp/articles/google-cloud-generative-ai](https://zenn.dev/google_cloud_jp/articles/google-cloud-generative-ai)

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.**