Google Cloud

Professional Machine Learning Engineer

認定試験ガイド

Professional Machine Learning Engineer 認定試験ガイドの新しいバージョンです。8 月 24 日以降 に日本語で Professional Machine Learning Engineer 認定試験の受験を予定している場合は、こちら の認定試験ガイドを確認ください。8 月 24 日より前に受験を予定している場合は、現在のバージョンをご確認ください

Professional Machine Learning（ML）Engineer は、Google Cloud の機能と一般的な ML アプローチに関する知識を生かして、AI ソリューションを構築、評価、運用化、最適化します。ML エンジニアは、大規模で複雑なデータセットを処理して、再現可能かつ再利用可能なコードを作成します。ML エンジニアは、基盤モデルに基づいて AI ソリューションを設計、運用化します。ML エンジニアは、責任ある AI への取り組みについて検討し、他の職務と密接に連携して、AI ベースのアプリケーションの長期的な成功を確実にします。ML エンジニアには、優れたプログラミング スキルと、データ プラットフォームや分散データ処理ツールに関する経験があります。ML エンジニアは、モデル アーキテクチャ、データと ML のパイプライン作成、MLOps、指標の解釈に精通しています。ML エンジニアは、プロンプト エンジニアリングとコンテキスト エンジニアリング、アプリケーション開発、インフラストラクチャ管理、データ エンジニアリング、データ ガバナンスの基本コンセプトに精通しています。ML エンジニアは、組織全体のチームが AI ソリューションを使用できるようにします。ML エンジニアは、従来の AI モデルと生成 AI（GenAI）モデルのトレーニング、再トレーニング、デプロイ、スケジューリング、チューニング、モニタリング、改善を行い、スケーラブルでパフォーマンスの高いソリューションを設計、作成します。

*注: この試験はコーディングのスキルを直接評価するものではありません。Python と SQL の基礎知識があれば、コード スニペットを含むいかなる質問も読み解くことができるはずです。

セクション 1: ローコード AI ソリューションの構築（試験内容の約 13%）

1.1 Gemini Enterprise Agent Platform で BigQuery ML または AutoML を使用した ML モデルの開発。以下のような点を考察します。

- BigQuery ML または Agent Platform AutoML における、ビジネス課題に応じた各種モデル（分類、回帰、予測、クラスタリングなど）の構築
- BigQuery ML を使用した特徴量エンジニアリングまたは特徴量選択の実行
- BigQuery ML を使用した予測の生成
- Agent Platform AutoML を使用したモデルのトレーニング

- BigQuery を使用した Gemini モデルのファインチューニング

1. 2 Google Cloud AI API または基盤モデルを使用した AI ソリューションの構築。以下のような点を考察します。

- Gemini Enterprise Agent Platform の Model Garden から、特定のタスクに応じたモデルの評価と選択
- 業界固有の API（Document AI API、Vision API、Translate API など）を使用したアプリケーションの構築
- 特定のユースケース向けのソリューションの構築とモデルのチューニング（Gemini、Imagen、Veo、Model Garden のサービスとしてのモデルなど）
- 費用、レイテンシ、可用性を考慮した Gemini ベースのアプリケーションの最適化

セクション **2:** チーム内およびチーム間の連携によるデータとモデルの管理（試験内容の約 **16%**）

2.1 ML 用データの探索と前処理。以下のような点を考察します。

- 効率的なテスト、トレーニング、サービングのための、さまざまなデータタイプ（表形式、テキスト、画像など）の整理と探索
- 規模と複雑さに応じた、適切なデータ前処理ツールの選択（BigQuery [SQL]、Dataflow、Apache Spark、インメモリ Python フレームワークなど）
- Gemini Enterprise Agent Platform Feature Store での特徴量の作成と統合
- データ プライバシーの確保と機密情報（個人を特定できる情報 [PII] など）の取り扱い

2.2 ノートブック（Gemini Enterprise Agent Platform Workbench、Colab Enterprise など）を使用したモデルのプロトタイピング。以下のような点を考察します。

- ノートブック環境の構築と実行時におけるコラボレーションとセキュリティのベスト プラクティスの適用
- Agent Platform Workbench または Colab Enterprise ノートブックでの一般的なフレームワーク（PyTorch、sklearn、JAX など）を使用したモデル開発
- ノートブック環境における、Model Garden のさまざまな基盤モデルとオープンソースモデルを使用したプロトタイプモデルの作成

2.3 ML テストのトラッキングと実行。以下のような点を考察します。

- 開発とテストに適した Google Cloud 環境（Gemini Enterprise Agent Platform の Experiments、Gemini Enterprise Agent Platform Pipelines、Kubeflow Pipelines など）をフレームワークに応じて選択
- 予測 AI ソリューションと生成 AI ソリューションの評価（モデル評価指標、LLM-as-a-Judge など）
- モデルのアーティファクト、バージョン、リネージの追跡と比較（Agent Platform の Experiments、Gemini Enterprise Agent Platform ML Metadata など）

セクション **3:** プロトタイプの **ML** モデルへのスケーリング（試験内容の約 **21%**）

3.1 費用、複雑さ、レイテンシ、スケーラビリティを考慮し、タスクに応じたモデルの構築。以下のような点を考察します。

- モデルタイプの選択（ARIMA、DNN、LLM など）
- プロダクトの選択（Agent Platform AutoML、BigQuery ML、Agent Platform Pipelines など）
- デプロイ戦略の選択
- 解釈可能性要件のあるモデル手法

3.2 モデルのトレーニング。以下のような点を考察します。

- Google Cloud（Cloud Storage、BigQuery など）上のトレーニング データ（表形式、テキスト、音声、画像、動画など）の整理
- さまざまなソースからの構造化データと非構造化データのトレーニング パイプラインへの取り込み
- さまざまなソフトウェア開発キット（SDK）を使用したモデルのトレーニング（Agent Platform カスタム トレーニング、Google Kubernetes Engine（GKE）上の Kubeflow、Agent Platform AutoML、表形式ワークフローなど）と、Google Cloud でのトレーニングの整理
- ML モデルのトレーニング失敗のトラブルシューティング
- ハイパーパラメータ チューニング
- Agent Platform や Model Garden の基盤モデルのファインチューニングと、チューニングを検討すべきタイミング

3.3 トレーニングに適したハードウェアの選択。以下のような点を考察します。

- コンピューティング オプションとアクセラレータ オプションの評価（CPU、GPU、TPU など）
- データ並列処理戦略とモデル並列処理戦略を用いた、GPU と TPU の分散トレーニングにおけるオプションの把握

セクション **4:** モデルのサービングとスケーリング（試験内容の約 **20%**）

4.1 モデルのサービング。以下のような点を考察します。

- 適切なサービス（Agent Platform、Model Garden、Cloud Run、GKE など）を使用した、バッチ推論とオンライン推論用のモデルのデプロイ
- ビルド済みコンテナとカスタム コンテナを使用した、さまざまなフレームワーク（PyTorch、XGBoost など）のモデルのパッケージ化とサービング
- Gemini Enterprise Agent Platform Model Registry でのモデルの整理とバージョン管理
- モデル バージョンを比較するためのモデル ロールアウト戦略（A/B テストやカナリア デプロイなど）の実装
- 推論の前処理と後処理のためのソリューションの開発

4.2 オンライン モデル サービングのスケーリング。以下のような点を考察します。

- Agent Platform Feature Store を使用した特徴の管理とサービング
- パブリック エンドポイントとプライベート エンドポイントへのモデルのデプロイ
- 適切なハードウェア（CPU、GPU、TPU、エッジなど）の選択
- スループットに基づいたサービング バックエンドのスケーリング（Gemini Enterprise Agent Platform Inference、コンテナ化されたサービングなど）
- 本番環境でのトレーニングとサービングのための ML モデルのチューニング

セクション **5: ML** パイプラインの自動化とオーケストレーション（試験内容の約 **18%**）

5.1 エンドツーエンドの ML パイプラインの開発。以下のような点を考察します。

- データとモデルの検証
- マネージド サービスまたは非マネージド サービスを使用し、テンプレートまたはカスタム ソリューションから行うパイプラインの構築とオーケストレーション（Agent Platform Pipelines、Managed Service for Apache Airflow、Ray on Gemini Enterprise Agent Platform など）
- トレーニングとサービングの間で一貫したデータ前処理の保証

5.2 モデル再トレーニングの自動化。以下のような点を考察します。

- 適切な再トレーニング ポリシーの決定

- 継続的インテグレーション、継続的デリバリー、継続的トレーニング（CI / CD / CT）パイプライン（Cloud Build など）を使用したモデルのデプロイ

セクション **6: AI** ソリューションのモニタリング（試験内容の約 **13%**）

6.1 AI ソリューションに対するリスクの特定。以下のような点を考察します。

- 適切なセキュリティ ツール（正規表現、安全フィルタ、Model Armor など）を使用し、データやモデルの意図しない悪用や漏洩（データの引き出し、悪意のあるプロンプト、LLM とのセンシティブ データの共有など）を防ぐ、安全な AI システムの構築
- 責任ある AI への取り組みへの準拠（バイアスのモニタリングなど）
- Agent Platform におけるモデルの説明可能性（Agent Platform Inference など）

6.2 AI ソリューションのモニタリング、テスト、トラブルシューティング。以下のような点を考察します。

- Gemini Enterprise Agent Platform における Model Monitoring の設定と活用による、本番環境モデルの継続評価指標の確立
- 一般的な問題（トレーニング サービング スキュー、データドリフト、コンセプト ドリフト、特徴アトリビューションのドリフトなど）のモニタリング
- 生成 AI ソリューションのモニタリング、テスト、評価