AIお役立ち情報/基本情報

2024-05-20

Gemini 1.5 Flashとは？その機能や使い方、料金体系を徹底解説

この記事のポイント

- 最大100万トークンの長いコンテキストウィンドウを持ち、マルチモーダルデータを効率処理
- 翻訳、コーディング、推論など様々なタスクで高いベンチマーク結果
- 同等性能を10分の1のコストで利用可能なコストパフォーマンスの高さ
- 倫理的な使用とセキュリティ、プライバシー保護のための措置を重視

監修者プロフィール
坂本 将磨

Microsoft AIパートナー、LinkX Japan代表。東京工業大学大学院で技術経営修士取得、研究領域：自然言語処理、金融工学。NHK放送技術研究所でAI、ブロックチェーン研究に従事。学会発表、国際ジャーナル投稿、経営情報学会全国研究発表大会にて優秀賞受賞。シンガポールでのIT、Web3事業の創業と経営を経て、LinkX Japan株式会社を創業。

Geminiの最新モデル「Gemini Flash」は、速度と効率性を重視した軽量・高性能モデルです。
最大100万トークンの長いコンテキストウィンドウを持ち、画像や音声などマルチモーダルデータの統合的処理が可能。

さらに、翻訳やコーディングなど様々なタスクで高い性能を示しながら、**従来モデルの10分の1のコスト**で利用できるコストパフォーマンスの高さも魅力です。

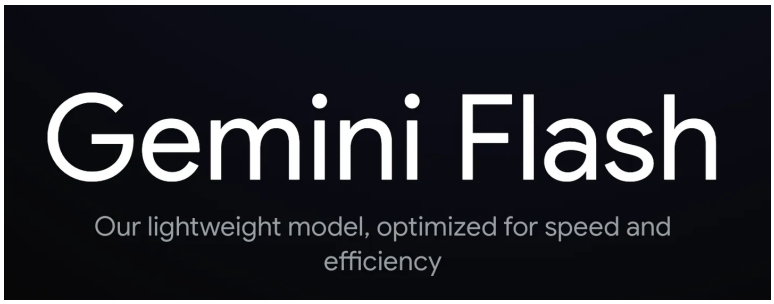
本記事では、その特徴や使い方、価格体系、活用例を詳しく解説します。

ピックアップ

新着記事

目次

Gemini Flashとは？



Gemini Flashは、Googleが開発したGeminiモデルファミリーの一部であり、特に速度と効率に重点を置いて設計されたモデルです。

Geminiモデルファミリーには、1.5 ProやNanoなどの他のモデルも含まれ、それぞれが異なる用途やニーズに応じた性能を提供しています。

Gemini Flashは、これらの中でも特に軽量でありながら、高性能な処理能力を持つモデルとして設計されています。

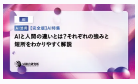
Gemini Flashの主な特徴


Adobe Fireflyとは？主要機能や使い方、料金体系を…
2024-05-20


Gemini 1.5 Flashとは？その機能や使い方、料金体…
2024-05-20


ChatGPTとは？その機能や日本語での使い方を徹底…
2024-05-20

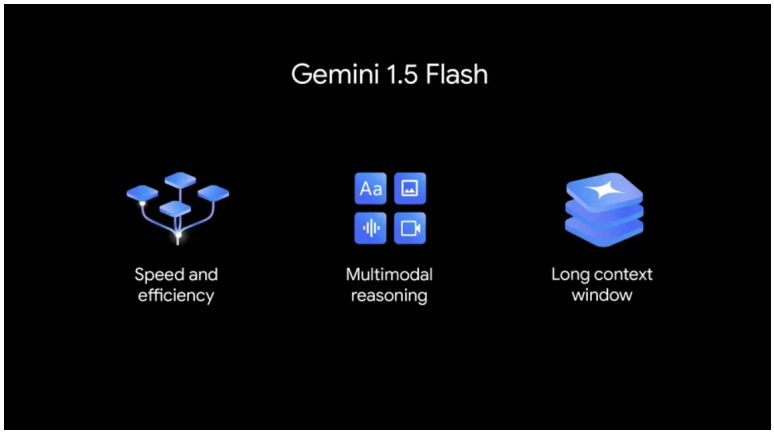

チャットボットを予約システムに活用！そのステッ…
2024-05-20


AIと人間の違いとは？それぞれの強みと短所をわか…
2024-05-20

- Gemini Flashとは？
- Gemini Flashの主な特徴
- 100万トークンのコンテキストウィンドウ
- 軽量かつ高性能なアーキテクチャ
- 優れたベンチマーク結果
- コストパフォーマンス
- Gemini Flashの料金体系
- Gemini Flashの使い方
- Google AI Stuioの場合
- Vertex AIの場合
- Gemini Flashの活用例
- 幅広い業界での応用が可能
- 開発者向けの統合
- コンプライアンスとセキュリティ
- 責任ある使用に向けた設計
- セキュリティ
- まとめ

Gemini Flashは、長いコンテキストウィンドウ、軽量・高性能なアーキテクチャ、優れたベンチマーク結果など、優れた特徴を備えています。

以下では、それぞれについて詳しく解説します。



100万トークンのコンテキストウィンドウ

Gemini Flashの主要な特徴の一つは、**最大100万トークンの長いコンテキストウィンドウを持つこと**です。これにより、長時間の動画や大規模なコードベースなど、膨大なデータの処理が可能となります。

また、画像、音声、テキストなど多様なデータを統合的に処理するマルチモーダル能力も持ち合わせています。

この効率化は、Gemini 1.5 Proから知識を抽出する「**ディスティレーション**」技術を通じて実現されており、速度とコスト効率の両方を大幅に向上させています。

> ディスティレーションとは、大規模なモデルから最も重要な知識とスキルを抽出し、より小さく効率的なモデルに転送するプロセスのことです。

軽量かつ高性能なアーキテクチャ

Gemini Flashは、Gemini 1.5 Proからのディスティレーションプロセスにより、**重要な知識とスキルが小型モデルに転送**されています。
このプロセスにより、軽量でありながら高性能なモデルが実現されています。

また、1.5 Proと比べても、ほとんどの開発者やエンタープライズのユースケースにおいて、**平均初回トークンレイテンシーがサブセカンド(1秒未満)である**など、高速な応答が可能です。

優れたベンチマーク結果

Gemini Flashは、翻訳、コーディング、推論など各種タスクにおいて優れたベンチマーク結果を示しています。

カ	ベンチマーク	形容	ジェミニ1.0プロ	ジェミニ1.0ウルトラ	ジェミニ1.5プロ (2024年2月現在)	ジェミニ1.5フラッシュ	ジェミニ1.5プロ (2024年5月現在)
全般	MMLUの	57科目(STEM、人文科学などを含む)の質問の表現	71.8%	83.7%	81.9%	78.9%	85.9%
コード	ナチュラル2コード	Python コード生成。HumanEvalのようなデータセットを差し出し、Web上ではリークしていません	69.6%	74.9%	77.7%	77.2%	82.6%
数学	数学	挑戦的な数学の問題(代数、幾何学、微積分学などを含む)	32.6%	53.2%	58.5%	54.9%	67.7%
推論	GPQA(メイン)	生物学、物理学、化学の各分野の専門家が書いた挑戦的な質問のデータセット	27.9%	35.7%	41.5%	39.5%	46.2%
	ビッグベンチハード	多段階の推論を必要とする多様な困難なタスク	75.0%	83.6%	84.0%	85.5%	89.2%
多言語	WMT23型	言語翻訳	71.7	74.4	75.2	74.1	75.3
画像	MMMUの	学際的な大学レベルの推論問題	47.9%	59.4%	58.5%	56.1%	62.2%
	MathVistaの	視覚的文脈における数学的推論	46.6%	53.0%	54.7%	58.4%	63.9%
オーディオ	FLEURS(55言語)	自動音声認識(単語エラー率に基づき、低いほど良い)	6.4%	6.0%	6.6%	9.8%	6.5%
ビデオ	エゴスキーマ	ビデオによる質問応答	55.7%	61.5%	65.1%	65.7%	72.2%

Geminiファミリーのパフォーマンス比較

例えば、PythonコードのHumanEval-likeデータセットでは、Gemini 1.5 Proが77.7%の精度を達成しているのに対し、Gemini Flashは77.2%とほぼ同等の精度を維持しています。

また、物理や化学、生物学の専門家が作成した質問に回答するGPQAタスクでは、Gemini 1.5 Proが41.5%の正答率であるのに対し、Gemini Flashは39.5%と、わずかに劣るものの高い精度を示しています。

コストパフォーマンス

Gemini Flashは、従来モデルのGemini Proと同等の性能を持ちながら、**10分の1のコストで利用可能**という点も大きな特徴です。

「Gemini 1.5 Flash」と「Gemini 1.5 Pro」の料金比較表です。

項目	Gemini 1.5 Flash	Gemini 1.5 Pro
入力価格（100万トークンあたり）	$0.35（128,000トークンまで） $0.70（128,000以降）	$3.50（128,000トークンまで） $7.00（128,000以降）
出力価格（100万トークンあたり）	$1.05（128,000トークンまで） $2.10（128,000以降）	$10.50（128,000トークンまで） $21.00（128,000以降）

コストを抑えつつ高性能なAI機能を導入したい企業などに最適なソリューションと言えます。

Gemini Flashの料金体系

Gemini Flashの料金体系は、以下の2つのプランに分かれています。

項目	無料プラン	従量課金制（ボドルの料金）
レート制限	- 15 RPM（1分あたりのリクエスト数）	- 360 RPM（1分あたりのリクエスト数）
	- 100万TPM（1分あたりのトークン数）	- 1,000万TPM（1分あたりのトークン数）
	- 1,500 RPD（1日あたりのリクエスト数）	- 10,000 RPD（1日あたりのリクエスト数）
価格（入力）	無料	- 100万トークンあたり$0.35（最大128,000トークンまでのプロンプト）
		- 100万トークンあたり$0.70（128,000を超えるプロンプトの場合）
コンテキストキャッシュ	近日提供予定、現在は該当なし	近日提供予定、詳細な料金は該当なし
価格（出力）	無料	- 100万トークンあたり$1.05（最大128,000トークンまでのプロンプト）
		- 100万トークンあたり$2.10（128,000を超えるプロンプトの場合）
プロンプト/回答の利用	○ (Googleサービスの改善のために利用)	×

Gemini Flashの使い方

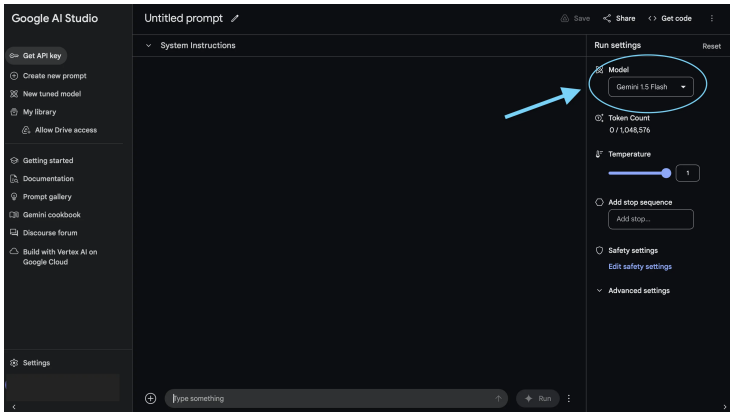
Gemini FlashはGoogle AI Studio、もしくはVertex AIで利用可能です。

Google AI Stuioの場合

1. Google AI Studioのページにアクセスします。

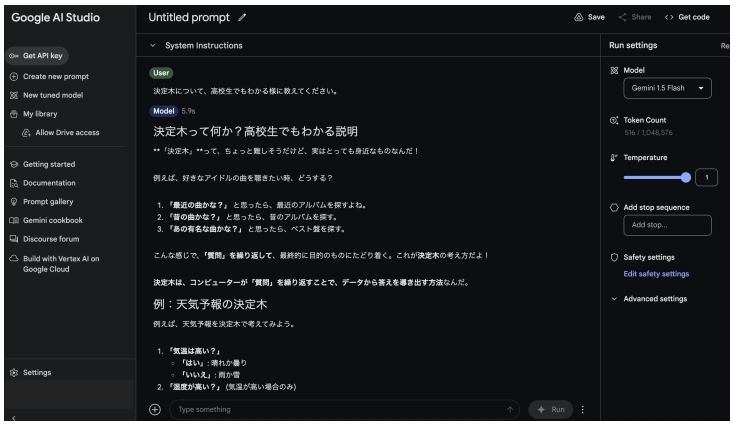
2. 「利用規約の確認」にチェックを入れ、*続行*を選択します。

3. 「モデルの選択(画像の矢印)」から、**Gemini Flash**を選択すれば、準備完了です。



モデル選択

4. 下のチャット欄に文章を入力する事で、Gemini FLashとの会話が可能です。



Vertex AIの場合

1.Google Cloudのアクセスし、アカウントを作成します。

> 新規ユーザー向けに、300クレジット分の無料トライアルが提供されています。

2.Vertex AIを有効にする

Vertex AIは、AIモデルの開発、トレーニング、デプロイを支援するサービスです。

3.Gemini Flash APIを有効にする

これは、Vertex AIコンソールのメニューから簡単に行うことができます。

4.APIリクエストを送信する

Gemini Flashを使用するには、APIリクエストを送信する必要があります。APIリクエストには、処理したいデータや必要な機能などを指定します。

5.レスポンスを処理する

APIリクエストを送信すると、Gemini Flashからレスポンスが返されます。レスポンスには、処理結果やその他の情報が含まれています。

以下は、Gemini Flashを使用してテキストを要約するためのサンプルコードです。

```
# ライブラリのインポート
import requests

# APIキーの設定
API_KEY = "YOUR_API_KEY"

# リクエストの作成
url = "https://vertex.ai/v1/projects/{project_id}/location
s/{location}/models/{model_id}"
data = {
  "inputs": [
    {
      "text": "This is a long piece of text that I want to
summarize."
    }
  ]
}
headers = {
  "Authorization": "Bearer " + API_KEY,
  "Content-Type": "application/json"
}

# リクエストの送信
response = requests.post(url, json=data, headers=headers)

# レスポンスの処理
if response.status_code == 200:
  summary = response.json()["outputs"][0]["text"]
  print(summary)
else:
  print("Error:", response.status_code)
```

このコードを実行すると、summary変数に要約されたテキストが格納されます。

Gemini Flashの活用例

Gemini Flashは、その優れた性能と柔軟性により、様々な業界や用途で活用できます。

特に、大量のデータを高速かつ効率的に処理する必要があるタスクに適しています。

幅広い業界での応用が可能

Gemini Flashは、大量かつ高頻度のタスクに最適です。
例えば、要約、チャットアプリケーション、画像および動画キャプション、長いドキュメントや表からのデータ抽出など、多様な応用が可能です。

特に、その長いコンテキストウィンドウとマルチモーダル推論の能力により、これらのタスクを迅速かつ効率的に処理することができます。

具体的な使用例としては、**1時間の動画、11時間の音声、3万行以上のコードベース、70万語以上の文章**など、膨大なデータを処理することが可能です。
これにより、幅広い業界や用途において活用できます。

開発者向けの統合

Gemini Flashは、主にGoogle AI StudioおよびVertex AIを通じて使用できます。開発者はAPIを利用して、簡単にプロジェクトに統合することができます。

また、データサイエンティストや機械学習エンジニア向けには、Vertex AIが提供する専用ツールも用意されています。
これらのツールを使用することで、より高度なカスタマイズや最適化が可能となります。

コンプライアンスとセキュリティ

Gemini Flashは、AIモデルの倫理的な使用と安全性を重視しています。

以下では、責任ある使用に向けた設計とセキュリティ対策について詳しく説明します。

責任ある使用に向けた設計

AI技術の展開において、倫理的な使用が重要です。Gemini Flashは、様々なアプリケーションにおいて倫理的に展開されるよう設計されています。

また、AIモデルのバイアスと公平性に対する対策も講じられています。

具体的には、トレーニングデータの多様性を確保し、**不適切なバイアスを排除するための手法**が取り入れられています。また、モデルの出力をモニタリングし、問題が発生した場合には速やかに対処できる体制が整えられています。

セキュリティ

Gemini Flashは、**安全な使用のための保護措置とプロトコル**が整備されています。これにより、ユーザーデータのプライバシーが守られ、安全に利用することができます。

具体的には、データの暗号化、アクセス制御、監査ログの記録など、多層的なセキュリティ対策が講じられています。
また、個人情報の取り扱いに関しては、厳格なポリシーと手順が定められており、適切に管理されています。

まとめ

本記事では、Googleが開発した軽量で高性能なAIモデル「Gemini Flash」について詳しく解説しました。Gemini Flashは、100万トークンの長いコンテキストウィンドウ、マルチモーダル推論、優れたパフォーマンスなどの特徴を持ち、要約やチャットアプリ、画像・動画キャプション、データ抽出など幅広い用途で活用できます。

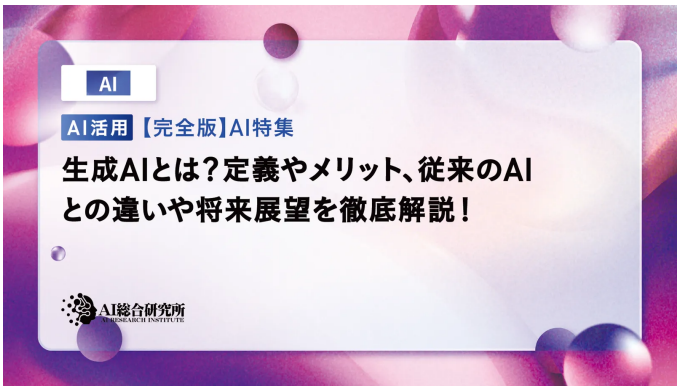
倫理的な側面にも配慮され、責任ある使用やセキュリティ、プライバシー保護のための措置が講じられています。
今後も更なる改良と拡張が予定されており、Gemini Flashは AI技術の発展に欠かせない存在として、私たちの生活やビジネスに大きな影響を与え続けるでしょう。

監修者

坂本 将磨

Microsoft AIパートナー、LinkX Japan代表。東京工業大学大学院で技術経営修士取得、研究領域：自然言語処理、金融工学。NHK放送技術研究所でAI、ブロックチェーン研究に従事。学会発表、国際ジャーナル投稿、経営情報学会全国研究発表大会にて優秀賞受賞。シンガポールでのIT、Web3事業の創業と経営を経て、LinkX Japan株式会社を創業。

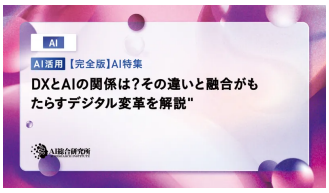
関連記事



AIお役立ち情報/使い方

生成AIとは？定義やメリット、従来のAIとの違いや将来展望を徹底解説！

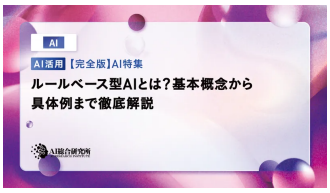
2024-05-18



AIお役立ち情報/基本情報

DXとAIの関係は？その違いと融合がもたらすデジタル変革を解説

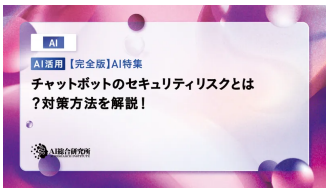
2024-05-19



AIお役立ち情報/使い方

ルールベース型AIとは？基本概念から具体例まで徹底解説

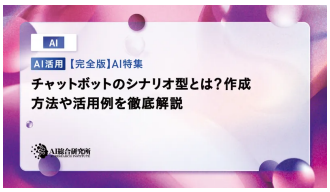
2024-04-20



AIお役立ち情報/使い方

チャットボットのセキュリティリスクとは？対策方法を解説！

2024-05-06

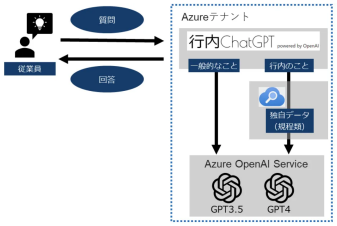


AIお役立ち情報/使い方

チャットボットのシナリオ型とは？作成方法や活用例を徹底解説

2024-05-06

あなたにおすすめの事例

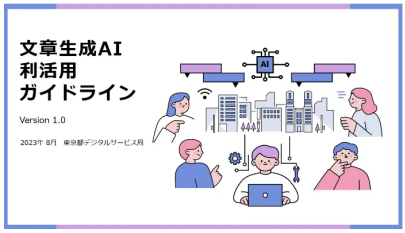


金融・保険

横浜銀行と東日本銀行による自動生成AI「行内ChatGPT」の導入で生産性を飛躍的に向上

IT・システム開発

Google Pixel 8登場: 最新Tensor G3搭載と画期的な「音声消しゴムマジック」機能



全般

東京都、職員向け「文章生成AI利活用ガイドライン」を策定

もっと見る

業種別にみるAI活用事例