Proprietary + Confidential

Google Cloud

金融高度化セミナー

生成 AI における自社データの活用

グーグル・クラウド・ジャパン合同会社
フィナンシャルサービス事業本部
執行役員 事業本部長
綱田 和功

2024/4/25

Proprietary + Confidential

綱田 和功（つなだ かずのり）

グーグル・クラウド・ジャパン合同会社
執行役員 フィナンシャルサービス事業本部長

セコム株式会社で、業務システムの開発に約10年間従事。基幹システムの刷新プロジェクトを主導。

1999年に日本マイクロソフトに入社。ITコンサルタントとして15年間、様々な企業の業務システムの企画、設計、開発、導入プロジェクトに参画。大手金融機関の大型プロジェクトを多数経験。その後、業務執行役員金融サービス営業統括本部長。金融機関のクラウドを活用したDXを推進。

2022年よりグーグル・クラウド・ジャパン合同会社で執行役員フィナンシャルサービス事業本部長。金融機関向けのクラウド事業に携わる。

Proprietary + Confidential

Agenda

01 検索技術をベースに、生成AI を活用する

02 大規模言語モデルの進化：マルチモーダルモデル

03 マルチモーダルモデルの銀行業へのインパクト

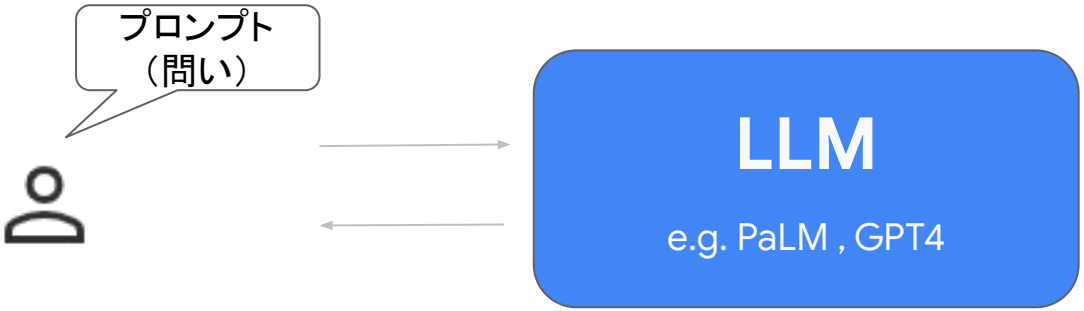
04 自社データの活用におけるセキュリティ確保

検索技術をベースに、
生成 AI を活用する

01

Proprietary + Confidential

大規模言語モデル(LLM)を使った問い合わせ業務で多くの人が行き詰まる点



知識が豊富な LLM を使うと、手始めに社内のヘルプデスクなどの問い合わせ業務で活用できそうと考える人は多い

ただし、やってみると、

- **自社独自のデータに基づいて回答できない**
- **ハルシネーション（幻覚）を起こす -> LLMの知識にない嘘の答えが返る**

この2つが課題になるので、解決する必要があります

le Cloud

Proprietary + Confidential

ソリューション：ファインチューニング？

	プロンプトチューニング	ファインチューニング	人による強化学習 (RLHF)	Distilling Step-by-Step
	Pre-trained Model Frozen	Tuned Adapter / Pre-trained Model Frozen	Tuned Adapter / Pre-trained Model Frozen	Update Student Model Weights
どのように機能するか？	Carefully author an input prompt or dialog to the model	Supervised model tuning can teach model new skills and change its behaviour	Predict rewards consistent w/ preferences (e.g. thumbs up/down), and align the model behavior with the preference	Extracts rationales from larger teacher models as additional supervision for smaller student models
トレーニングデータ	1-10 labeled examples	Hundreds labeled examples	5K-10K pairwise examples (A is better than B) and 1K-3K unlabeled prompts	Hundreds or thousands of of unlabeled prompts
対応モデル	All foundational models	Text-bison, chat-bison, code-bison, codechat-bison	Text-bison T5-S, L, XL, and XXL, Llama2	Teacher models: text-bison, text-unicorn Student model: text-gecko

Proprietary + Confidential

ファインチューニング のユースケース

Scenarios to use text model tuning

Prompt design works well when the desired output or task can be easily and concisely defined in a prompt and the prompt consistently produces the desired output. Tuning is required when you want the model to learn something niche or specific that deviates from general language patterns. For example, you can use model tuning to teach the model the following:

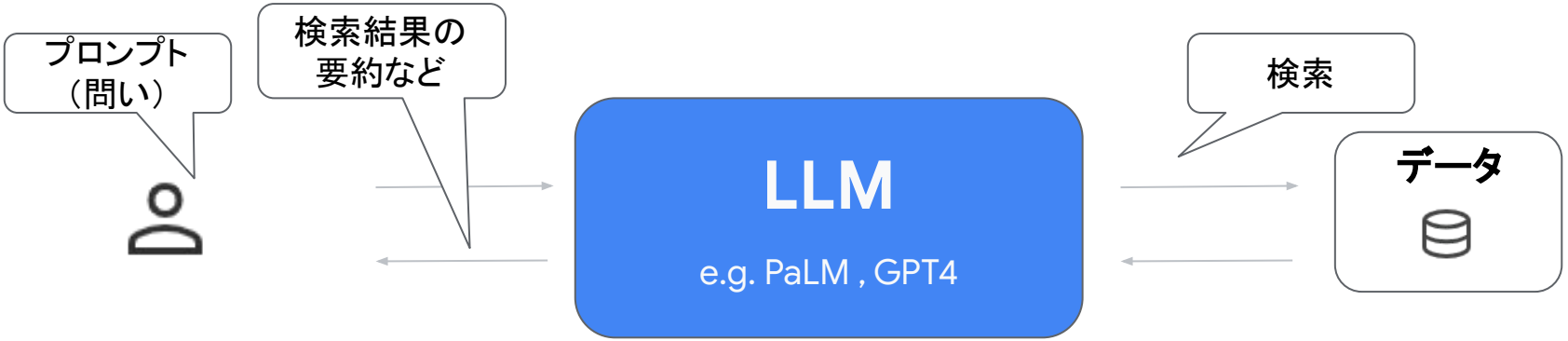
- Specific structures or formats for generating output.
- Specific behaviors such as when to provide a terse or verbose output.
- Specific customized outputs for specific types of inputs.

Tune text models | Vertex AI | Google Cloud より

つまり出力のフォーマットを整えたい、用語のちょっとした変換など、 LLMの振る舞いを安定させるための要素がかなり強いと言えます。
したがって自社データに基づいた結果を得たいというニーズに必ずしも答えられないという現実があります。

ソリューション：RAG (検索拡張生成)

Proprietary + Confidential



質問に対して、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）

LLM に外部データを参照させることを Grounding 、検索と LLM を連携させることでよりリッチな検索体験を作ることは RAG (Retrieval Augmented Generation) と呼ばれている

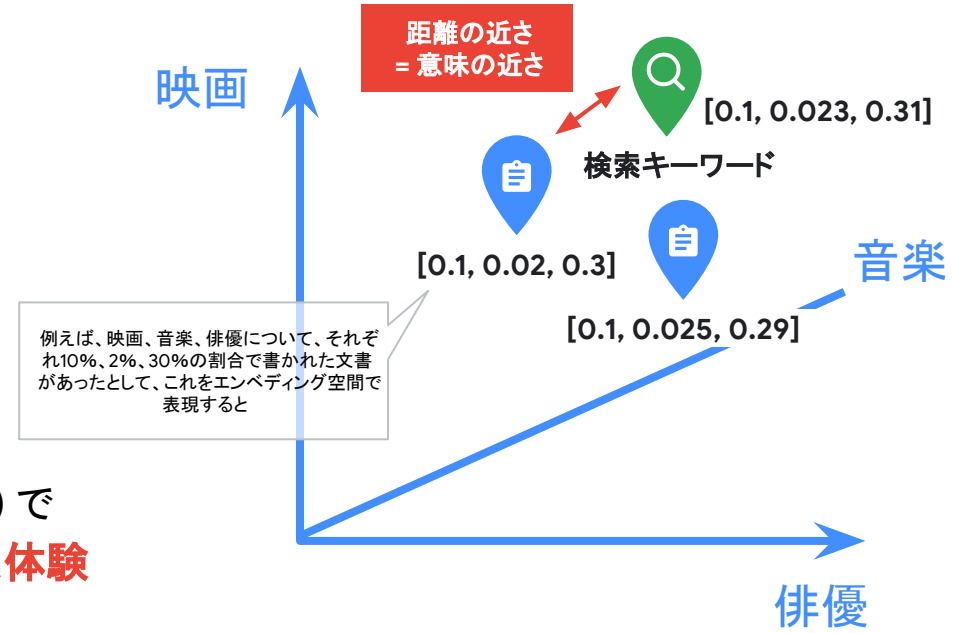
RAG の一つの実装例： Google 検索

Proprietary + Confidential

Proprietary + Confidential

エンベディング (埋め込み) と ベクトル検索 (意味検索)

- コンテンツを**埋め込み空間のベクトル**に変換
- 検索キーワードを**埋め込み空間のベクトル**に変換して検索
- 検索キーワードと**意味が近いコンテンツ**が検索できる



従来の手法 (AND 検索や OR 検索など) では難しかった、**検索意図を理解した検索体験**が提供できる

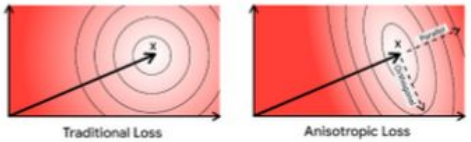
Proprietary + Confidential

高速かつスケーラブルなベクトル検索（意味検索）

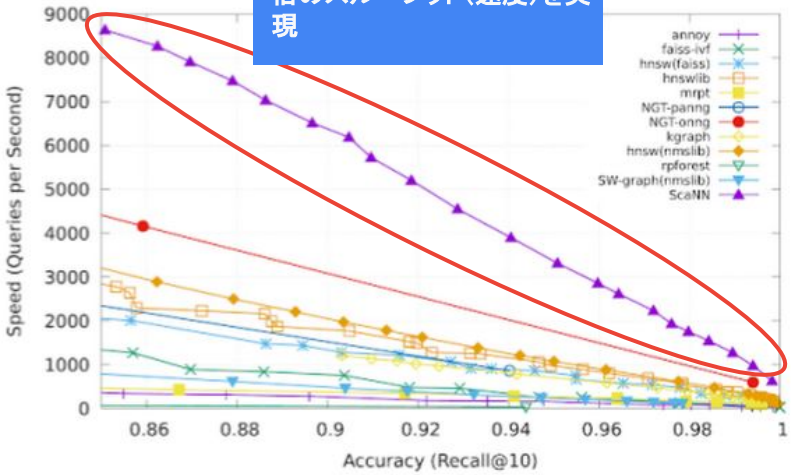
2020 年、Google Research は ScaNN という新しい ベクトル検索アルゴリズムを発表しました。これは**業界で最も優れた ベクトル検索アルゴリズム**の一つとされており、**Google 検索**や **YouTube** などの主要な Google サービスにおける基盤として重要な役割を担っています。

ScaNN: The vector search service for Google Search, YouTube and Play

Achieving higher accuracy with shorter latency



近似値を求める場合、他のアルゴリズムより数倍から10倍のスループット（速度）を実現



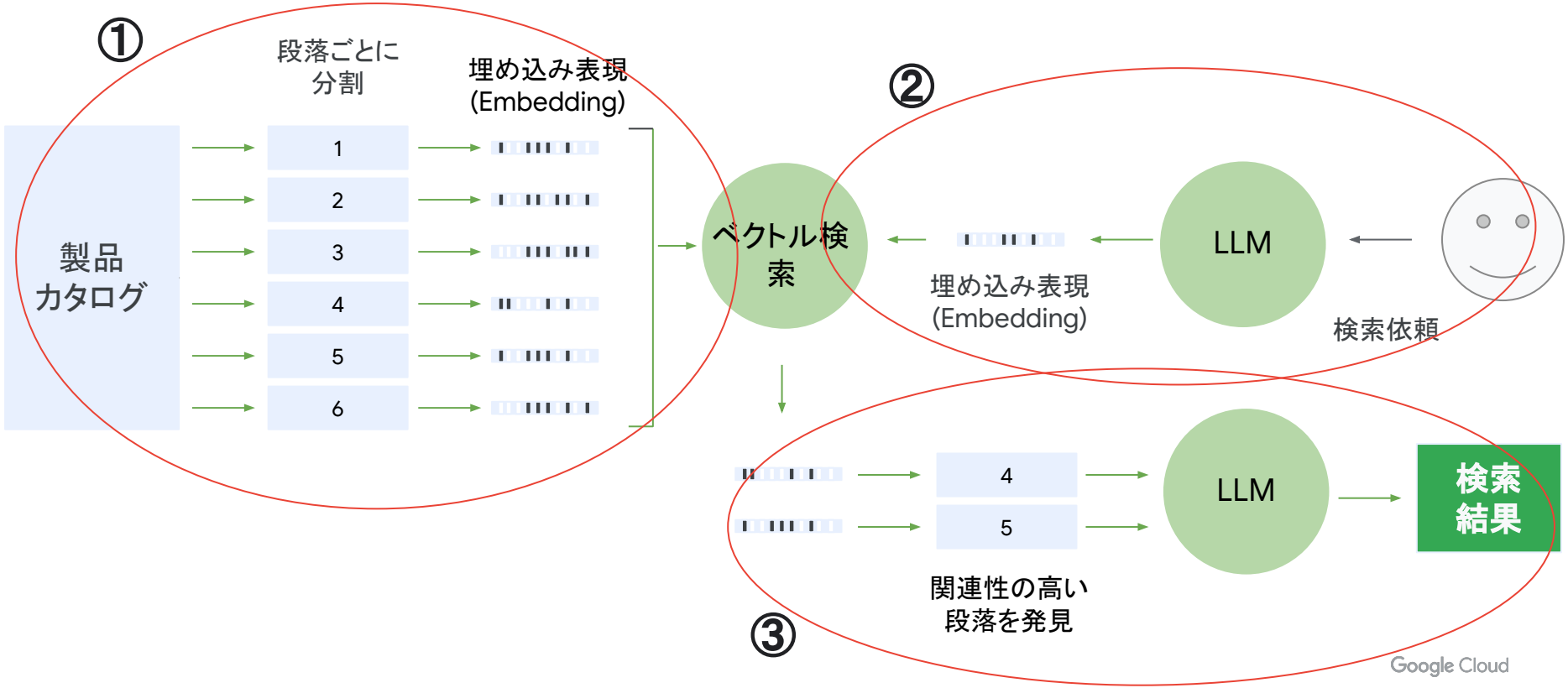
Proprietary + Confidential

Google 検索の強み:
エンベディング (埋め込み) と ベクトル検索 (意味検索)

擬似メルカリサイトにおける商品検索のデモ

Proprietary + Confidential

RAGの実装



Google Cloud

Proprietary + Confidential

RAGの実装

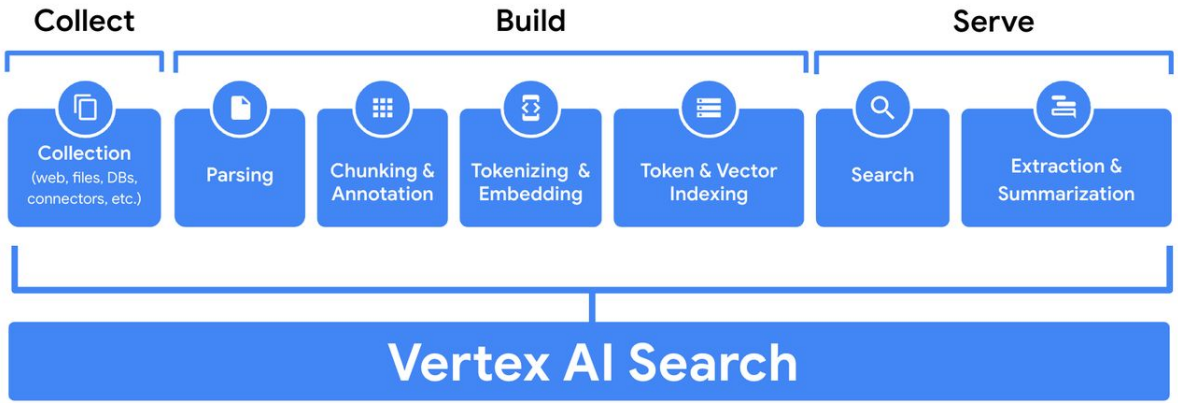
Confidential & Proprietary

Vertex AI Search: Google 検索テクノロジーを活用した RAG

https://cloud.google.com/blog/products/ai-machine-learning/rags-powered-by-google-search-technology-part-1?hl=en

Google や他の企業や研究組織の AI 研究者や開発者は、**テキスト チャンク、クエリ拡張、ハイブリッド検索、ナレッジ グラフ、リランキングなどを含む、多数の RAG テクノロジを研究してきました。**つまり、LLM には、でっちあげにならないように適切なデータが必要であり、最適なデータを取得するには最適な検索テクノロジーが必要です。

Google 検索テクノロジーは、大規模な情報検索の先駆者として 20 年以上の経験に裏付けられており、RAG ソリューションに強力な情報検索機能を提供します。Vertex AI Searchを使用すると、開発者は LLM と Google 検索テクノロジーを簡単に組み合わせることができるため、**RAG システム用の独自の高度な検索エンジンの設計と構築に数か月かける代わりに、AI アプリの構築に集中できます。**



大規模言語モデルの進化
マルチモーダルモデル

02

Gemini は

初期から複数のデータ種類で
トレーニングを行うことにより、会話をコード、画像、動画、
音声、テキストの間をより
シームレスに扱うことが可能に。

※これまでの伝統的なマルチモーダルの開発は、はテキスト、コード、画像などを異なるTransfomer をつかってから組み合わせて実現をしていた

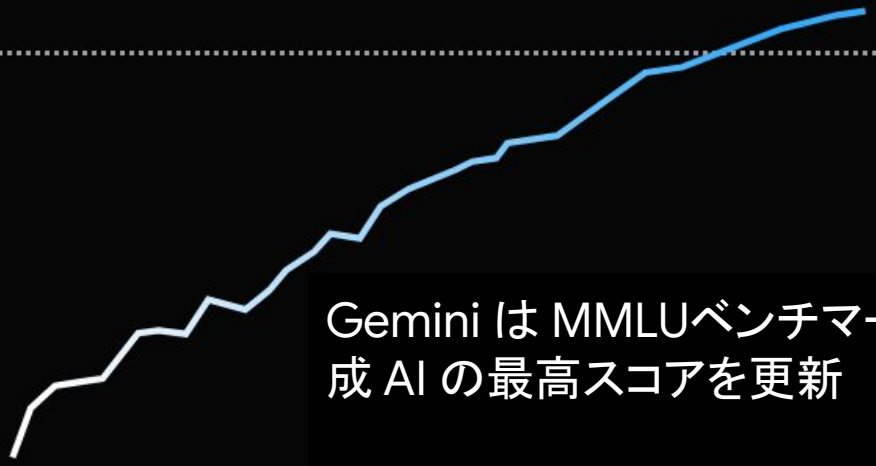
Gemini Ultra

90.0%

CoT@32*

89.8%

Human expert (MMLU)



86.4%

5-shot* (reported)
Previous SOTA (GPT-4)

*Note that evaluations of previous SOTA models use different prompting techniques.

Gemini は MMLUベンチマーク ※においても、生成 AI の最高スコアを更新

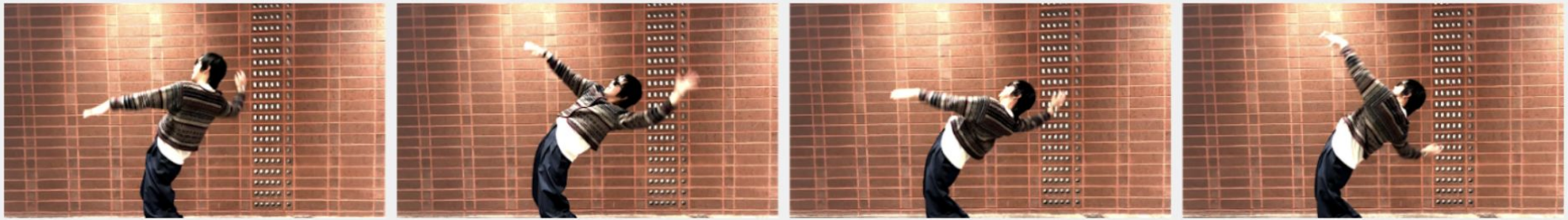
人間のエキスパートのスコアを超え、
90%を達成した初めての AI モデル

※ 57の領域の専門知識を問うベンチマーク。自然科学、文化人類学などが含まれる

https://deepmind.google/technologies/gemini/#capabilities

マルチモーダル プロンプトの例

複数の画像とプロンプトより推論することで経緯を推論



Guess what movie I'm acting out.（なんの映画をまねてると思う？）

Gemini: The Matrix (マトリックスだね)

Nice! But which part specifically? Look at my body movements.（いいね、でもどの部分が？体の動きをみて）

Gemini: The part where Neo dodges bullets. （Neo が弾丸を避けるシーンだね）

そのプロンプトの入出力例
https://developers.googleblog.com/2023/12/how-its-made-gemini-multimodal-prompting.html

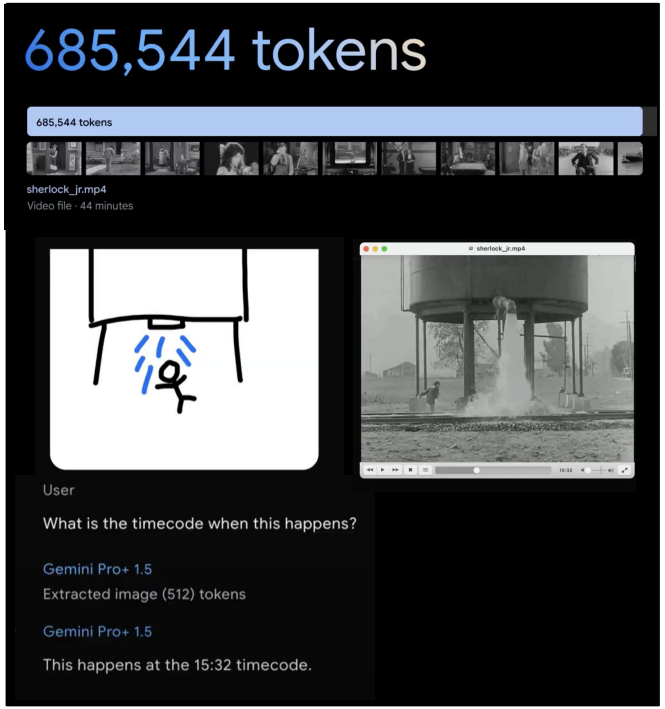
Gemini 1.5 Pro

画期的なロングコンテキストの理解を備えた中規模のマルチモーダル　モデル

Private Preview

Gemini 1.5 Pro は劇的に強化されたパフォーマンスを提供し、次のような基盤モデルのアプローチに大きな変化をもたらします ：

- モデルのパフォーマンスを向上させながら、より効率的なトレーニングとサービスを提供する **新しいエキスパート混合（MoE）アーキテクチャ**
- 膨大な量の情報にわたる複雑な推論のための **拡張されたコンテキスト ウィンドウ**（最大 100 万トークン）
- テキスト、コード、画像、オーディオ、ビデオを含む **モダリティ全体での理解と推論の向上**
- 安全性リスクに関する新しい研究に基づいて構築され、レッドチーム技術を活用してさまざまな潜在的な危害をテストする **広範な倫理および安全性テスト**



Proprietary + Confidential

LLMのコンテキストウィンドウとは何ですか？

回答案を表示

LLMにおけるコンテキストウィンドウとは

コンテキストウィンドウは、大規模言語モデル (LLM) が一度に処理できる **トークンの最大数** を指します。これは、LLMが文章を理解し、応答を生成する際に参照できる情報量に相当します。

コンテキストウィンドウが大きいほど、LLMは長い文章を理解し、より複雑なタスクを実行することができます。しかし、コンテキストウィンドウサイズが大きくなると、モデルの **計算量** も増大します。そのため、各LLMは、処理速度と精度の間でバランスを取ったコンテキストウィンドウサイズを採用しています。

具体的な例

- **GPT-3.5**: 4,096トークンから16,384トークン
- **Megatron-Turing NLG**: 5,376トークン
- **WuDao 2.0**: 10,000トークン
- **BLOOM**: 176,000トークン
- **PaLM**: 6,144トークン
- **PaLM 2**: 8,000トークン
- **Gemini**: 最大100万トークン (デフォルト: 128,000トークン)

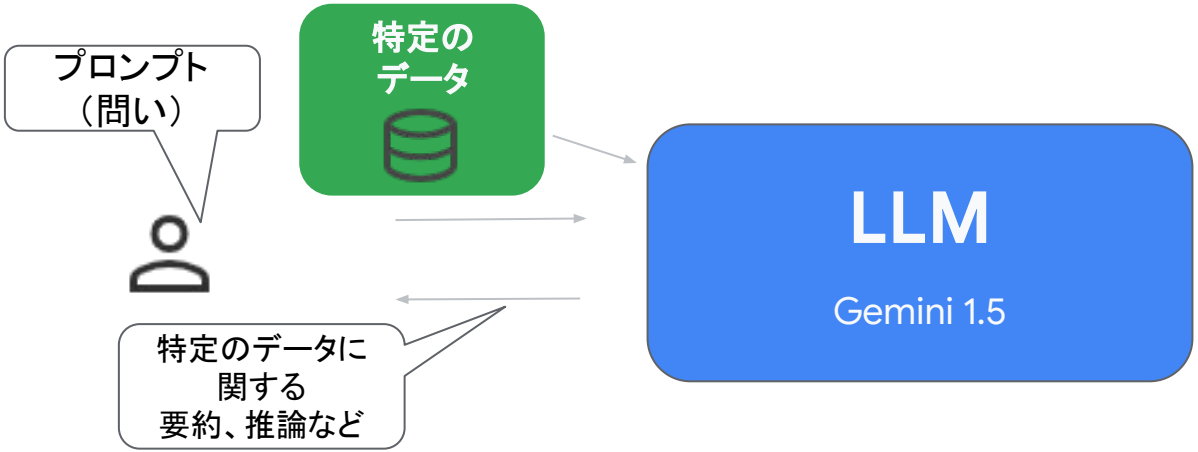
Gemini 1.5 Pro デモ: 402 ページの記録を通した推論

これは、402 ページの PDF トランスクリプトと一連のマルチモーダル プロンプトを使用した、
最新モデル Gemini 1.5 Pro の実験的な機能である、□いコンテキストの理解のデモです。

https://ntrs.nasa.gov/api/citations/20160014392/downloads/20160014392.pdf

Proprietary + Confidential

RAGの実装に変化をもたらす Gemini 1.5



特定のデータ（文書、動画、音声など）について問い合わせをしたい場合は、外部データの検索を行うのではなく、そのデータに基づき LLMが一発で回答を生成することができる

- コンテキストウィンドウの大幅な拡張（最大 100万トークン）により実現
- 検索を使わないので、より正確な情報を元にした回答の生成が可能

マルチモーダルモデルの銀行業へのインパクト

03

Confidential & Proprietary

マルチモーダルモデル Gemini が銀行業に与えるインパクトを教えて？ (1/3)

マルチモーダルモデル Gemini は、テキストだけでなく、画像、音声、動画などの複数の種類のデータ (モーダリティ) を学習し、高度な理解と分析が可能になるモデルです。銀行業において、Gemini は以下の分野に大きなインパクトを与える可能性があります。

1. 顧客体験の向上

- **チャットボット: 顧客の表情や声のトーンを認識し、より自然な会話ができるチャットボットを提供**
- **パーソナライズされたサービス: 顧客の行動データに基づき、顧客一人ひとりに最適な金融商品・サービスを提案**
- **これらによる顧客満足度向上、顧客離脱防止**

2. 業務効率化

- **書類処理: 画像データから必要な情報を自動的に抽出し、データ入力作業を効率化**
- **音声認識: 顧客との対話内容を自動的に記録し、議事録作成を効率化**
- **これらによる業務時間短縮、人材の有効活用**

Confidential & Proprietary

マルチモーダルモデル Gemini が銀行業に与えるインパクトを教えて？ (2/3)

3. リスク管理

- 不正検知: 監視カメラの映像や音声データから、不正取引をリアルタイムで検知
- 顔認証: 生体認証技術と組み合わせることで、セキュリティを強化
- これらによる不正取引の防止、被害額の削減

4. 新たなビジネスモデルの創出

- 音声バンキング: 音声認識技術を用いて、電話で銀行取引を行うサービス
- ウェアラブルデバイス: スマートウォッチなどのウェアラブルデバイスで金融サービスを提供
- これらによる新たな収益源の創出

Confidential & Proprietary

マルチモーダルモデル Gemini が銀行業に与えるインパクトを教えて？(3/3)

以下は、Gemini 導入の成功事例です。

- **銀行 A**: チャットボットに Gemini を導入し、顧客満足度が 20% 向上
- **銀行 B**: 音声認識システムに Gemini を導入し、業務時間を 10% 短縮
- **銀行 C**: 顔認証システムに Gemini を導入し、不正取引を 50% 削減

これらの事例は、Gemini が銀行業に大きなインパクトを与え得ることを示しています。

Gemini 導入にあたり、銀行固有の、データ収集・管理の課題、モデル開発・運用コストの課題、AI倫理の課題など、克服すべき課題も存在します。

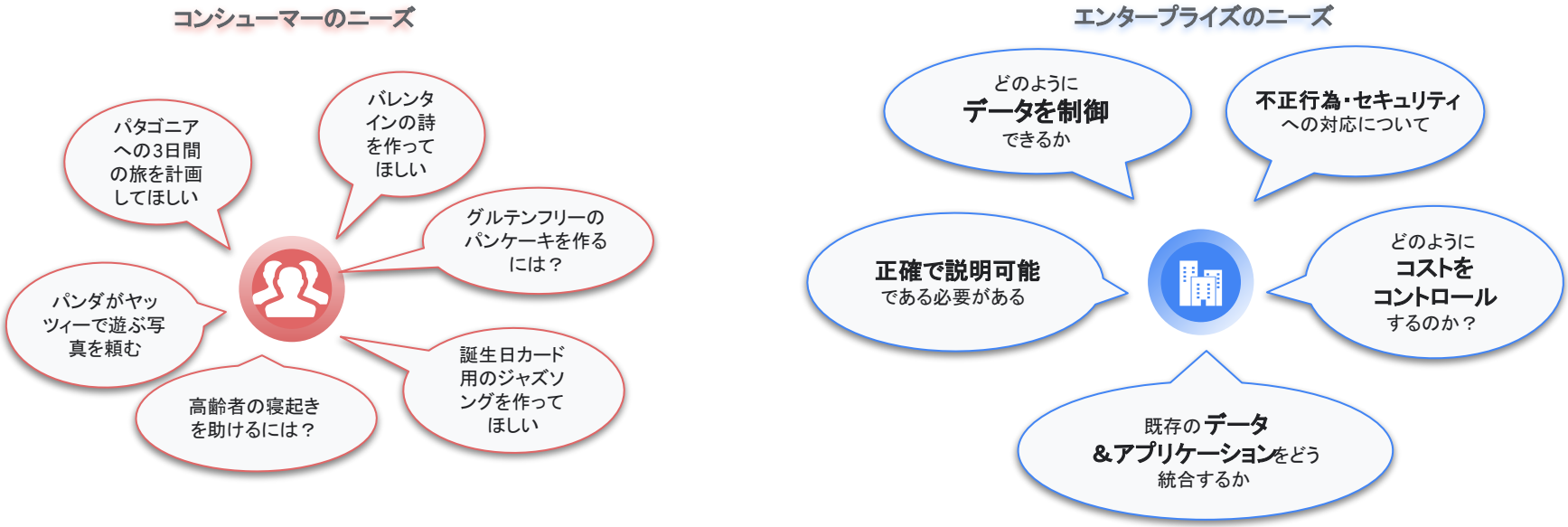
銀行の皆様がこれらの課題を克服し、マルチモーダルモデル Gemini を効果的に活用することで、競争力強化、顧客満足度向上、業務効率化を実現できると信じております。

自社データの活用におけるセキュリティ確保

04

Proprietary + Confidential

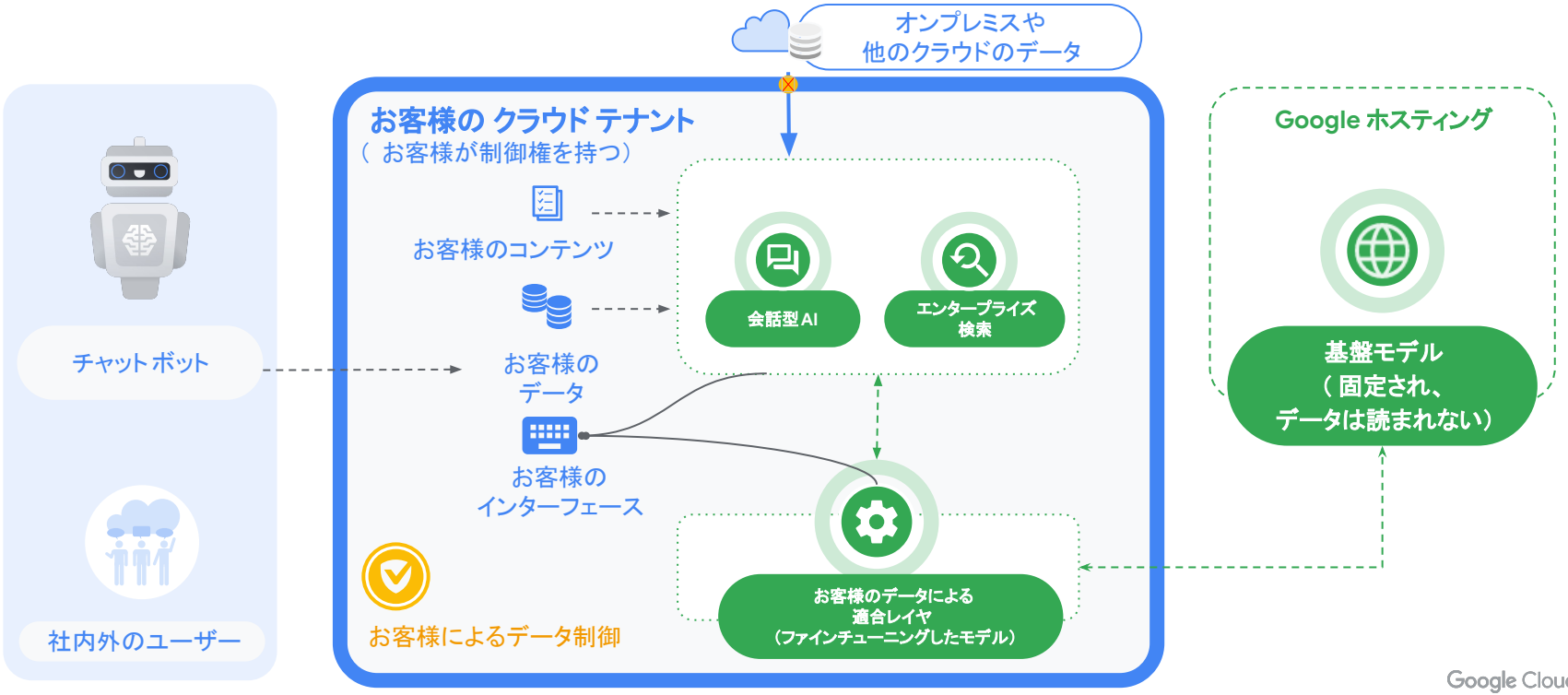
生成 AI: 一般ユーザーと企業とで異なるニーズ



Proprietary + Confidential

生成 AI の企業利用での注意点
データ保護のエンタープライズ アーキテクチャが必須

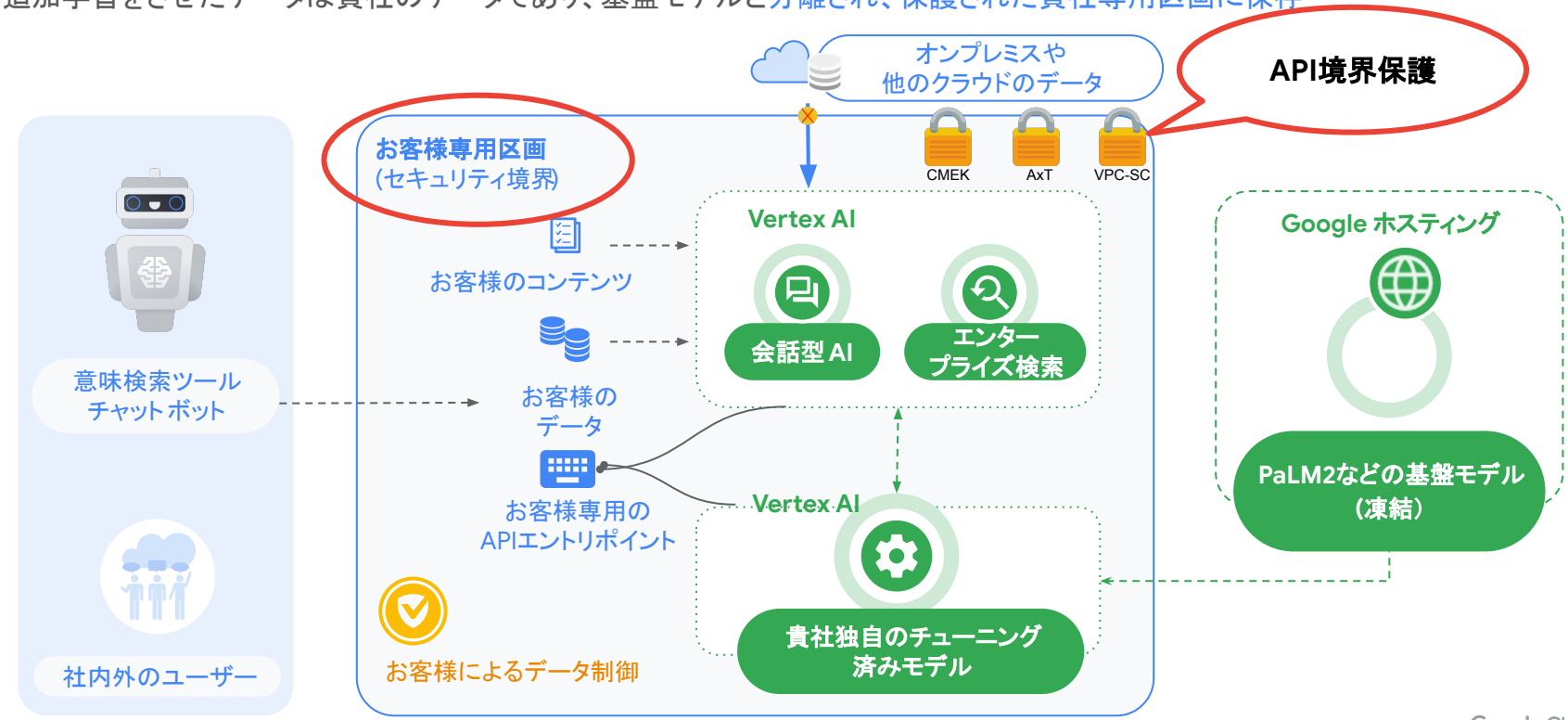
追加学習をさせたデータは自社のデータであり、基盤モデルと異なる分離、保護がされたプラットフォームが必須



Proprietary + Confidential

Google は、お客様データを保護するアーキテクチャを標準装備

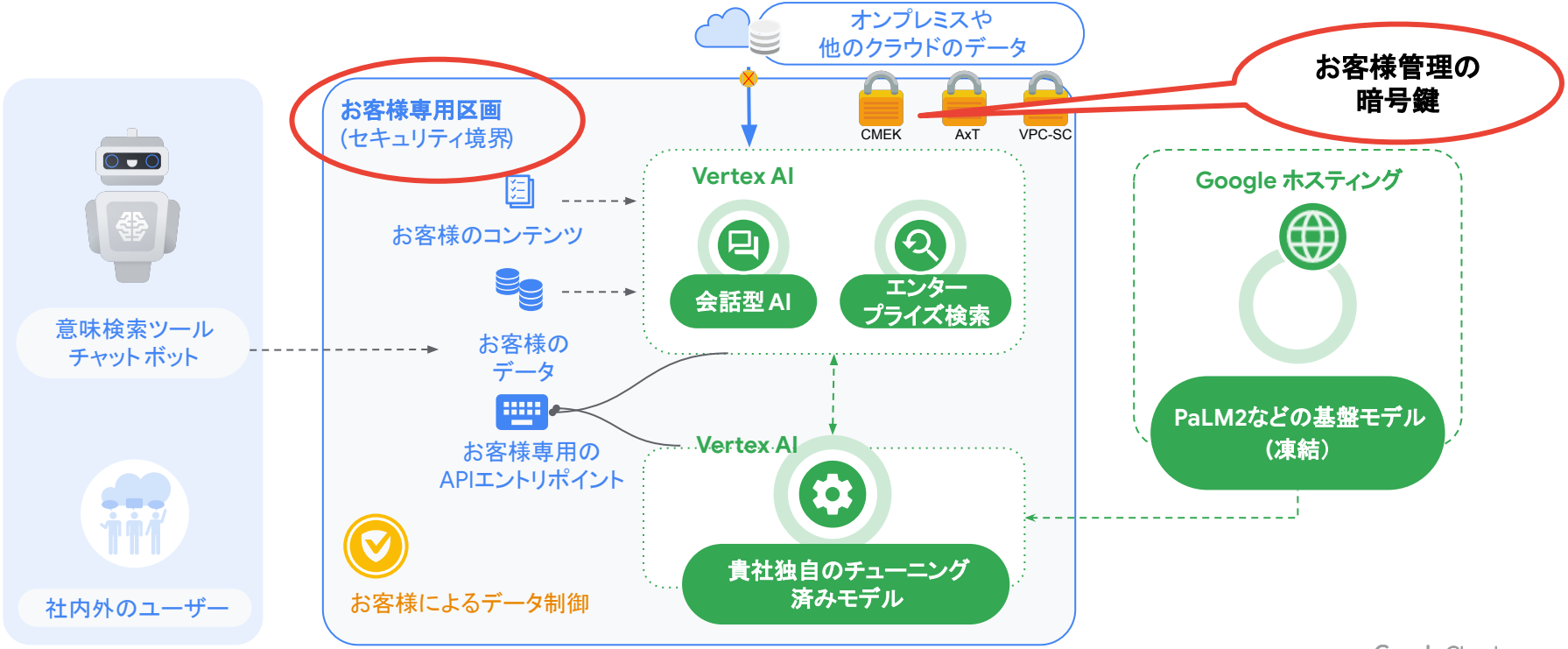
追加学習をさせたデータは貴社のデータであり、基盤モデルと分離され、保護された貴社専用区画に保存



Proprietary + Confidential

Google は、お客様データを保護するアーキテクチャを標準装備

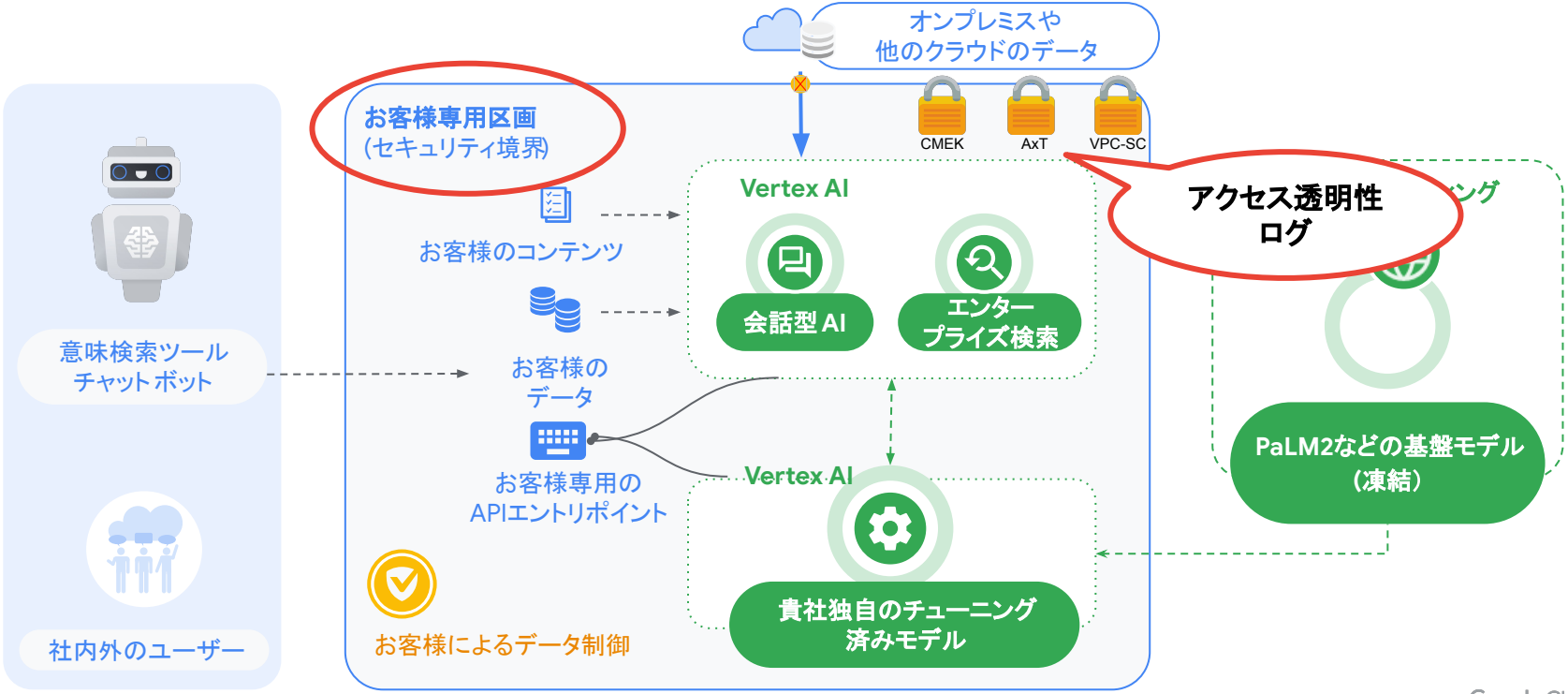
追加学習をさせたデータは貴社のデータであり、基盤モデルと分離され、保護された貴社専用区画に保存



Proprietary + Confidential

Google は、お客様データを保護するアーキテクチャを標準装備

追加学習をさせたデータは貴社のデータであり、基盤モデルと分離され、保護された貴社専用区画に保存



Confidential & Proprietary

[2023/12]

Vertex AI におけるお客さまデータの取り扱い

データ ガバナンスと生成 AI (https://cloud.google.com/vertex-ai/docs/generative-ai/data-governance?hl=ja)

基盤モデルのトレーニング

デフォルトでは、**Google Cloud は基盤モデルのトレーニングに顧客データを使用しません。**お客様は Google Cloud の基盤モデルを使用できますが、その際、**お客様のプロンプト、レスポンス、およびアダプタモデルのトレーニング データは、基盤モデルのトレーニングに使用されません。**

アダプタモデルのトレーニング

Vertex AI には、お客様がアダプタモデルをトレーニングできるサービスが用意されています。**アダプタモデルのトレーニング データは顧客データであり、保存されません。**また、**Google Cloud の基盤モデルの改善にも顧客データは使用されません。**アダプタモデルは、アダプタモデルをトレーニングしたお客様のみが使用できます。Google は、アダプタモデルが既存の Google の知的財産を使用する範囲を除き、アダプタモデルの所有権を主張しません。デフォルトでは、顧客データは暗号化された形式で保存され、転送中も暗号化されます。また、お客様は、顧客管理の暗号鍵（CMEK）を使用してアダプタモデルの暗号化を制御し、いつでもアダプタモデルを削除できます。

Confidential & Proprietary

[2023/12]

Vertex AI におけるお客さまデータの取り扱い

データ ガバナンスと生成 AI (https://cloud.google.com/vertex-ai/docs/generative-ai/data-governance?hl=ja)

予測

予測中に基盤モデル、アダプタモデル、安全性分類器によって処理される入力と出力は、顧客データです。**顧客データは、お客様の出力の生成に必要な期間を超えて Google に保存されることはなく、Google の基盤モデルのトレーニングに使用されることもありません。**たとえば、Vertex AI 基盤モデルと安全分類器を使用してお客様が作成したドキュメント要約アプリケーションの場合、Google は、入力ドキュメントも生成された要約も保存せず、基盤モデルと安全分類器は変更されません。

Vertex AI へInput したデータや Output されたデータはお客様からの明示的な許可がない限り保存されることはありません。**お客様からの明示的な許可を得てデータを保存するケースとしては、Gemini モデルのキャッシュが挙げられます。Gemini モデルは後続のプロンプトへの応答速度を上げるため、最大で 24 時間のキャッシュを保持します。本キャッシュは無効化することが可能です。**

Google Cloud

Thank you!