Azure OpenAI Service における主要 GPT モデルの比較分析：料金、トークン制限、マルチモーダル機能

はじめに

目的

本レポートは、Microsoft Azure OpenAI Service 上で利用可能な主要な大規模言語モデル（LLM）およびマルチモーダルモデルについて、技術的および経済的な観点から比較分析を提供することを目的とします。特に、利用者の関心が高いと考えられる gpt-4o, gpt-4o-mini, gpt-4, gpt-3.5-turbo, および関連する oシリーズ モデル (o1, o3, o4-mini, o3-mini, o1-mini 等) に焦点を当て、現時点（本レポート執筆時点）での利用料金、トークン制限（コンテキストウィンドウと最大出力）、マルチモーダル機能（テキスト以外のデータ処理能力）を整理します。

背景

Azure OpenAI Service は、OpenAI の先進的な AI モデルを Microsoft Azure の堅牢なクラウドインフラストラクチャ上で提供するサービスです [1]。これにより、企業は Azure のセキュリティ、コンプライアンス、スケーラビリティ、他の Azure サービスとの統合といった利点を享受しながら、最先端の AI 機能をアプリケーションに組み込むことが可能になります [3]。提供されるモデルは多岐にわたり、それぞれ異なる能力と価格帯を持つため、利用目的に応じた適切なモデル選択は、コスト、性能、機能要件のトレードオフを伴う重要な意思決定となります。

スコープ

本レポートは、公開されている情報に基づき、Azure OpenAI Service における指定モデルの現状を分析します。分析対象は、利用料金体系、トークン処理能力、マルチモーダル対応状況に限定します。将来的な価格や仕様の変更に関する注意喚起は、利用者の要望に基づき省略します。Azure 固有のデプロイメントタイプ（Standard (Pay-As-You-Go), Provisioned Throughput Units (PTU), Batch API）やリージョンによる差異にも可能な範囲で触れますが、主眼は標準的な従量課金制（Pay-As-You-Go）モデルとします [2]。

1. Azure OpenAI Service モデルランドスケープ

1.1. 利用可能な主要モデルの特定

Azure OpenAI Service では、OpenAI によって開発された多様なモデルファミリーが提供されています。これには、自然言語処理と生成に優れた GPT-4 シリーズ、GPT-3.5 シリーズ、画像生成に特化した DALL-E シリーズ、テキストのベクトル表現を生成する Embeddings モデル、音声認識と翻訳を行う Whisper モデル、そして近年登場した高度な推論能力を持つ o シリーズなどが含まれます [2]。

利用者が一般的に使用する可能性のある名称や略称と、Azure OpenAI Service 上で実際に

デプロイされる際の正式なモデル名およびバージョンを対応付けることは、混乱を避け、正確な比較を行う上で重要です。以下の表に、本レポートで主に比較対象とするモデルのマッピングを示します。

表 **1: Azure OpenAI** モデルマッピング

利用者照会用語 **(**可能性**)**	対応する **Azure OpenAI** モデルファミリー	**Azure** 上の代表的な最新デプロイメント名**/**バージョン **(GA)**
4o	GPT-4o	gpt-4o-2024-11-20
4o mini	GPT-4o mini	gpt-4o-mini-0718
o4 / o4mini	o4-mini	o4-mini-2025-04-16
o3	o3	o3-2025-04-16
o3mini	o3-mini	o3-mini-2025-01-31
o1	o1	o1-2024-12-17
o1mini	o1-mini	o1-mini-2024-09-12
(GPT-4)	GPT-4	gpt-4-turbo-2024-04-09 (GPT-4 Turbo with Vision)
(GPT-3.5)	GPT-3.5-Turbo	gpt-3.5-turbo-0125

出典: 4
注: 本表は代表的な最新の一般提供（GA）バージョンを示しており、プレビュー版や旧バージョンも存在する場合があります。o4-mini, o3, o1-mini のバージョンは 6 等に基づきます。

1.2. モデルバージョンの管理とアップデート

Azure OpenAI Service では、各モデルはリリース日などを示す特定のバージョン名（例: gpt-4o-2024-11-20）で管理され、OpenAI との連携のもと、定期的に新しいバージョンがリリースされます [8]。利用者はモデルをデプロイする際に、アップデートポリシーを選択できます。これには、新しいデフォルトバージョンがリリースされた際に自動で更新される「自動更新」、現在使用中のバージョンが廃止される際に自動で更新される「期限切れ時にアップグレード」、自動更新を行わずバージョンを固定する（ただし廃止時には利用不可となる）「自動アップグ

レードなし」といった選択肢があります [9]。

新しいモデルバージョンがリリースされる際には、Azure から通知があり、通常、新バージョンがデフォルトになる少なくとも 2 週間前に告知されます。また、主要な旧バージョンも廃止日まで利用可能に保たれるため、必要に応じて切り替えることが可能です [9]。

留意点として、Azure OpenAI Service におけるモデルのアップデートサイクルは、OpenAI 本体での最新モデルのリリースと必ずしも完全に一致するわけではありません [10]。これは、Azure がエンタープライズ利用を前提としており、プラットフォーム上での安定性、セキュリティ検証、コンプライアンス要件への適合、既存の Azure サービスとの統合テストなどを経て提供されるためです [5]。このプロセスにより、OpenAI 本体での発表から Azure での一般提供開始までに時間差が生じることがあります [10]。結果として、Azure 利用者は、最新鋭の機能を即座に利用できない可能性がある一方で、より検証され、安定した環境でモデルを利用できるという側面も持ちます。

さらに重要な点として、モデルバージョンの違いは、単なる機能追加や性能向上だけでなく、応答の挙動、API の互換性、トークン制限、そして利用料金にも影響を与える可能性があります [4]。例えば、新しいバージョンでは指示追従性が向上したり [12]、特定のパラメータの挙動が変わったりすることがあります。また、価格体系やコンテキストウィンドウサイズが見直されることもあります [4]。したがって、特定のモデルバージョンに依存して構築されたアプリケーションは、バージョンアップグレード時に互換性テストや、必要に応じたプロンプトの調整、あるいはコードの修正が必要になる場合があるため、バージョン管理方針を慎重に検討する必要があります [9]。

2. Azure OpenAI モデルの比較分析

2.1. 利用料金体系

Azure OpenAI Service の標準的な課金モデルは、従量課金制（Pay-As-You-Go）であり、モデルが処理した「トークン」の数に基づいて料金が計算されます [1]。トークンは、モデルがテキストを処理する際の基本単位であり、単語や文字の一部に対応します。目安として、英語では 1 トークンあたり約 4 文字、または約 3/4 語に相当するとされています [1]。日本語の場合、漢字、ひらがな、カタカナ、アルファベットなどが混在するため、単純な文字数や単語数との換算は難しいですが、一般的に英語よりも多くのトークンを消費する傾向があります。

課金においては、モデルへの入力（プロンプト）に使われたトークン数と、モデルからの出力（補完、Completion）として生成されたトークン数が別々にカウントされ、それぞれに異なる単価が設定されていることが一般的です [4]。例えば、1,000 トークンのプロンプトを入力し、1,000 トークンの応答を得た場合、合計 2,000 トークン分の料金が発生しますが、入力と出力の単価が異なれば、その合計額は単純な 2,000 トークン分の料金とは異なります [10]。

以下の表は、本レポート執筆時点における主要モデルの 100 万トークンあたりの利用料金（Azure Global Standard Deployment、Pay-As-You-Go）を比較したものです。

表 2: Azure OpenAI モデル別 100 万トークンあたりの料金比較 (USD)

モデルファミリー	**Azure デプロイメント名/バージョン**	**100万入力トークン**	**100万キャッシュ入力トークン**	**100万出力トークン**	**Batch API (入力/出力, 1M)**	備考
GPT-4o	gpt-4o-2024-11-20	$2.50	$1.25	$10.00	$1.25 / $5.00	最新大規模GAモデル。初期版(0513)は$5/$15 [4]。Realtime APIは別価格 [4]。
GPT-4o mini	gpt-4o-mini-0718	$0.15	$0.075	$0.60	$0.075 / $0.30	高コスト効率モデル [4]。Realtime APIは別価格 [4]。
GPT-4	gpt-4-turbo-2024-04-09	$10.00	-	$30.00	-	GPT-4 Turbo with Vision。旧GPT-4 (8K)は$30/$60, (32K)は$60/$120 [15]。
GPT-3.5-Turbo	gpt-3.5-turbo-0125	$0.50	-	$1.50	-	最新GAモデル。Instruct版は$1.50/$2.00 [10]。旧版(0613

						4K)は $1.50/$2.00 [17]。
o4-mini	o4-mini-2025-04-16	$1.10	$0.275	$4.40	-	Azure価格未公開。OpenAI API価格を参考 [18]。推論特化。
o3	o3-2025-04-16	$10.00	$2.50	$40.00	-	Azure価格未公開。OpenAI API価格を参考 [18]。最上位推論モデル。
o3-mini	o3-mini-2025-01-31	$1.10	$0.55	$4.40	$0.55 / $2.20	推論モデルのコスト効率版 [4]。
o1	o1-2024-12-17	$15.00	$7.50	$60.00	N/A	高度推論モデル [4]。
o1-mini	o1-mini-2024-09-12	$1.21	$0.605	$4.84	N/A	US/EUデータゾーン価格 [4]。Global価格はN/A。OpenAI API価格($1.10/$4.40)と近似 [18]。

出典: 4

注: 上記は Global Standard Deployment の Pay-As-You-Go 価格（USD）です。リージョン（US/EU データゾーン、その他リージョン）によっては価格が異なる場合があります（通常、Global より若干高価）4。キャッシュ入力価格や Batch API 価格は利用可能なモデル/バージョンが限られます 4。価格は変更される可能性があるため、最新情報は公式価格ページで確認することが推奨されます 10。価格設定に関して注目すべき点として、必ずしも最新世代のモデルが最も高価であるとは限

らない傾向が見られます。例えば、GPT-4o mini は GPT-3.5 Turbo の一部バージョン（例: gpt-3.5-turbo-instruct）よりも安価であり、GPT-4o (1120) は初期の GPT-4 (32k) よりも大幅に安価です [4]。これは、モデル開発における性能向上と並行して、推論効率の改善やアーキテクチャの最適化が進んでいること、また、普及を促進するための戦略的な価格設定が行われていることを示唆しています [2]。したがって、利用者は「新しいモデル＝高価」という単純な前提にとらわれず、具体的なモデルバージョンごとの価格と、自身のタスクで要求される性能レベルを比較検討することが重要です。コストパフォーマンスは、モデル世代だけでなく、特定のバージョンや利用シナリオによって変動します。

さらに、近年の価格表には「キャッシュ入力 (Cached Input)」という項目が登場しています [4]。これは、API コール間で繰り返し使用される入力トークン（例えば、チャット履歴の一部や RAG (Retrieval-Augmented Generation) で検索された文書チャンクなど）に対して、通常の入力トークン価格よりも割引された価格（通常 50% 割引）が適用される仕組みです [4]。多くの AI アプリケーション、特に会話型 AI や RAG を利用するシステムでは、過去のコンテキスト情報を維持するために同じ内容をプロンプトに含めて繰り返し送信するケースが少なくありません [13]。キャッシュ入力価格の導入は、このようなシナリオにおいて、入力トークンコストを大幅に削減できる可能性を示唆しています。これは、単に性能や機能でモデルを選択するだけでなく、API の利用パターンを最適化すること（例えば、キャッシュが有効に機能するようなセッション管理やコンテキスト管理の工夫）が、運用コスト管理においてますます重要になることを意味します。

その他、料金に関する補足事項として、以下の点が挙げられます。

- ファインチューニング: 既存モデルを特定のデータセットで追加学習させるファインチューニングには、トークン処理料金とは別に、トレーニング時間またはトレーニングに使用されたトークン数に基づくコストと、ファインチューニング済みモデルをデプロイしておくためのホスティング時間コストが発生します [10]。
- **Provisioned Throughput Units (PTU):** 大規模かつ予測可能なワークロードを持つ利用者向けに、一定量のモデル処理能力を予約する PTU という購入オプションがあります。これにより、安定したスループットと予測可能なコストが実現できますが、従量課金制とは異なる価格体系となります [2]。
- **Batch API:** 大量のデータを非同期で処理する場合、通常のリアルタイム API よりも割引された価格が適用されることがあります [4]。
- 高レベル **API** とツール利用: Assistants API のような高レベル API を利用する場合、API が内部で Code Interpreter や File Search といったツールを呼び出すと、それらのツール利用に対して追加のセッション料金やストレージ料金が発生することがあります [4]。
- 関連 **Azure** サービス: Azure OpenAI Service の利用に伴い、監視のための Azure Monitor Logs や、ネットワーク設定、データストレージなど、他の Azure サービスの利用料金が別途発生する可能性があります [10]。

2.2. トークン制限

各 Azure OpenAI モデルには、一度の API コールで処理できる最大の入力トークン数（コンテキストウィンドウサイズ）と、一度に生成できる最大の出力トークン数が定められています [8]。

コンテキストウィンドウは、モデルが一度に考慮できる情報の量を決定します。このサイズが大きいほど、より長い文書全体を読み込ませたり、より複雑な指示を与えたり、より長期にわたる会話履歴を維持したりすることが可能になります [12]。これにより、文脈に基づいたより正確な応答や、一貫性のある対話が期待できます。

最大出力トークン数は、モデルが一度のリクエストに対して生成できるテキストの最大長を制限します [8]。この制限を超える長さの応答が必要な場合は、複数回に分けて生成するなどの工夫が必要になります。

以下の表は、主要モデルのトークン制限を比較したものです。

表 **3: Azure OpenAI** モデル別 トークン制限比較

モデルファミリー	**Azure** デプロイメント名 **/**バージョン	最大入力コンテキストウィンドウ **(**トークン数**)**	最大出力トークン数 **(**トークン数**)**
GPT-4o	gpt-4o-2024-11-20	128,000	16,384
GPT-4o mini	gpt-4o-mini-0718	128,000	16,384
GPT-4	gpt-4-turbo-2024-04-09	128,000	4,096
GPT-3.5-Turbo	gpt-3.5-turbo-0125	16,385	4,096
o4-mini	o4-mini-2025-04-16	200,000	100,000
o3	o3-2025-04-16	200,000	100,000
o3-mini	o3-mini-2025-01-31	200,000	100,000
o1	o1-2024-12-17	200,000	100,000
o1-mini	o1-mini-2024-09-12	128,000	65,536

出典: 8
注: 上記は代表的なバージョンの値です。旧バージョンでは異なる場合があります（例: 初期 GPT-4 は 8K/32K、初期 GPT-3.5 Turbo は 4K）。
近年のモデル、特に GPT-4 Turbo、GPT-4o ファミリー、o シリーズ、そして GPT-4.1 シリーズでは、128,000 トークンや 200,000 トークンといった非常に大きなコンテキストウィンドウが提供されるようになっています [8]。これは、初期の GPT-3.5 Turbo (4K/16K) や GPT-4 (8K/32K) と比較して、扱える情報量が飛躍的に増大したことを意味します。大規模なコンテキストウィンドウは、書籍のような長文コンテンツの分析、複数文書にまたがる情報の統合、非常に長い会話履歴の維持、複雑な Few-shot プロンプティングなどを可能にし、AI アプリケーションの可能性を大きく広げます [12]。長文処理や複雑な背景理解が求められるタスクにおいて、これらの最新モデルは旧モデルに対して明確な技術的優位性を持っています。ただし、コンテキストウィンドウが大きくなると、それに比例して API コールあたりの潜在的なコストや応答時間が増加する可能性もあるため、タスクの要件に応じて必要十分なコンテキストを持つモデルを選択することが依然として重要です [13]。

一方、最大出力トークン数も増加傾向にはありますが、入力コンテキストウィンドウほどの劇的な拡大は見られません。例えば、最新の GPT-4o や GPT-4o mini は 16K トークンの出力が可能ですが、GPT-4 Turbo や初期の GPT-4o (0513) は 4K トークンに制限されています [8]。o シリーズは 100K という大きな出力が可能ですが、これは推論タスクにおける中間ステップの生成なども考慮されている可能性があります [8]。出力トークン数の制限が入力ほど大きくない背景には、非常に長いテキストを一貫性を保ちながら高品質に生成することの技術的な難しさや、長大な応答生成に伴う計算コストおよびレイテンシへの配慮があると考えられます。多くのインタラクティブなユースケースでは、極端に長い単一の応答よりも、より短い応答を対話的に繰り返す方が効果的な場合もあります [25]。したがって、非常に長いレポートやコード全体を一度に生成する必要があるような特定のタスクでは、最大出力トークン数が制約となる可能性があります。その場合、出力を複数ステップに分割して生成し、後で結合するといったアプリケーション側の工夫が必要になるかもしれません。モデルを選択する際には、入力コンテキストウィンドウだけでなく、最大出力トークン数もユースケースの要件と照らし合わせて評価する必要があります。

2.3. マルチモーダル機能

マルチモーダル AI とは、従来のテキストデータに加えて、画像、音声、将来的には動画など、複数の異なる種類のデータ（モダリティ）を入力として受け付けたり、出力として生成したりできる AI モデルを指します [2]。これにより、よりリッチで現実に近いインタラクションや分析が可能になります。

Azure OpenAI Service において、マルチモーダル機能を備えている主要なモデルファミリーは、GPT-4o、GPT-4o mini、GPT-4 Turbo with Vision、そして o シリーズです [6]。

- 画像入力: GPT-4o、GPT-4o mini、GPT-4 Turbo with Vision、o1 などが対応しています [6]。これらのモデルは、入力された画像の内容を理解し、説明文を生成したり、画像に関する質問に答えたりすることができます。画像入力に対するトークンコストは、テキストとは別に計算されます。画像はまず特定のサイズ（例: 2048x2048 ピクセル内）にリサイズされ、さらに 512x512 ピクセルのタイルに分割されます。各タイルに対して一定のトークンコスト（例: GPT-4o/Turbo は 170 トークン、GPT-4o mini は 5667 トークン）が課され、さらにベースとなる固定トークン（例: GPT-4o/Turbo は 85 トークン、GPT-4o mini は 2833 トークン）が加算されるという計算方式が採用されています [6]。
- 音声入力/出力: 現時点では、主に GPT-4o ファミリーが音声機能を担っています [8]。具体的には、低遅延でのリアルタイム音声対話を実現する gpt-4o-realtime-preview や gpt-4o-mini-realtime-preview モデル [4]、音声ファイルからの文字起こしを行う gpt-4o-transcribe や gpt-4o-mini-transcribe モデル [23]、テキストから自然な音声を合成する gpt-4o-mini-tts モデルなどが提供されています [8]。これらの音声処理機能にも、テキストとは異なる独自の価格設定（例: リアルタイム API の音声入力/出力トークン単価、文字起こし/音声合成の分単位課金）が適用されます [4]。
- 画像出力: 画像生成は、GPT モデルではなく、DALL-E シリーズのモデルが担当します [6]。
- **GPT-3.5 Turbo:** 基本的にテキストのみを扱うモデルであり、画像や音声の入出力機能はサポートされていません [8]。

以下の表は、主要モデルのマルチモーダル対応状況をまとめたものです。

表 **4: Azure OpenAI** モデル別 マルチモーダル機能比較

モデルファミリー	**Azure** デプロイメント名/バージョン	テキスト入力	画像入力	音声入力	テキスト出力	画像出力	音声出力	備考
GPT-4o	gpt-4o-2024-11-20	Yes	Yes	Yes (専用モデル経由/Preview)	Yes	No	Yes (専用モデル経由/Preview)	単一モデルでテキスト・画像を統合処理 [8]。音声は realtime, transcr

								ibe, audio-preview 等で対応[4]。
GPT-4o mini	gpt-4o-mini-0718	Yes	Yes	Yes (専用モデル経由/Preview)	Yes	No	Yes (専用モデル経由/Preview)	GPT-4o の軽量版。音声は realtime, transcribe, tts 等で対応[4]。
GPT-4	gpt-4-turbo-2024-04-09	Yes	Yes (Turbo Vision)	限定的	Yes	No	No	turbo-2024-04-09 が Vision 対応[8]。Azure AI Vision 拡張機能は非対応[8]。ベース GPT-4 はテキストのみ。
GPT-3.5-Turbo	gpt-3.5-turbo-0125	Yes	No	No	Yes	No	No	テキスト処理に特化[8]。
o4-mini	o4-mini-2025-04-16	Yes	Yes	No	Yes	No	No	推論特化。画像入力対応[8]。音声対

								応の言及なし。
o3	o3-2025-04-16	Yes	Yes	No	Yes	No	No	推論特化。画像入力対応[8]。音声対応の言及なし。
o3-mini	o3-mini-2025-01-31	Yes	Yes	No	Yes	No	No	推論特化。画像入力対応[8]。音声対応の言及なし。
o1	o1-2024-12-17	Yes	Yes	No	Yes	No	No	推論特化。画像入力対応[8]。音声対応の言及なし。
o1-mini	o1-mini-2024-09-12	Yes	Yes	No	Yes	No	No	推論特化。画像入力対応[8]。音声対応の言及なし。

出典: 4

注:「音声入力/出力」の「専用モデル経由」は、gpt-4o-realtime-preview, gpt-4o-transcribe, gpt-4o-mini-tts などの特定のモデル/API を指します。これらは GPT-4o ファミリーの一部として提供されています。o シリーズについては、現時点で Azure 上での明確な音声機能の提供に関する情報は見当たりませんでした。画像出力は DALL-E が担当します。

マルチモーダル機能は急速に進化しており、特に GPT-4o ファミリーがその最前線にいます[2]。これらのモデルは、テキスト、画像、音声を単一のニューラルネットワーク内で統合的に処理する能力を持つように設計されており[8]、従来のように画像認識モデル、音声認識モデル、テ

キスト生成モデルを個別に用意し、それらをパイプラインで繋ぎ合わせるアプローチ [26] と比較して、よりシームレスで、レイテンシが低く、モダリティ間の文脈理解に基づいた高精度な応答を提供する可能性があります [24]。この統合的なアプローチは、開発の複雑さを軽減し、カスタマーサービス、コンテンツ作成、データ分析など、多様な分野での新しいアプリケーションの創出を加速させると期待されています [2]。

一方で、マルチモーダル機能を利用する際には、コスト計算の複雑化という側面も考慮する必要があります。前述のように、画像入力はピクセル数に基づくタイル計算 [6]、音声処理は時間単位または専用のトークン単価 [4] といったように、モダリティごとに異なる課金体系が適用されます。テキスト処理のトークンコストに加えてこれらのコストが発生するため、マルチモーダルアプリケーション全体の運用コストを正確に見積もることは、テキストのみの場合よりも難しくなります。特に、大量の高解像度画像や長時間の音声データを処理する場合には、コストが予想以上に増加する可能性があるため、利用量に応じた慎重なコストシミュレーションと管理が不可欠です。

3. 主要な考慮事項と推奨事項

3.1. コスト vs. 性能 vs. 機能

Azure OpenAI Service で利用可能なモデルは多岐にわたり、それぞれコスト、性能（推論能力、精度、速度）、機能（コンテキスト長、マルチモーダル対応）の点で特徴が異なります。最適なモデルを選択するには、これらの要素間のトレードオフを理解し、特定のユースケース要件と照らし合わせることが不可欠です。

- トレードオフの分析:
  - **GPT-3.5-Turbo:** 最も低コストな選択肢の一つであり、応答速度も比較的速い傾向があります [13]。基本的なテキスト生成、簡単な要約、FAQ ベースの応答など、高度な推論や創造性、長い文脈の理解を必要としないタスクに適しています。ただし、コンテキストウィンドウは最新版でも 16K トークンと他の最新モデルに比べて小さく、マルチモーダル機能は提供されていません [8]。
  - **GPT-4 (旧世代):** GPT-3.5 Turbo と比較して、格段に高い推論能力、精度、創造性を持ち、複雑な指示への追従性も優れています [1]。専門的な文章作成、高度なコーディング支援、複雑な問題解決に適していますが、利用コストは GPT-3.5 Turbo の数倍から数十倍と高価になります [13]。初期のモデル (0613 など) はコンテキストウィンドウが 8K または 32K に限られます [8]。
  - **GPT-4 Turbo (with Vision):** GPT-4 の高い能力を維持しながら、128K トークンという巨大なコンテキストウィンドウと画像入力機能を提供します [8]。コストは旧世代の GPT-4 (特に 32K 版) よりも抑えられている傾向があります [4]。長大な文書の読解・要約や、画像を含むレポート作成などのタスクに適しています。
  - **GPT-4o:** 最新のフラッグシップモデルであり、GPT-4 Turbo と同等以上のテキストおよびコーディング性能を持ちながら、特に非英語言語での性能が向上し、さらにテキ

スト、画像、音声の各モダリティを統合的に扱えるように設計されています [2]。利用コストは GPT-4 Turbo と同程度か、最新版では若干低く設定されていますが、音声処理には追加コストが発生します [4]。多様な入出力形式を扱う最先端のアプリケーションや、より自然な対話体験が求められる場合に最適です。
    - **GPT-4o mini:** GPT-4o の持つマルチモーダル対応（テキスト、画像、音声）や 128K コンテキストといった特徴を維持しつつ、大幅に低コスト化（GPT-4o の約 1/10 以下）と高速化を実現したモデルです [2]。性能面でも GPT-3.5 Turbo を凌駕すると評価されており [14]、コストと性能のバランスが非常に優れています。多くの標準的なタスクにおいて、有力な第一候補となり得るモデルです。
    - **o シリーズ (o1, o3, o4-mini, o3-mini, o1-mini):** これらのモデルは、特に複雑な推論、多段階の問題解決、計画立案、数学、科学、コーディングといったタスクに最適化されています [2]。GPT-4 シリーズとは異なる価格帯と特性を持ち、より高度な自律型エージェント機能や専門的な分析が求められる場合に強みを発揮します。画像入力にも対応しています [8]。mini 版は、o シリーズの高い推論能力をより低コストで利用したい場合に適しています [4]。
- 推奨事項:
    - コスト最優先、高速応答: gpt-3.5-turbo-0125 または gpt-4o-mini-0718。タスクの複雑性が低ければ GPT-3.5、多少複雑でもコストを抑えたいなら GPT-4o mini が有力です。
    - 最高の性能（テキスト・コード）、コスト許容: gpt-4o-2024-11-20 または gpt-4-turbo-2024-04-09。特に高度な推論や計画が必要な場合は o3 や o1 も検討対象となります。
    - 長文処理、大規模コンテキスト: 128K 以上のコンテキストを持つ gpt-4-turbo-2024-04-09, gpt-4o-2024-11-20, gpt-4o-mini-0718, o1-mini, または 200K コンテキストを持つ o1, o3, o4-mini, o3-mini。GPT-4.1 シリーズも 1M トークン対応 [12]。
    - 画像入力が必要: gpt-4-turbo-2024-04-09, gpt-4o-2024-11-20, gpt-4o-mini-0718, o1, o3, o4-mini, o3-mini, o1-mini。
    - 音声対話・処理が必要: gpt-4o ファミリー（関連する realtime, transcribe, tts モデルを利用）。
    - コストと性能のバランス重視: gpt-4o-mini-0718 は、多くのシナリオにおいて非常に魅力的な選択肢となる可能性が高いです [14]。

3.2. Azure OpenAI Service 特有の利点と考慮点

Azure OpenAI Service を利用する際には、OpenAI API を直接利用する場合と比較して、Azure プラットフォーム固有の利点と考慮すべき点があります。

- 利点:
    - エンタープライズグレードのセキュリティとプライバシー: Azure が提供する堅牢なセ

キュリティ機能、例えば Virtual Network (VNet) 統合によるプライベートネットワーク接続、Managed Identity (Microsoft Entra ID 連携) による安全な認証、リージョン指定によるデータ所在地の管理などが利用可能です [2]。これにより、特に機密性の高いデータを扱う企業にとって、データプライバシーとセキュリティを強化できます。
    - コンプライアンスとガバナンス: Azure が取得している多数の国際的および業界固有のコンプライアンス認証を活用できます [5]。また、Azure Policy を用いた利用制限などのガバナンス強化も可能です。
    - 責任ある **AI (Responsible AI)** 機能: 不適切または有害なコンテンツの生成リスクを低減するためのコンテンツフィルタリング機能がデフォルトで有効化されており、組織のポリシーに合わせてカスタマイズすることも可能です [5]。これは、AI の安全な利用を支援する重要な機能です。
    - **Azure** エコシステムとの統合: Azure AI Search (旧 Cognitive Search) を用いた RAG の構築、Azure AI Vision や Azure AI Speech との連携、Azure Machine Learning を用いた MLOps パイプラインへの組み込み、Azure Functions や Logic Apps によるワークフロー自動化、Cosmos DB や Azure SQL Database とのデータ連携など、他の Azure サービスとのシームレスな統合が容易です [2]。これにより、既存の Azure 資産を活用した包括的なソリューション構築が可能になります。
    - リージョン展開と可用性: 世界中の多くの Azure リージョンでサービスが提供されており、データレジデンシー要件を満たすために特定のリージョンを選択してデプロイできます [4]。また、Global Standard Deployment や Data Zone Deployment といった、可用性やデータ処理地域を考慮したデプロイオプションも提供されています [4]。
- 考慮点:
    - 最新モデル提供のタイムラグ: 前述の通り、OpenAI 本体で発表された最新のモデルや機能が Azure OpenAI Service で利用可能になるまでには、検証や統合プロセスを経るため、時間差が生じる場合があります [10]。
    - 応答挙動の微妙な差異: 同じモデル名（例: GPT-4o）であっても、Azure 上のデプロイメントでは、Azure 独自のシステムプロンプトの適用や、より厳格な安全フィルターの実装などにより、OpenAI API を直接利用した場合と応答の傾向が完全に同一にならない可能性があります [11]。
    - コスト構造の違い: 基本的なトークンあたりの価格は OpenAI API と一致させる方針が示されていますが [10]、Azure 上での利用には、VNet や Azure Monitor といった関連インフラストラクチャの利用料金が追加で発生する可能性があります [10]。また、ファインチューニングの課金体系（時間課金からトークン課金への変更など [23]）や PTU の価格設定は Azure 独自のものとなる場合があります [2]。

これらの点を踏まえると、Azure OpenAI Service の提供価値は、単に OpenAI のモデルへのアクセスを提供する点に留まらず、Azure プラットフォームが持つエンタープライズ向けの機能群（セキュリティ、コンプライアンス、ガバナンス、監視、他の Azure サービスとの統合容易性）と組み合わせることで最大化されると言えます。企業が AI を導入する際には、モデルの性能

だけでなく、データの安全性、運用管理の効率性、既存システムとの連携といった側面も極めて重要になります [2]。Azure OpenAI Service は、Azure AD (Entra ID) による厳格なアクセス管理、VNet 統合や Private Link によるネットワークレベルでの分離、Azure Policy による利用統制、Azure Monitor による詳細な利用状況監視といった、OpenAI API 単体では得られない、あるいは限定的な付加価値を提供します [6]。したがって、特に規制の厳しい業界の企業や、既に Azure を基盤として活用している企業にとっては、運用・管理・セキュリティ要件全体を考慮した場合、Azure OpenAI Service が OpenAI API よりも適した選択肢となる可能性が高いと考えられます。

3.3. マルチモーダル機能の活用と課題

GPT-4o ファミリーや GPT-4 Turbo with Vision、o シリーズなどのマルチモーダル対応モデルは、テキスト、画像、音声といった複数の情報源を統合的に扱うことで、新たなアプリケーションの可能性を切り開きます。

- 活用例:
  - コンテンツ生成: 製品画像と説明テキストを組み合わせて魅力的な広告コピーを生成する、図表を含むレポートを自動作成する、動画コンテンツの要約とハイライトを生成する（将来的な動画対応を見据えて）。
  - データ分析: 財務レポート内のグラフとテキストを同時に解釈してインサイトを抽出する、顧客からの問い合わせメールに添付された画像の内容を理解して応答を生成する、音声通話記録を文字起こしし、感情分析や要点抽出を行う [2]。
  - インタラクション: 画像をアップロードしてそれについて質問する、音声で指示を出して応答を得る、視覚障がい者向けに画像の内容を音声で説明するなど、より自然で多様なインターフェースを実現する [2]。
- 課題:
  - プロンプトエンジニアリングの高度化: テキストだけでなく、画像や音声をどのように効果的にモデルに提示し、それらを組み合わせた意図を正確に伝えるかが新たな課題となります [6]。例えば、画像内の特定の部分に注目させる指示や、音声のトーンを考慮した応答を求める指示など、より複雑なプロンプト設計が必要になる場合があります。
  - コスト管理の複雑化: 前述の通り、テキスト、画像、音声でそれぞれ異なる課金体系が存在するため、アプリケーション全体のコスト予測と管理が複雑になります [4]。利用するモダリティとデータ量に応じてコストが変動するため、予算管理には注意が必要です。
  - 性能評価の難しさ: マルチモーダルな応答（例: 画像の内容を踏まえたテキスト生成）の品質を客観的かつ定量的に評価するための標準的な指標や方法論は、テキストのみの場合と比較してまだ発展途上です。
  - データ準備と前処理: 画像や音声データをモデルが効率的に処理できる形式（解像度、フォーマット、サイズなど）に変換・準備する前処理ステップが必要になる場合が

あります [26]。

マルチモーダル AI の真価は、単に複数のデータタイプを扱えるという点に留まりません。むしろ、異なるモダリティ間に存在する関係性を深く理解し、それらを統合して新たな洞察や表現を生み出す能力にこそ、その本質があります。例えば、単に画像に写っているオブジェクトをリストアップするだけでなく、画像全体の雰囲気（例:「楽しそう」「厳粛」）を捉え、その雰囲気に合ったテキスト（詩、キャッチコピー、説明文など）を生成することや、音声対話において、話されている言葉の内容だけでなく、声のトーンや抑揚から感情を読み取り、それに応じた共感的な応答を生成するといった応用が考えられます。GPT-4o のような統合型マルチモーダルモデルは、モダリティ間の相互作用をより深く捉える潜在能力を持っていると考えられ [8]、「この画像のレトロな雰囲気に合わせて、製品紹介文を書いてください」や「顧客の不満そうな声色を考慮して、丁寧な謝罪文を作成してください」といった、より高度でニュアンスに富んだ指示への対応が期待されます。このような能力を最大限に引き出すためには、モダリティ間の連携を意識した、より洗練されたプロンプト設計やアプリケーションアーキテクチャの検討が求められます。

4. まとめ

Azure OpenAI Service は、多様な性能、価格、機能を持つ先進的な AI モデル群を提供しており、利用者は自身の特定の要件と制約に基づいて最適なモデルを選択することが可能です。

- モデル選択のポイント:
  - **GPT-4o ファミリー (GPT-4o, GPT-4o mini):** 最新世代であり、テキスト、画像、音声（専用モデル経由）を扱えるマルチモーダル機能と高い性能が特徴です。特に GPT-4o mini は、優れたコストパフォーマンスにより、多くのユースケースで魅力的な選択肢となります。
  - **o シリーズ (o1, o3, o4-mini など):** 複雑な推論、問題解決、コーディングタスクに特化しており、高度なエージェント機能や専門分析に適しています。画像入力にも対応します。
  - **GPT-4 Turbo (with Vision):** 依然として強力なモデルであり、特に 128K という広大なコンテキストウィンドウと画像入力機能が必要な場合に有効です。
  - **GPT-3.5 Turbo:** コスト効率と応答速度を最優先する場合に適した、実績のある選択肢です。
- 評価軸: モデル選択にあたっては、以下の点を総合的に評価することが重要です。
  - 利用料金: 入力・出力トークン単価、キャッシュ入力価格の有無、マルチモーダル機能（画像、音声）利用時の追加コスト。
  - トークン制限: アプリケーションが必要とする情報量を扱えるだけの最大入力コンテキストウィンドウ、および必要な応答長を生成できる最大出力トークン数。
  - 機能要件: テキスト処理能力（推論、創造性、指示追従性）、マルチモーダル対応（画像入力、音声入出力）、特定のタスクへの最適化（例: o シリーズの推論能力）。

- **Azure 利用の意義:** Azure OpenAI Service を利用する最大の利点は、OpenAI の先進モデルを、Azure プラットフォームが提供するエンタープライズグレードのセキュリティ、コンプライアンス、ガバナンス、監視、そして他の Azure サービスとのシームレスな統合環境下で活用できる点にあります。ただし、最新モデルの提供タイミングの遅延や、OpenAI API との応答挙動のわずかな差異には留意が必要です。

本レポートで提供した情報が、Azure OpenAI Service におけるモデル選択プロセスの一助となり、利用者のプロジェクト成功に貢献できれば幸いです。AI 技術は急速に進化しており、価格や性能、機能は今後も変化していくことが予想されるため、継続的な情報収集と評価が推奨されます。

引用文献

1. Azure OpenAI Service: pricing and differences with Copilot - Dev4Side, 4月 20, 2025にアクセス、https://www.dev4side.com/en/blog/azure-openai-service
2. Azure OpenAI Service, 4月 20, 2025にアクセス、https://azure.microsoft.com/en-us/products/ai-services/openai-service
3. Azure OpenAI Service documentation - Quickstarts, Tutorials, API Reference, 4月 20, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/
4. Azure OpenAI Service - Pricing, 4月 20, 2025にアクセス、https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/
5. Updated 2025! Azure OpenAI Service - an introduction - LicenseQ, 4月 20, 2025にアクセス、https://licenseq.com/azure-openai-service-an-introduction/
6. What is Azure OpenAI Service? - Learn Microsoft, 4月 20, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/overview
7. Maximizing Efficiency with Azure OpenAI Service for Business Solutions - ProServeIT, 4月 20, 2025にアクセス、https://www.proserveit.com/blog/introduction-to-microsoft-new-azure-openai-service
8. Azure OpenAI Service models - Learn Microsoft, 4月 20, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models
9. Azure OpenAI Service model versions - Learn Microsoft, 4月 20, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/model-versions
10. Calculating OpenAI and Azure OpenAI Service Model Usage Costs | Blog - DevRain, 4月 20, 2025にアクセス、https://devrain.com/blog/calculating-openai-and-azure-openai-service-model-usage-costs
11. Differences Between Azure OpenAI GPT-4o and OpenAI's Public GPT-4o - Learn Microsoft, 4月 20, 2025にアクセス、https://learn.microsoft.com/en-us/answers/questions/2153786/differences-between-azure-openai-gpt-4o-and-openai
12. Announcing the GPT-4.1 model series for Azure AI Foundry and GitHub

developers, 4月 20, 2025にアクセス、https://azure.microsoft.com/en-us/blog/announcing-the-gpt-4-1-model-series-for-azure-ai-foundry-developers/
13. Azure OpenAI: what are the real costs for prompts and responses? - ClearPeople, 4月 20, 2025にアクセス、https://www.clearpeople.com/blog/what-are-the-real-costs-for-generating-prompts-and-responses-in-azure-openai
14. Microsoft Azure AI: GPT-4o Mini available - schneider it management, 4月 20, 2025にアクセス、https://www.schneider.im/microsoft-azure-ai-gpt-4o-mini-available/
15. Introducing GPT-4 in Azure OpenAI Service, 4月 20, 2025にアクセス、https://azure.microsoft.com/en-us/blog/introducing-gpt4-in-azure-openai-service/
16. LLM API Pricing - BotGenuity, 4月 20, 2025にアクセス、https://www.botgenuity.com/tools/llm-pricing
17. Azure gpt-3.5-turbo Pricing Calculator | API Cost Estimation - Helicone, 4月 20, 2025にアクセス、https://www.helicone.ai/llm-cost/provider/azure/model/gpt-3.5-turbo
18. API Pricing - OpenAI, 4月 20, 2025にアクセス、https://openai.com/api/pricing/
19. Pricing - OpenAI API, 4月 20, 2025にアクセス、https://platform.openai.com/docs/pricing
20. GPT-4o mini: API Provider Performance Benchmarking & Price Analysis, 4月 20, 2025にアクセス、https://artificialanalysis.ai/models/gpt-4o-mini/providers
21. GPT-4 Turbo: API Provider Performance Benchmarking & Price Analysis, 4月 20, 2025にアクセス、https://artificialanalysis.ai/models/gpt-4-turbo/providers
22. Azure OpenAI Service - Pricing, 4月 20, 2025にアクセス、https://azure.microsoft.com/en-au/pricing/details/cognitive-services/openai-service/
23. What's new in Azure OpenAI Service? - Learn Microsoft, 4月 20, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/whats-new
24. Introducing GPT-4o: OpenAI's new flagship multimodal model now ..., 4月 20, 2025にアクセス、https://azure.microsoft.com/en-us/blog/introducing-gpt-4o-openais-new-flagship-multimodal-model-now-in-preview-on-azure/
25. Work with chat completion models - Azure OpenAI Service - Learn Microsoft, 4月 20, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/chatgpt
26. Azure/multimodal-ai-llm-processing-accelerator - GitHub, 4月 20, 2025にアクセス、https://github.com/Azure/multimodal-ai-llm-processing-accelerator
27. Azure OpenAI cost went to $163/day with no real usage - any way I can dig deeper into the costs?, 4月 20, 2025にアクセス、https://www.reddit.com/r/AZURE/comments/1gpmrz7/azure_openai_cost_went_to_163day_with_no_real/
28. Title: Unexpected $50K Azure Bill for OpenAI Service Used for Only an Hour - Reddit, 4月 20, 2025にアクセス、

https://www.reddit.com/r/AZURE/comments/1g0mkwi/title_unexpected_50k_azure_bill_for_openai/
29. Decoding Azure Open AI Costs and Capacity Challenges - YouTube, 4月 20, 2025にアクセス、https://www.youtube.com/watch?v=kFBQWFFkpiw
30. GPT 3.5 and GPT 4 Turbo 1106 Question for Output Context Length - API, 4月 20, 2025にアクセス、https://community.openai.com/t/gpt-3-5-and-gpt-4-turbo-1106-question-for-output-context-length/598903
31. Announcing the o1 model in Azure OpenAI Service: Multimodal reasoning with “astounding” analysis, 4月 20, 2025にアクセス、https://azure.microsoft.com/en-us/blog/announcing-the-o1-model-in-azure-openai-service-multimodal-reasoning-with-astounding-analysis/
32. Azure OpenAI Service Announces Multimodal Innovations at Microsoft Build 2024, 4月 20, 2025にアクセス、https://techcommunity.microsoft.com/blog/azure-ai-services-blog/announcing-multimodal-innovations-in-generative-ai-with-azure-openai-service-mic/4146804
33. GPT-4o Now Generally Available in Azure: Revolutionizing AI with Multimodal Capabilities, 4月 20, 2025にアクセス、https://www.serverless-solutions.com/gpt-4o-now-generally-available-in-azure-revolutionizing-ai-with-multimodal-capabilities/
34. The new gpt-4o-audio-preview in Azure OpenAI is awesome! But how much will it actually cost me? - Clemens Siebler's Blog, 4月 20, 2025にアクセス、https://clemenssiebler.com/posts/azure-openai-gpt4o-audio-api-cost-analysis/
35. Announcing new multi-modal capabilities with Azure AI Speech | Microsoft Community Hub, 4月 20, 2025にアクセス、https://techcommunity.microsoft.com/blog/azure-ai-services-blog/announcing-new-multi-modal-capabilities-with-azure-ai-speech/4144400
36. Comparing OpenAI vs. Azure OpenAI Services - Private AI, 4月 20, 2025にアクセス、https://private-ai.com/en/2024/01/09/openai-vs-azure-openai/