

最新の生成 AI モデル へのアップデートに必 要な LLMOps

Google
Cloud
Next

Tokyo

Proprietary



牧 允皓

Google Cloud
AI Solutions Architect



“

今、皆さんが使っている生成 AI のモデルいつまで使えるかご存知ですか？ ”

Gemini の場合

モデル ID	リリース日	廃止日	詳細
gemini-2.5-pro	2025 年 6 月 17 日	2026 年 6 月 17 日	
gemini-2.5-flash	2025 年 6 月 17 日	2026 年 6 月 17 日	
gemini-2.0-flash-001	2025 年 2 月 5 日	2026 年 2 月 5 日	Gemini 2.0: Flash、Flash-Lite、Pro - Google デベロッパー ブログ
gemini-2.0-flash-lite-001	2025 年 2 月 25 日	2026 年 2 月 25 日	Gemini 2.0: Flash、Flash-Lite、Pro - Google デベロッパー ブログ

日々進化する生成 AI を使いこなすために

LLM (大規模言語モデル) の進化と普及により、
AI を使ったシステムの運用も複雑化

常に最新のモデルを使い、高い精度によって
ビジネスの価値を生むには**運用の設計**が重要

LLMOps への道程

01

精度劣化のリスク

02

LLMOps の概要

03

Vertex AI の機能

Vertex AI Studio

Prompt Optimizer

Gen AI Evaluation Service

01. 精度劣化のリスク

カスタマー サポートの半自動化（仮想）

PoC

データを収集し、実際に精度が出るか、価値が出るかを検証

例: カスタマー サポートの待ち時間が長く、生成 AI による半自動化を目指す
オフラインの検証で
精度 95% を実現

リリース

カスタマー サポートの一次受付を AI にさせ、より緊急度の高い問い合わせを人間が対応

サービスの離脱率が
5% 軽減

運用

特に問題がなさそうなので、**ログの記録とマシンの監視だけ継続**

サービス劣化

待ち時間が長くなったと SNS で話題になっていることを観測し、システムを確認

新商品・新サービスに関する問い合わせが増え、一次受付の AI が正しく優先順位をつけられずに
ユーザが離脱

モニタリングの例

入力データの変化

PoC で用意したデータセットと、実際にリリースした後のデータは必ずしも一致しない

新しい商品名、想定していなかった言語など、様々な要因で入力データは変化する

モデル精度

入力データが変化するとモデルの精度も変化する可能性が高い

評価用データセットの準備、更新、再評価は重要

サービス自体の品質

モデルの精度では説明できない要因は、離脱率、CTR、顧客満足度などのビジネス KPI で評価することも重要

02. LLMOps の概要

“

LLMOps(大規模言語モデル運用)とは、大規模言語モデル(LLM)の管理と運用に関連する手法とプロセスを指します。LLM は、テキストやコードの膨大なデータセットでトレーニングされた AI モデルで、テキストの生成、翻訳、質問への回答など、言語関連のさまざまなタスクを実行できます。”

データ マネジメント

高品質なデータを使用

LLM を効果的にトレーニングするには、大量の高品質なデータが必要。

組織は、トレーニングに使用するデータがクリーンで正確であり、目的のユースケースに関連していることを確認する必要がある。

データを効率的に管理

LLM は、トレーニングと推論中に膨大な量のデータを生成できる。組織は、ストレージと取得を最適化するために、データ圧縮やデータパーティショニングなどの効率的なデータ マネジメント戦略を実装する必要がある。

データ ガバナンスの確立

LLMOps のライフサイクル全体を通じてデータの安全かつ責任ある使用を確保するために、明確なデータ ガバナンス ポリシーと手順を確立する必要がある。

モデルのトレーニング

適切なトレーニング アルゴリズムを選択

LLM やタスクの種類によって、適切なトレーニング アルゴリズムがある。組織は、利用可能なトレーニング アルゴリズムを慎重に評価し、特定の要件に最も適したものを選択する必要がある。

トレーニング パラメータを最適化

ハイパー パラメータ調整は、LLM のパフォーマンスを最適化するために重要。学習率やバッチサイズなどのさまざまなトレーニング パラメータを試して、モデルに最適な設定を見つける。

トレーニングの進行状況をモニタリングする

潜在的な問題を特定し、必要な調整を行うには、トレーニングの進行状況を定期的にモニタリングすることが不可欠。組織は、損失や精度などの主要なトレーニング指標を追跡するために、指標とダッシュボードを実装する必要がある。

デプロイ

適切なデプロイ戦略を選択

LLM は、クラウドベースのサービス、オンプレミスのインフラストラクチャ、エッジデバイスなど、さまざまな方法でデプロイできる。お客様の具体的な要件を慎重に検討し、お客様のニーズに最も適したデプロイ戦略を選択。

デプロイのパフォーマンスを最適化

デプロイ後は、LLM をモニタリングしてパフォーマンスを最適化する必要がある。これには、リソースのスケールリング、モデル パラメータの調整、レスポンス時間を改善するためのキャッシュ メカニズムの実装が含まれる場合がある。

セキュリティを確保

LLM と LLM が処理するデータを保護するために、強力なセキュリティ対策を実装する必要がある。これには、アクセス制御、データ暗号化、定期的なセキュリティ監査などがある。

モニタリング

モニタリング指標を 確立

LLM の健全性とパフォーマンスをモニタリングするために、重要業績評価指標 (KPI) を確立する必要がある。これらの指標には、精度、レイテンシ、リソース使用率などがある。

リアルタイム モニタリングの実装

運用中に発生する可能性のある問題や異常を検出して対応できるように、リアルタイム モニタリング システムを実装する必要がある。

モニタリング データの 分析

モニタリング データは定期的に分析して、傾向、パターン、改善の余地を特定する必要がある。この分析は、LLMOps プロセスの最適化と、高品質の LLM の継続的なデリバリーを保証するのに役立つ。

03. Vertex AI の機能

LLMOps の一例

プロンプト管理

目的のタスクを解くためにプロンプトをデザインする。

プロンプトの試行錯誤を管理し、実験を適切に記録する。

Vertex AI Studio

プロンプト最適化

新しいモデルに乗り換えることで精度向上を図る。モデルの変化に適応するため、プロンプトを最適化する。

Vertex AI
Prompt Optimizer

モデル評価

モデルの精度を定量的に評価することで、データに基づく意思決定が可能。

Vertex AI
Gen AI Evaluation
Service

自動化

定型化できる操作は、自動化することで運用コストとヒューマンエラーを減す

Vertex AI Pipelines

LLMOps の一例

プロンプト管理

目的のタスクを解くためにプロンプトをデザインする。

プロンプトの試行錯誤を管理し、実験を適切に記録する。

Vertex AI Studio

プロンプト最適化

新しいモデルに乗り換えることで精度向上を図る。モデルの変化に適応するため、プロンプトを最適化する。

Vertex AI
Prompt Optimizer

モデル評価

モデルの精度を定量的に評価することで、データに基づく意思決定が可能。

Vertex AI
Gen AI Evaluation
Service

自動化

定型化できる操作は、自動化することで運用コストとヒューマンエラーを減す

Vertex AI Pipelines

Vertex AI Studio



Vertex AI Studio - プロンプト作成

The screenshot displays the Vertex AI Studio interface for creating a prompt. The main workspace is titled "Customer Support Classification". It contains two prompt blocks, each highlighted with a red border. The first block is a "System Instruction" (システム指示) that sets the role to a customer support center receptionist. The second block is a "User Message" (ユーザーメッセージ) that provides a classification task with three categories: product, delivery, and contract. A red box highlights a context menu that appears when clicking on the prompt blocks, offering actions like "履歴" (History), "メモ" (Memo), "削除" (Delete), "API リファレンス" (API Reference), and "評価" (Evaluate). On the right side, there are settings for "出力" (Output) and "ツール" (Tools), including options for "構造化出力" (Structured Output), "思考予算" (Thought Budget), and "グラウンディング" (Grounding) using Google or Vertex AI Search.

Customer Support Classification

システム指示 ⓘ 🗑️

あなたは熟練のカスタマーサポートセンターの受付です。丁寧な対応を心がけてください。

次の問い合わせを、3つの分類候補から適切なものを選び、お客様に返信しなさい。
分類候補: {商品に関するお問い合わせ, 配送に関するお問い合わせ, ご契約に関するお問い合わせ}
問い合わせ: 先週買った冷蔵庫が壊れた。

72 個のトークン

履歴
メモ
削除
API リファレンス ⓘ
評価 ⓘ

出力

構造化出力 ⓘ

思考予算
自動 ⓘ

ツール

グラウンディング: Google
Google 検索または Google マップによるグラウンディング。

グラウンディング: 未指定
RAG Engine, Vertex AI Search,

フィードバックを共有 >

Vertex AI Studio - メモの編集

Customer Support Classification

システム指示 ⓘ ⓘ

あなたは熟練のカスタマーサポートセンターの受付です。丁寧な対応を心がけてください。

聞いてモデルが要約した考えを表示する

お問い合わせありがとうございます。

この度は、先週ご購入いただいた冷蔵庫の故障とのこと、大変ご不便をおかけし申し訳ございません。

お客様のお問い合わせは、以下の分類候補より『商品に関するお問い合わせ』として承ります。

お客様のお困り事を一日も早く解決できるよう、詳細を確認させていただきたく存じます。恐れ入りますが、よろしければ製品の具体（例：電源が入らない、冷えないなど）や、ご購入時の情報（購入日、ご購入店、製品の型番など）をお伺いしてもよろしいでしょうか。

お手数をおかけいたしますが、ご協力いただけますと幸いです。
引き続き、何卒よろしくお願い申し上げます。

📄 👍 🗣️ ✏️ 回答の質を高める

ここにプロンプトを入力します

521 個のトークン

v0
アプリケーション開発者がMVPで使っていたサンプルのプロンプトを流用。

38 / 2000

メモは管理目的でのみ使用され、モデルには送信されません。

確認 キャンセル

Vertex AI Studio - 履歷

Customer Support Classification_2025_07_03_181426

他)

問い合わせ: 先週買った冷蔵庫が壊れた。

AI

▼ 処理

聞いてモデルが要約した考えを表示する

この度は、お問い合わせいただき、誠にありがとうございます。受付を担当させていただきます〇〇(受付名)でございます。

先週ご購入いただきました冷蔵庫の故障とのこと、大変ご不便、ご心配をおかけし、誠に申し訳ございません。

お申し出の内容は、商品に関するお問い合わせとして承りました。お客様の生活に直結する大変緊急性の高い内容と認識しております。

つきましては、担当のオペレーターが詳細をお伺いするため、追ってお客様にご連絡させていただきます。恐れ入りますが、今しばらくお待ちいただけますようお願い申し上げます。

この度は、ご迷惑をおかけいたしますことを重ねてお詫び申し上げます。何卒よろしくようお願い申し上げます。

▼ メモ

オペレーターから連絡する旨を追記。

履歴

Customer Support Classification_2025_07_03_181426

2025/07/04 10:14:26

Customer Support Classification_2025_07_03_180944

2025/07/04 10:09:44

Customer Support Classification_2025_07_03_175803

2025/07/04 9:58:03

詳細

モデル	gemini-2.5-flash
温度	1
トークンの上限	65535
停止シーケンス	N/A
最大レスポンス	N/A
Top-K	N/A
Top-P	1
グラウンディング	N/A
シード	0

プロンプト、出力結果、メモ

パラメータ

プロンプト履歴

Vertex AI Studio - プロンプト比較

比較 プロンプトを送信 評価

ドキュメント

Customer Support Classification

次の問い合わせを、3つの分類候補から適切なものを選び、お客様に返信しなさい。

この度は、お問い合わせいただき、誠にありがとうございます。受付を担当させていただきます〇〇（受付名）でございます。

先週ご購入いただきました冷蔵庫の故障とのこと、大変ご不便、ご心配をおかけし、誠に申し訳ございません。

お申し出の内容は、商品に関するお問い合わせとして承りました。お客様の状況に合わせたサポートをご提供いたします。

このリージョンでは評価指標を利用できません

更新

新しいプロンプト

次の問い合わせを、3つの分類候補から適切なものを選び、お客様に返信しなさい。

お問い合わせありがとうございます。以下のように対応いたしました。

分類:
商品に関するお問い合わせ

お客様への返信:

お問い合わせいただき、誠にありがとうございます。

このリージョンでは評価指標を利用できません

新増促左

比較を追加して、さまざまなプロンプト、モデル、パラメータ設定によってモデルの出力がどのように変化するかを確認します。

- + 新しいプロンプトを比較
- + 保存したプロンプトを比較
- + グラウンドトゥールズ

LLMOps の一例

プロンプト管理

目的のタスクを解くためにプロンプトをデザインする。

プロンプトの試行錯誤を管理し、実験を適切に記録する。

Vertex AI Studio

プロンプト最適化

新しいモデルに乗り換えることで精度向上を図る。モデルの変化に適応するため、プロンプトを最適化する。

Vertex AI
Prompt Optimizer

モデル評価

モデルの精度を定量的に評価することで、データに基づく意思決定が可能。

Vertex AI
Gen AI Evaluation
Service

自動化

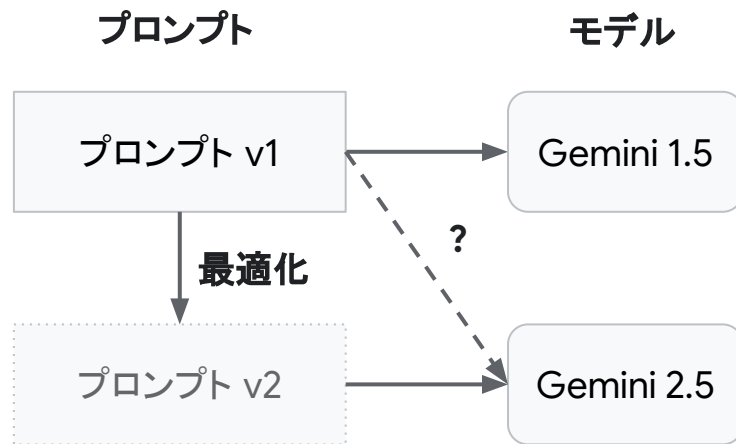
定型化できる操作は、自動化することで運用コストとヒューマンエラーを減す

Vertex AI Pipelines

モデル アップデートに必要な技術

日々新しいモデルがリリースされ、アプリケーション側のアップデートが精度の向上にもつながる

生成 AI のモデルを移行するには、プロンプトを新しいモデルに合わせて最適化することが重要



Prompt Optimizer でプロンプトの最適化

 プロンプト管理 ドキュメント

- 保存済みのプロンプトを表示し、最適化ツールにアクセスしてプロンプトを改善します。

 プロンプト作成サポート
目的を説明してください。必要な結果が得られるようにするため、プロンプトが生成されるか、既存のプロンプトが改善されます。
[開く](#)

 プロンプトの比較
調整可能なパラメータを使用してプロンプトを並べて比較し、どちらがより良い結果を生み出すかを確認してください。
[開く](#)

 プロンプトの最適化 プレビュー
Colab Enterprise ノートブックでラベル付きの例を使って Google モデルのプロンプトを最適化します。
[ノートブックを開きます](#)

 評価
データに対するモデルのパフォーマンスを把握できます。
[ノートブックを開く](#)

[プロンプトを作成](#) [プロンプトをインポート](#)

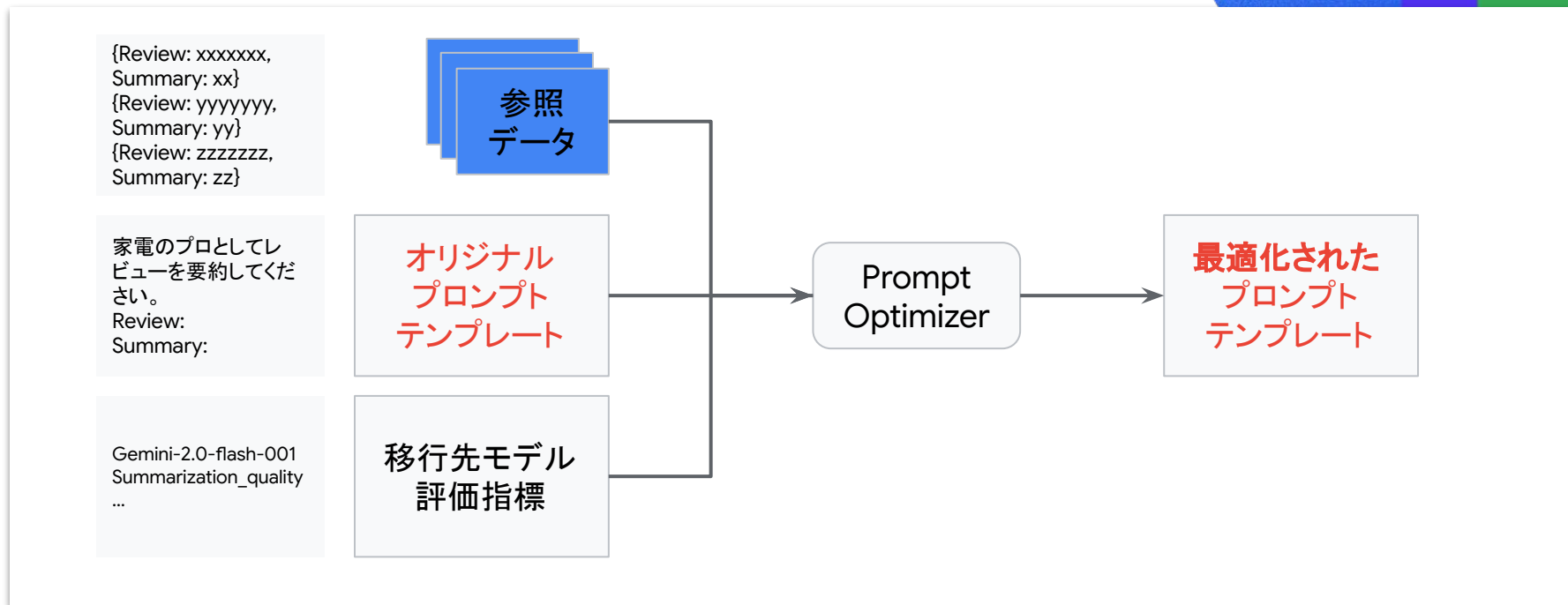
フィルタ プロンプトをフィルタします

☐	プロンプト名	メディア	メモ	最終更新	モデル ID	
☐	Japanese Customer...	いいえ		2025/07/04	gemini-2.5-flash	
☐	Customer Support...	いいえ	オペレーターから連絡する旨を追記。	2025/07/04	gemini-2.5-flash	

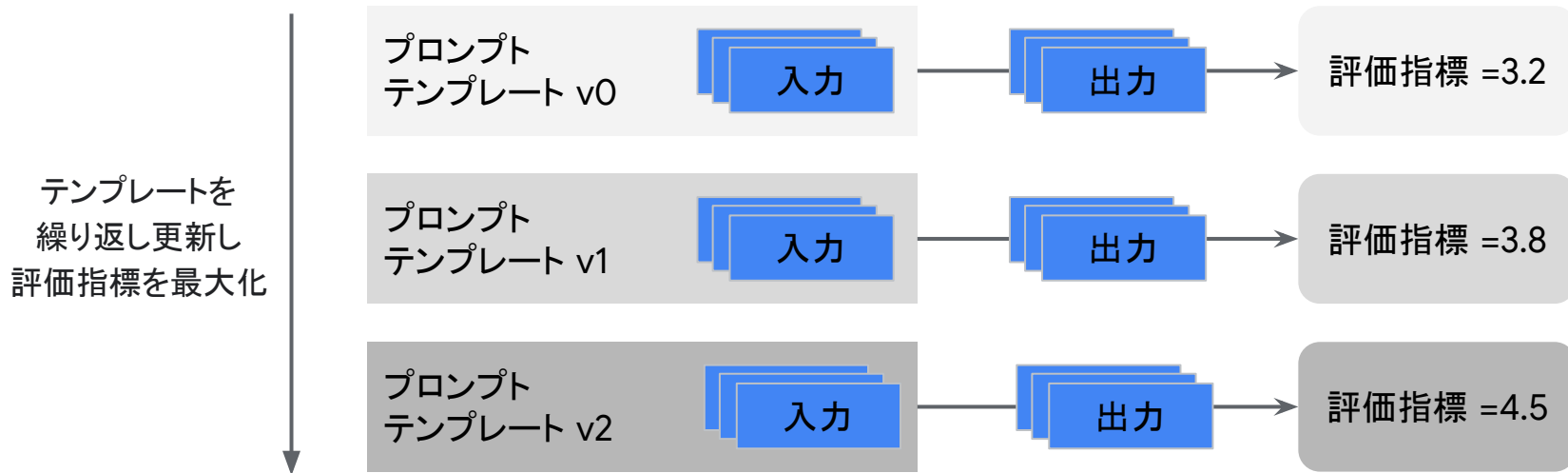
Google Cloud Next Tokyo

026

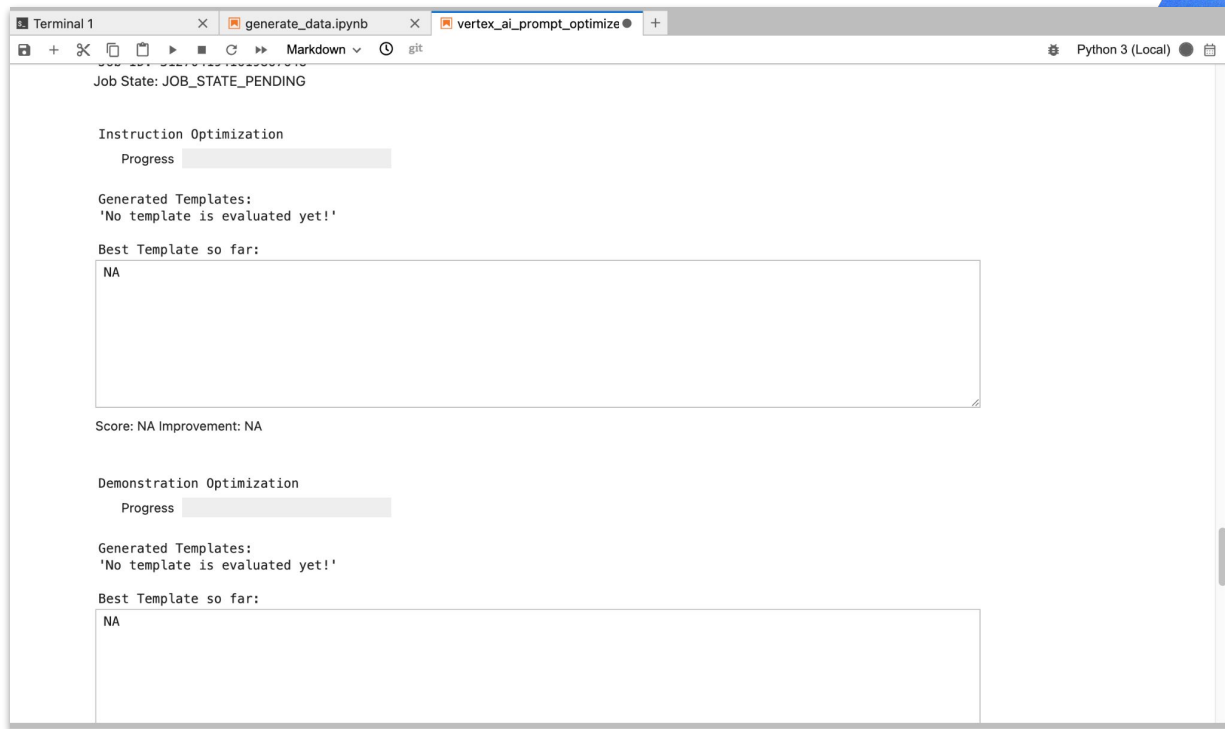
Prompt Optimizer に必要なデータ



Prompt Optimizer の仕組み



Prompt Optimizer の job



```
Terminal 1 | generate_data.ipynb | vertex_ai_prompt_optimizer | +
Markdown | git | Python 3 (Local) |

Job State: JOB_STATE_PENDING

Instruction Optimization
Progress [ ]

Generated Templates:
'No template is evaluated yet!'

Best Template so far:
NA

Score: NA Improvement: NA

Demonstration Optimization
Progress [ ]

Generated Templates:
'No template is evaluated yet!'

Best Template so far:
NA
```

プロンプト テンプレートの最適化

Instruction Optimization

Progress

Generated Templates:

step	prompt	metrics.summarization_quality/mean
0	タスク:\n以下のレビューを簡単にいうと?	4.8
1	タスク:\n以下のレビューを要約してください。クエリ内の「要点」セクションは、生成すべき要約の形式、詳細度、および長さを判断するための参考として活用してください。レビューの重要な情報を網羅しつつ、簡潔にまとめてください。	5.0
1	タスク:\n以下の「レビューテキスト」の内容を要約してください。「要点」は、要約の形式、長さ、および含めるべき情報の粒度の参考として使用してください。	5.0

全に一致するように作成してください。「要点」に記載されている全ての主要な情報項目を、

10	タスク:\n以下のレビューを要約してください。クエリ内の「要点」セクションは、生成すべき要約の形式、詳細度、および長さの**具体的な基準**として活用してください。レビューの重要な情報を網羅しつつ、「要点」セクションが示す情報量、詳細度、および長さに**厳密に合致する**要約を生成してください。	4.8
----	--	-----

Best Template so far:

タスク:

以下のレビューを要約してください。クエリ内の「要点」セクションは、生成すべき要約の形式、詳細度、および長さを判断するための参考として活用してください。レビューの重要な情報を網羅しつつ、簡潔にまとめてください。

Score: 5.0 Improvement: 0.200

LLMOps の一例

プロンプト管理

目的のタスクを解くためにプロンプトをデザインする。

プロンプトの試行錯誤を管理し、実験を適切に記録する。

Vertex AI Studio

プロンプト最適化

新しいモデルに乗り換えることで精度向上を図る。モデルの変化に適応するため、プロンプトを最適化する。

Vertex AI
Prompt Optimizer

モデル評価

モデルの精度を定量的に評価することで、データに基づく意思決定が可能。

Vertex AI
Gen AI Evaluation
Service

自動化

定型化できる操作は、自動化することで運用コストとヒューマンエラーを減す

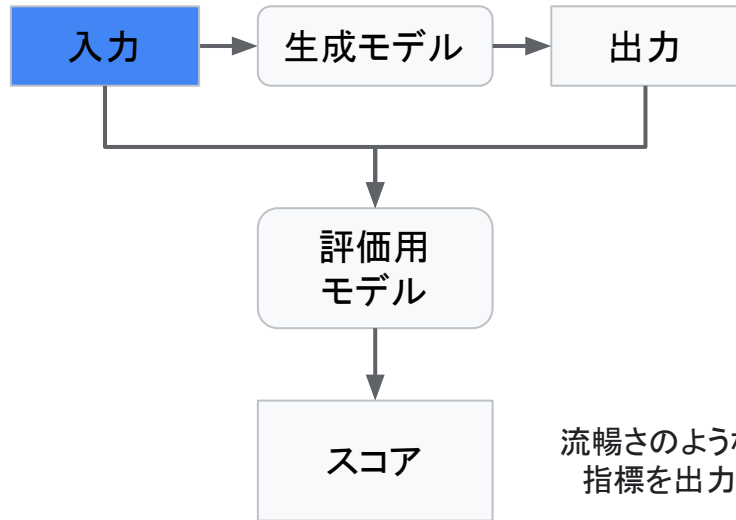
Vertex AI Pipelines

Vertex AI - Gen AI Evaluation Service

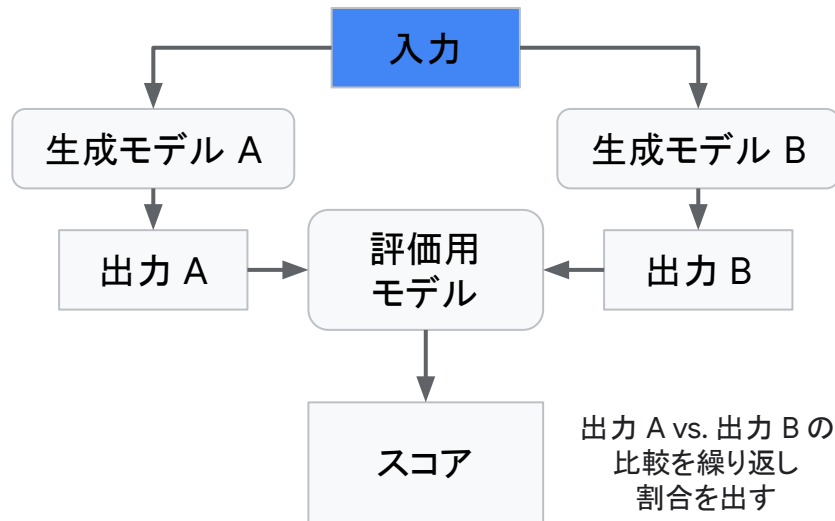
	評価のアプローチ	データ	費用と処理速度
モデルベースの指標	判定モデルを使用し、記述的な評価基準に基づいてパフォーマンスを評価	グラウンド トゥールースはなくてもかまわない	費用がやや高く低速
計算ベースの指標	数式を使用してパフォーマンスを評価	通常、グラウンド トゥールースが必要	費用が低く高速

モデルベースの指標

ポイントワイズ指標



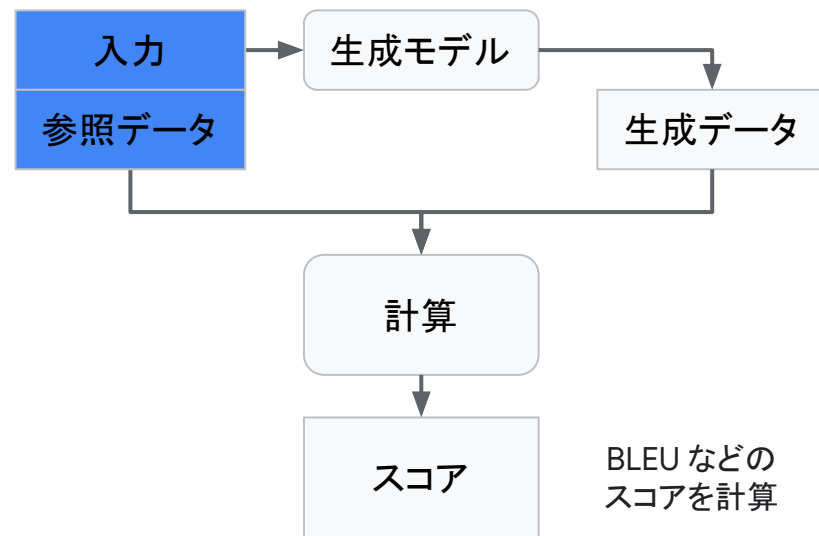
ペアワイズ指標



計算ベースの指標

モデルベースとは異なり、テキストに対して完全一致、BLEUなどのスコアを計算する

モデルベースよりも高速に計算できることが多く計算方法が具体的に定義されている。学术论文などでも度々使われる指標



LLMOps の一例

プロンプト管理

目的のタスクを解くためにプロンプトをデザインする。

プロンプトの試行錯誤を管理し、実験を適切に記録する。

Vertex AI Studio

プロンプト最適化

新しいモデルに乗り換えることで精度向上を図る。モデルの変化に適応するため、プロンプトを最適化する。

Vertex AI
Prompt Optimizer

モデル評価

モデルの精度を定量的に評価することで、データに基づく意思決定が可能。

Vertex AI
Gen AI Evaluation
Service

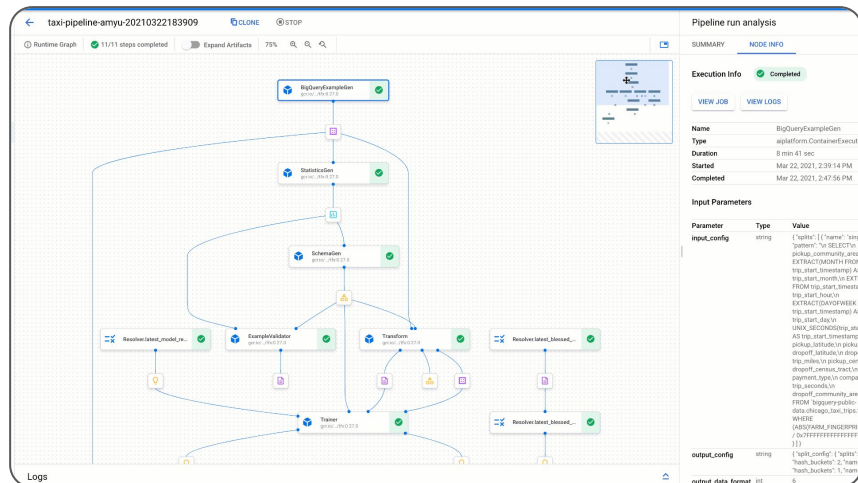
自動化

定型化できる操作は、自動化することで運用コストとヒューマンエラーを減す

Vertex AI Pipelines

様々な作業の自動化

Vertex AI Pipelines



Vertex AI Workbench - Executor

The screenshot shows the Vertex AI Workbench - Executor interface. The top navigation bar includes 'mchrestkha-sandbox' and 'Preview'. The 'Launcher' tab is selected, showing a grid of notebook icons for different environments: Python (Local), PySpark (Local), PySpark on cluster-5977-m, Python 3 on cluster-5977-m, Pytorch (Local), R (Local), and R on cluster-5977-m in us. A 'Create a managed notebook' dialog is open, displaying configuration options: Machine type (n1-standard-4), GPU type (NVIDIA Tesla T4), Number of GPUs (1), and a checkbox for 'Install NVIDIA GPU driver automatically for me'.

最新の生成 AI モデルへの アップデートに必要な LLMOps

プロンプト管理

目的のタスクを解くためにプロンプトをデザインする。

プロンプトの試行錯誤を管理し、実験を適切に記録する。

Vertex AI Studio

プロンプト最適化

新しいモデルに乗り換えることで精度向上を図る。モデルの変化に適応するため、プロンプトを最適化する。

Vertex AI
Prompt Optimizer

モデル評価

モデルの精度を定量的に評価することで、データに基づく意思決定が可能。

Vertex AI
Gen AI Evaluation
Service

自動化

定型化できる操作は、自動化することで運用コストとヒューマンエラーを減す

Vertex AI Pipelines