Enterprise Strategy Group by TechTarget

経済性に関するホワイト ペーパー

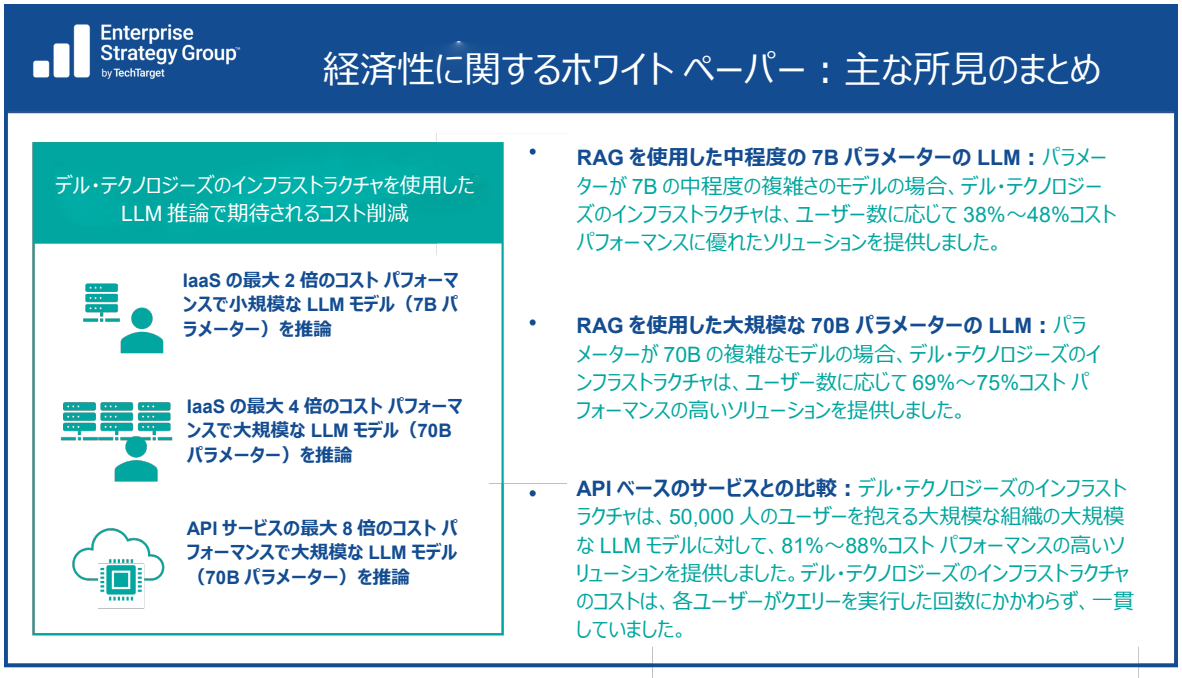
大規模言語モデルの推論にかかる総コストを理解する

デル・テクノロジーズのオンプレミス ソリューションの活用により、RAG を使用した LLM 推論のコスト パフォーマンスを、パブリッククラウドまたはトークンベースの API と比較して 38%～88%向上させる方法

著者：Enterprise Strategy Group、Practice Director 兼 Principal Validation Analyst、Aviv Kaufmann

2024 年 4 月

目次



はじめに

この経済性に関するホワイト ペーパーでは、テキストベースの生成AI (GenAI)機能を組織に提供するためのオプションと考慮事項の一部について説明します。TechTargetのEnterprise Strategy Groupは、大規模言語モデル(LLM)の推論について、オンプレミスのデル・テクノロジーズ インフラストラクチャで検索拡張生成(RAG)を使用した場合と、APIを介したネイティブ パブリッククラウド インフラストラクチャ アズ ア サービス(IaaS)またはOpenAI GPT-4 Turbo LLMモデル サービスを使用した場合で、予測されるコストをモデル化して比較しました。デル・テクノロジーズが、IaaSよりも最大4倍、GPT-4 Turbo APIより最大8倍のコスト パフォーマンスでLLM推論を提供できることがわかりました。

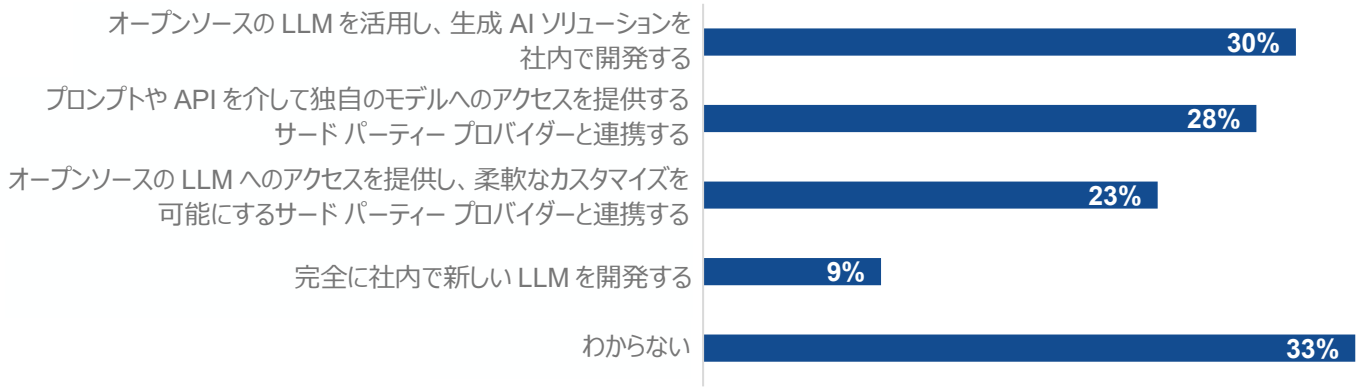
課題

組織は、企業固有のデータやその他の知的財産を活用する生成AIやLLMの機能を導入して、コンテンツ生成を自動化したり、質問に回答したり、意思決定者がすぐにインサイトを利用できるようにしたりしています。Enterprise Strategy Groupの調査によると、組織で生成AIを使用することの主なメリットとして、プロセスとワークフローの改善や自動化、データ分析とビジネス インテリジェンスのサポート、従業員の生産性の向上、運用効率の向上などが挙げられています。その他にも、多くのメリットについて回答が寄せられています。[1]

LLMの開発はコストがかかり、複雑になる可能性がありますが、組織はニーズに合わせて既存のオープンソースLLMを簡単に強化、微調整、カスタマイズできます。OpenAI GPTなどの既製のAPIベースのサービスは、よりシンプルなソリューションを提供しますが、特に大規模な組織やより複雑なLLMでは、推論（クエリー）のコストが急増する可能性があります。また、組織は、強力なGPU対応エンタープライズ サーバーまたは同等のGPU対応クラウド インスタンス、およびオープンソースLLMを実行するNVIDIAのAI Enterpriseなどの機械学習プラットフォーム上に、独自のLLM推論ソリューションを構築して制御することもできます。意外な結果ではありませんが、Enterprise Strategy Groupの調査により、組織がLLMでサポートされる生成AIを開発および使用するうえで最も一般的な戦略は、オープンソースのLLMを活用し、生成AIソリューションを社内で開発することであることがわかりました。[2]

図1：ほとんどの組織は独自の生成AIソリューションを社内で開発することを計画

あなたの組織では、大規模言語モデル(LLM)でサポートされる生成 AI をどのように開発または使用しますか？（回答者の割合、N=670、複数回答可）



出典：*Enterprise Strategy Group（TechTarget, Inc.の部門）*

[1]出典：Enterprise Strategy Group の調査レポート『*Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*』（2023 年 8 月）。

[2]同上。

LLM推論の主な考慮事項

テキストベースのLLMは、特定の業界、ユース ケース、および組織に合わせて調整可能なテキストベースのコンテンツ、回答、要約、および質問の学習、理解、および作成に重点を置いています。RAGは、追加のソースから取得したカスタム データで生成AIモデルの結果を補強し、モデルの精度を高めます。これらは企業で最も導入されているLLMであり、他の多くのユース ケースに加えて、チャットボット、Q&Aアシスタント、プロセスの改善と自動化に使用できます。また、カスタム ツールやアプリケーションに組み込まれた機能としても使用できます。LLMモデルを提供する場合、組織はトレーニング（効果的なモデルを構築するために必要なデータ集約型およびコンピューティング集約型の分析）、推論（トレーニング済みモデルでのユーザー インタラクションのサービス）、および微調整（モデルの継続的な更新と最適化）のためのインフラストラクチャを考慮する必要があります。このレポートでは、推論ワークロードを容易にするために必要なインフラストラクチャに焦点を当てています。LLMの推論に使用可能な導入方法はいくつかあります。

- **従来のインフラストラクチャ：**コンピューティング、メモリー、GPU、ストレージで構成される従来のインフラストラクチャを購入またはリースして、商用またはオープンソースのAIプラットフォームとともに導入および管理できるため、組織は導入のあらゆる側面を制御できます。この方法は、大規模で予測可能なワークロードに対して最もコスト パフォーマンスが高い可能性があります。
- **パブリッククラウドIaaS：**同様に、組織は、商用またはオープンソースのAIプラットフォームとともに、GPUとストレージを備えた同等のクラウド コンピューティング インスタンスを導入できます。この方法を使用すると、俊敏性が得られ、既存のツールと容易に統合できるため、プラットフォーム上で同様の制御が可能になります。この方法は、小規模な導入や、予測不能な要件または季節的な要件がある場合に最もコスト パフォーマンスが優れている可能性があります。
- **LLM APIサービス：**OpenAI GPTなどの確立されたサービスを使用すると、インフラストラクチャやAIプラットフォームを管理することなく、機能を迅速に提供できます。この方法は、調査時や開始時、小規模な導入、および大規模なカスタマイズや制御を必要としない導入に最適です。

LLMプラットフォームを決定する前に、組織は時間をかけて要件と機能を理解し、LLM推論用のプラットフォームの選択に関する次のような考慮事項について話し合う必要があります。

- **コスト/ROI：**組織は、すべてのテクノロジー投資の実装と使用に関するコストとメリットを考慮する必要があります。Enterprise Strategy Groupの調査によると、組織がAIイニシアティブの有効性を測定するために使用する最も一般的な指標はコスト削減とROIでした。[3]
- **パフォーマンスと拡張性：**プロセッサー、GPU、メモリー、ストレージに十分なリソースを確保してインフラストラクチャをサイジングすることは、通常の負荷とピーク負荷で予想される推論の同時実行処理を処理し、平均推論レイテンシーを短縮してユーザーにポジティブな体験を提供できるようにするために重要です。また組織は、推論プラットフォームに移行する前に、LLMのコンピューティング負荷の高いトレーニングを同じプラットフォームで行うか、より高性能な専用トレーニング プラットフォームで行うかを判断する必要もあります。
- **シンプルな管理：**オンプレミス インフラストラクチャをクラウド インフラストラクチャやサービスと比較する場合、組織は社内の能力を考慮し、インフラストラクチャとプラットフォームの運用コスト（管理、サポートとメンテナンス、電源および冷却など）を理解することが重要です。また、コロケーション オプションにより、組織は独自のデータセンターでホスティングするメリットの多くを得ながら、インフラストラクチャとプラットフォームの運用に必要なリソースとスキルの負担を軽減することができます。
- **ユーザーのワークロードの予測：**ツールにアクセスするユーザー数と、ユーザーが1日あたりに質問する頻度を理解して予測することは、ソリューションを選択する際に考慮すべき重要な指標です。需要が小さい場合はAPIサービスで十分かもしれませんが、組織がサポートするユーザーと推論が増えるにつれて、独自のプラットフォームを構築する方がコスト パフォーマンスが高まります。組織は、インフラストラクチャを適切にサイジングし、ビジネスのニーズに合わせて拡張できるように、時間の経過に伴う導入と使用頻度の増加予測を考慮することが重要です。

[3]出典：Enterprise Strategy Groupt 調査レポート『*Navigating the Evolving AI Infrastructure Landscape*』（2023 年 9 月）。

- **データ ガバナンス：**組織は、モデルのトレーニングと保守に必要なデータ ソースの場所とデータ ガバナンスの要件を考慮する必要があります。ハイブリッドクラウド インフラストラクチャは、データがローカルに存在するか、必要な場所で簡単に取得できる場合に最も効果的ですが、パブリッククラウドを使用すると、データの収集と一元化が容易になる場合があります。また、オンプレミス インスタンスを使用すると、組織はセキュリティをより適切に制御し、機密データのコンプライアンスを確保できます。最新かつ包括的で偏りのないデータをトレーニングして維持することで、一層優れたLLMが構築され、これまで以上に正確なインサイトが推論から得られます。

Enterprise Strategy Groupの経済性に関する分析

Enterprise Strategy Groupは、さまざまな複雑さのRAGを利用する複数のオープンソースLLM（パラメーターの数は7Bや70Bなど）と、さまざまな規模の組織（ユーザー数5,000から50,000）の推論を提供するうえで予測されるコストを比較する、経済性の分析を作成しました。このモデルは社内向けのテキストベースのQ&Aを提供しており、推論はデータが配置されている場所で発生するため、データ移行のコストは高くないと推定しました。この分析では、インフラストラクチャの提供と実行、システムの管理、必要に応じたクラウド サービスの支払いなど、3年間にわたるモデルの実行と推論に関連するすべてのコストを調べました。

デル・テクノロジーズのオンプレミス インフラストラクチャとパブリッククラウドIaaSの比較

当社のモデルでは、まず、従来のインフラストラクチャ（オンプレミス、コロケーション環境、エッジ ロケーションなど）でLLM推論を実行する場合と、Amazon EC2インスタンス上の同様に構成されたパブリッククラウドIaaSで実行した場合の予測コストを比較しました。推論ノード サーバーとNVIDIA H100 GPUの構成要件は、推論ベースライン テストの結果に基づいてワークロードごとにサイジングされました。これにより、通常の負荷とピーク負荷（最大要求数とモデル インスタンス数に基づく）において同時実行要件が処理可能になり、また予測されるワークロードごとに適切なレイテンシーとスループットが提供されるようにしました。次に、デル・テクノロジーズのインフラストラクチャと同等のEC2構成の両方について、表1に記載されている各コストをモデル化しました。

表1.LLM推論ワークロードの要件ごとにモデル化されたコストと前提条件

コスト カテゴリー	デル・テクノロジーズ（オンプレミス）	パブリッククラウドIaaS (Amazon EC2)
初期の取得コスト（ハードウェアおよびソフトウェア）	Dell PowerEdge R760xaおよびR660（ProDeployおよびProSupport付き）についてデル・テクノロジーズが提示する価格	該当なし
追加資本コスト（利息）および減価償却費（利益）	モデルに組み込み（WACC 8%、年間減価償却利益6%）	該当なし
電力コストと冷却コスト	システムの仕様に基づいて算出（0.173ドル/kWh）	該当なし
毎月のクラウド支出	該当なし	p5.48xlarge EC2インスタンスのコストは3年間の予約割引に基づいて算出
NVIDIA AI Enterpriseライセンス/GPU	5年間のライセンスに基づく（比例配分）	インスタンスごと/時間（16時間/日、週5日に基づいてコストを制限）
インフラストラクチャ/インスタンス管理	モデル化（ノード数に基づくシステム管理者の10%～100%）	オンプレミス モデルよりも66%低い
MLモデルとプラットフォーム管理	モデル化（モデル インスタンスの数に基づくMLエンジニアの20%～100%）	オンプレミス モデルと同じ

出典：*Enterprise Strategy Group（TechTarget, Inc.の部門）*

小型モデル：7BパラメーターのMistral 7B LLM

最初の比較では、オープンソースのMistral 7B LLMと同様に、約70億個のパラメーターを含む小規模なモデルを提供するためのコストをモデル化しました。要件をサイジングするために、要求あたり平均約0.4秒のレイテンシー、推論1秒あたりの推定スループット2.29～6.86を提供可能なサーバーとGPUの構成を予測したテスト結果に基づくサイジング ツールを使用しました。インスタンスとGPU数の大まかな前提条件を表2に示します。

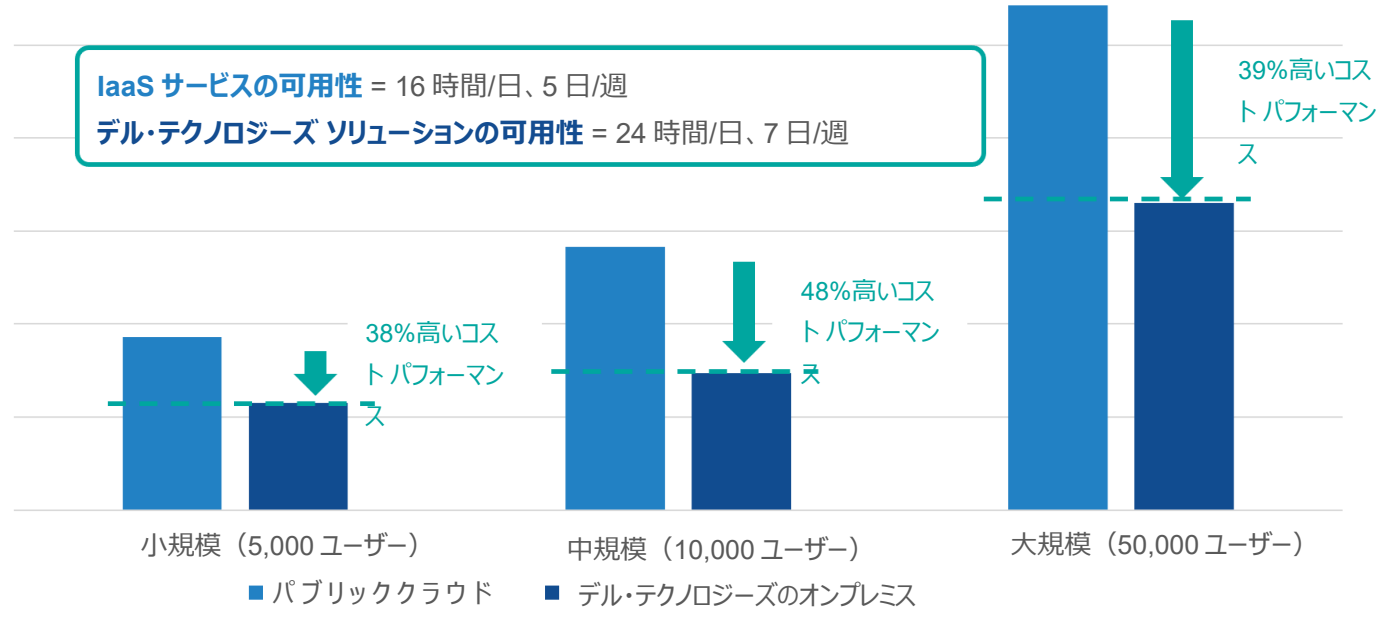
表2. Mistral 7Bパラメーター モデル推論の構成の前提条件

LLMモデル（パラメーター数）	ユーザー数	推論ノード/インスタンスの数	H100 GPUの数
Mistral (7B)	5,000	1	1
	10,000	1	2
	50,000	1	4

出典：*Enterprise Strategy Group（TechTarget, Inc.の部門）*

次に、表1にまとめたすべてのコストを構成ごとにモデル化しました。図2に示すように、デル・テクノロジーズのインフラストラクチャは、組織に推論を提供する際のコスト パフォーマンスが1.6～1.9倍（38%～48%）優れていると同時に、組織が24時間年中無休で利用できることがわかりました。

図2：RAGを使用した7BパラメーターMistral LLMの推論を提供する場合の予測コスト



出典：*Enterprise Strategy Group（TechTarget, Inc.の部門）*

大規模モデル：70BパラメーターのLlama 2 LLM

次に、オープンソースのLlama 2 70B LLMと同様に、700億個のパラメーターを持つ大規模なモデルを提供するために予想されるコストをモデル化しました。同じサイジング ツールを使用して要件を再度サイジングし、サーバーとGPUの構成を予測しました。この構成では、要求あたりの平均レイテンシーはわずかに高めの約1.8秒、推論1秒あたりの推定スループット2.29～22.86を提供可能です。インスタンスとGPU数の大まかな前提条件を表3に示します。

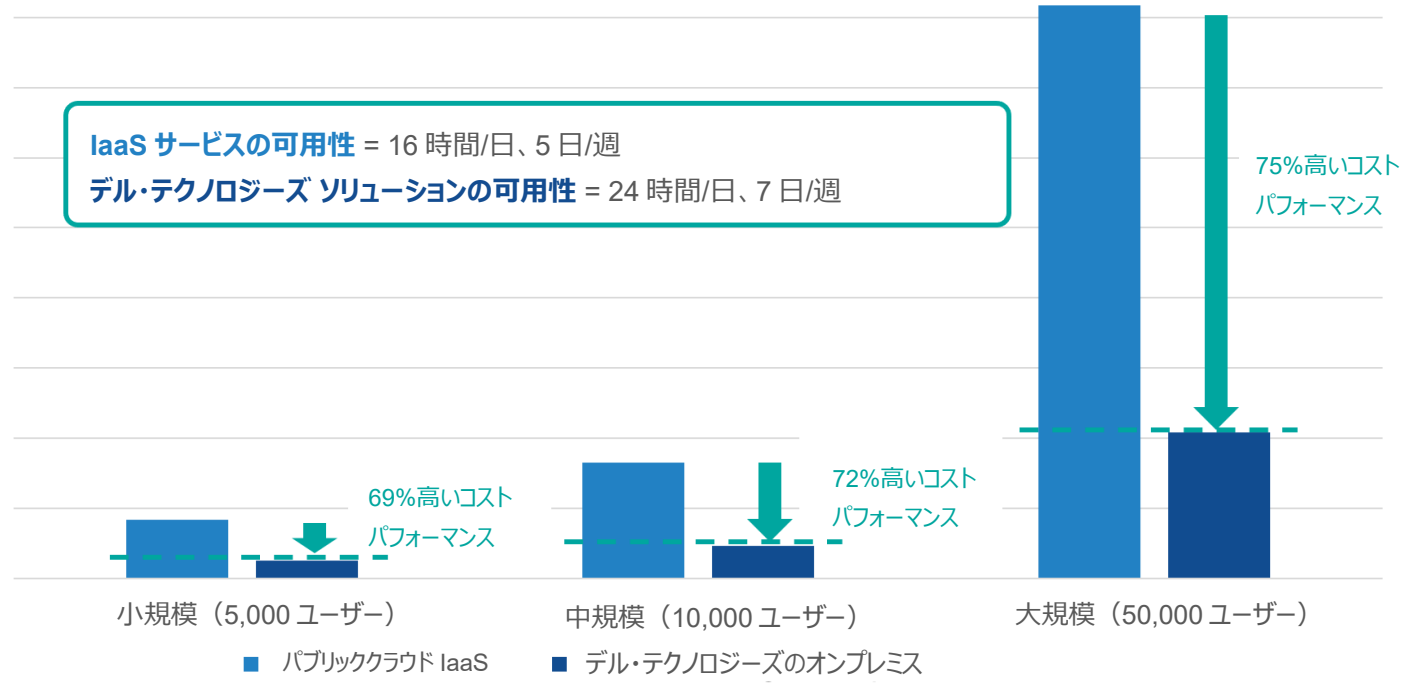
表3. Llama 2 70Bパラメーター モデル推論の構成の前提条件

LLMモデル（パラメーター数）	ユーザー数	推論ノード/インスタンスの数	H100 GPUの数
Llama 2 (70B)	5,000	2	8
	10,000	4	16
	50,000	20	80

出典：*Enterprise Strategy Group（TechTarget, Inc.の部門）*

上記の各構成について、表1にまとめたすべてのコストを再度モデル化した結果、デル・テクノロジーズのインフラストラクチャは、組織に推論を提供する際のコスト パフォーマンスが3.3～4倍（69%～75%）高いと同時に、組織が24時間年中無休で利用できることがわかりました。

図3：RAGを使用した70BパラメーターLlama 2 LLMの推論を提供する場合の予測コスト

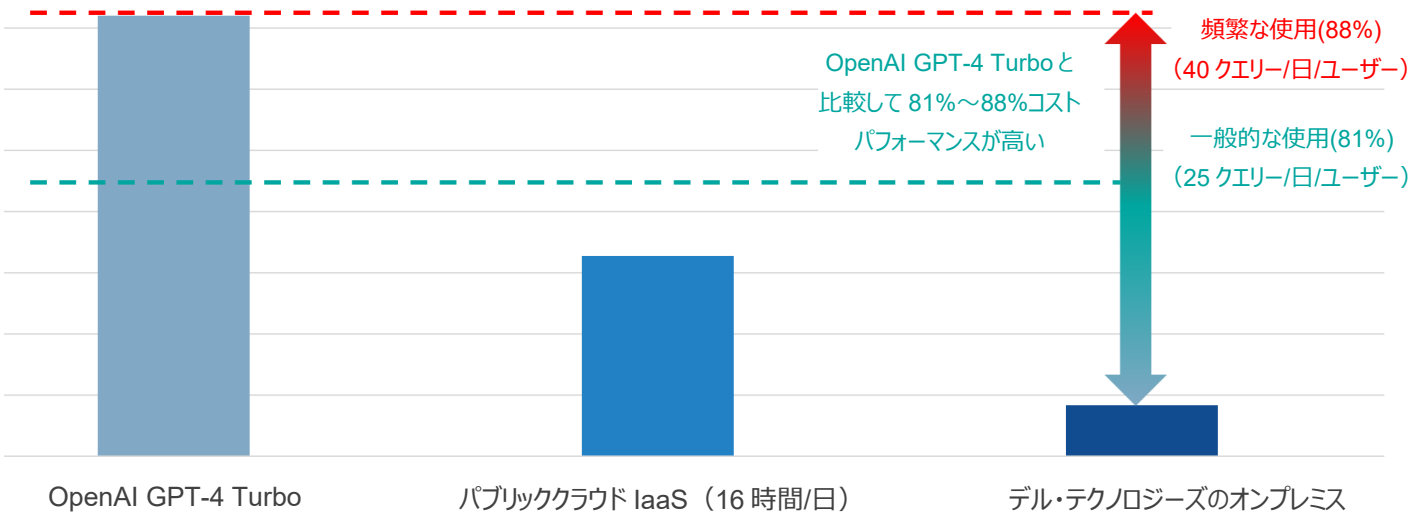


出典：*Enterprise Strategy Group（TechTarget, Inc.の部門）*

デル・テクノロジーズのオンプレミス インフラストラクチャとAPIベースの生成AIサービスの比較

次に、入力と出力の「トークン」ごとにコスト パフォーマンスの高い価格が設定されている、確立されたOpenAI APIベースの生成AIサービスGPT-4 Turboを使用して、大規模な組織が50,000人のユーザーに同等の70Bパラメーター モデルを提供する場合に予測されるコストを比較しました。テキストベースのQ&Aでは、クエリーごとに適度なトークン強度が必要であり、ピーク負荷の変動が大きくなく、必要な入力トークンと出力トークンの数のバランスが比較的均等になります。クエリーあたり合計1,500トークン（入力と出力）を想定し、オンプレミス ソリューションとAPIベースのソリューションの両方で、ユーザーごとに1日あたり平均約25クエリーを想定しました。公式声明の調査によると、これはユーザーごとのクエリー数が中程度の状態であり、比較的新興の組織ではユーザーごとのクエリー数が少なく、既存の安定した組織では、ユーザーごとの1日あたり平均クエリー数が40件であることがわかりました。GPT-4 Turboの計算では、コストはユーザーあたり月額約12.50ドルと予測されており、ユーザーあたり月額約30ドルのスイートベースのAI支援ツールと比較しても遜色ありません。これらの仮定に基づき、デル・テクノロジーズのオンプレミス インフラストラクチャは、APIベースのサービスを使用する場合よりも推論のコスト パフォーマンスが5.4～8.6倍（81%～88%）高く、生成AI機能をユーザーあたり月額約2.31ドルで提供できることがわかりました。

図4：70BパラメーターのLlama 2 LLMの推論を50,000人のユーザーに提供する場合の予測コスト



出典：*Enterprise Strategy Group（TechTarget, Inc.の部門）*

考慮すべき問題

Enterprise Strategy Groupのモデルは、保守的で信頼性が高く、検証済みの仮定に基づいて誠実に構築されていますが、1つのモデル化されたシナリオがすべての潜在的な環境を表すことはありません。お客様のコスト削減は、特定のユース ケース、データの性質、専門知識のレベル、モデルとインフラストラクチャの要件によって異なります。Enterprise Strategy Groupでは、利用可能な製品を独自に分析し、デル・テクノロジーズと相談して、自社独自の概念実証テストで実証済みのソリューションとの違いを理解し、検討することをお勧めします。

LLM推論向けのデル・テクノロジーズ ソリューション

デル・テクノロジーズは、組織がデータの場所を問わず簡単にデータにAIを導入できるよう支援します。これは、デスクトップからデータセンター、クラウドまで、AIサービスの幅広いポートフォリオを提供することを意味します。これにより、組織は適切な規模の投資を行い、データを活用してAIファクトリーを構築し、効率的かつ安全に、持続可能な方法でAIのユース ケースを実現できます。Dellは、包括的なサービス ポートフォリオと、広範でオープンなパートナーのエコシステムへのアクセスを提供することで、これを実現します。AI戦略の策定、生成AIへの投資の加速と拡大など、AIの導入のどの段階にあっても、組織を支援します。

Dell Professional Services for Generative AIは、データ セキュリティの脅威、コンプライアンス上の懸念、データ サイロ、未検証のデータ セットといった課題を抱える組織における取り組みを支援します。優先度の高いユース ケースに関するビジネス リーダーとITリーダーの合意形成、目標達成のための実用的なロードマップの提供、LLM統合に向けた企業データの準備、サイバーセキュリティの成熟度の向上、特定のビジネス ニーズに合わせたAIプラットフォームの確立などに役立ちます。さらに、Dell APEXを利用すると、組織はAIソリューションをサブスクライブし、マルチクラウドのユース ケースに合わせて最適化することができます。

Dellのソリューションの詳細については、DellのAIに関するWebページを参照してください。

結論

ビジネスのほぼすべての領域で生成AIの使用が拡大することは、運用の改善と将来の成功を確実にするうえで重要な要素です。Enterprise Strategy Groupの調査によると、組織が現在生成AIを適用している上位の分野には、研究、マーケティング、ソフトウェア開発、製品開発、IT運用などがあり、あらゆる分野での使用の可能性が高まることが予想されています。[4]組織は、独自にカスタマイズしたバージョンのLLMに対してトレーニングと推論を行うことで、よりインパクトのある有意義な結果を達成できます。

LLMの推論に使用できる導入の方法は複数あり、それぞれが特定のユース ケースや要件に対してメリットをもたらします。カスタマイズされたLLMに含まれる機能を活用する準備ができている数千人のユーザーを抱える大規模な組織の場合、デル・テクノロジーズのインフラストラクチャは、IaaSと比較して最大4倍、OpenAI GPT-4 Turboと比較して最大8倍コスト パフォーマンスに優れた、高性能なLLM推論を提供できます。Enterprise Strategy Groupは、LLMを実装して組織を強化する企業に対し、デル・テクノロジーズが提供するコスト パフォーマンスに優れたテクノロジーと知識豊富なサービスを活用することを強く推奨します。これにより、確かな成果を挙げて生成AIイニシアティブを加速させ、期待されるコスト削減の達成にかかる時間を短縮することができます。

[4]出典：*Enterprise Strategy Group* の調査レポート『Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns』（2023 年 8 月）。

Enterprise Strategy Group について

TechTarget の Enterprise Strategy Group は、焦点を絞った実践的なマーケット インテリジェンス、デマンドサイド調査、アナリスト アドバイザリー サービス、GTM 戦略ガイダンス、ソリューション検証、エンタープライズ テクノロジーの売買をサポートするカスタム コンテンツを提供しています。

contact@esg-global.com
www.esg-global.com