OpenAI APIパラメータおよび設定可能範囲の変更履歴に関するレポート（2023年1月～2025年4月）

1. はじめに

1.1. 目的

本レポートは、2023年1月1日から2025年4月25日までの期間におけるOpenAI APIのパラメータ、設定可能な値の範囲、および関連するモデルの進化を詳述することを目的とします。この分析は、急速な発展を遂げたこの期間におけるAPIの変遷を理解するための技術的参照資料として提供されます。

1.2. 焦点

分析は、公式の変更履歴および関連資料 [1] に記録されている具体的な変更点に焦点を当てます。主要な調査領域には、パラメータの追加・廃止、値の範囲（例：トークン制限、temperature設定）の変更、API利用に影響を与えるモデルの導入・廃止が含まれます。

1.3. 重要性

これらの履歴的な変更点を理解することは、既存アプリケーションを維持する開発者、移行を計画する開発者、新機能を評価する開発者、そして将来のAPIの方向性を予測する開発者にとって極めて重要です。レビュー対象期間は、OpenAIプラットフォームが著しい成熟期を迎えた時期であり、複雑性、能力、開発者による制御可能性の向上が特徴です。

1.4. 方法論

本レポートの情報は、提供されたリサーチスニペットから統合され、公式のOpenAIおよびAzure OpenAIのドキュメント、変更履歴、発表に重点を置いています。分析は、要求された通り、パラメータと範囲の変更を優先します。

2. コアテキスト生成API（CompletionsおよびChat Completions）の進化

2.1. レガシー**Completions API**からの移行

当初、text-davinci-003のようなモデルを使用する旧Completions API (/v1/completions) が主要なインターフェースでした。しかし、OpenAIは戦略的にChat Completions API (/v1/chat/completions) へと重点を移しました。

OpenAIは、ada、babbage、curie、davinci、text-davinci-003などの旧Completionモデルおよびedits APIの廃止を発表し、2024年1月4日を廃止日と設定しました [19]。Completions API自体も「レガシー」と位置づけられ、将来の開発はChat Completionsに集中することになりました [20]。これらのレガシーベースモデルに基づくファインチューニング済みモデルも移行が

必要となりました [19]。

多くのレガシーinstructモデルの代替として、gpt-3.5-turbo-instructが指定されました [19]。この動きは、OpenAIのAPI戦略における根本的な転換を示しています。より構造化され、対話に適したChat Completions形式に開発努力を集約することで、モデルのパフォーマンス向上（Chatモデルは複数ターンの対話に最適化されている）、APIサーフェスの単純化、そしてメッセージロール構造を通じた安全性・アライメントの強化を意図していたと考えられます。Completions APIを明確に「レガシー」と位置づけ [20]、関連モデルを大規模に廃止したこと [19] は、この戦略的転換を裏付けています。代替としてgpt-3.5-turbo-instruct [20] を導入・推奨したことは、instructパラダイムに慣れたユーザーへの橋渡しを図りつつも、新しいAPI構造への移行を促す試みであったと解釈できます。この統合により、OpenAIのメンテナンス負担が軽減され、開発者の注意が推奨されるインタラクションモデルに集中することになります。

2.2. Chat Completions (v1/chat/completions) パラメータの進化

Chat Completions APIは、OpenAIプラットフォームの中核として急速に進化し、開発者がモデルの出力をより細かく制御できるように、多くのパラメータが追加・変更されました。

- **コアサンプリングパラメータ:** temperature（範囲0〜2）とtop_p（範囲0〜1）は、ランダム性と決定性のバランスを取るための主要な制御手段であり続け、通常はどちらか一方のみを変更することが推奨されました [32]。frequency_penaltyとpresence_penalty（範囲-2.0〜2.0）は、繰り返しを制御するためのより細かい調整を提供しました [32]。これらの基本的なパラメータは、テキスト生成の基本的な挙動を調整するために不可欠です。
- **Function CallingからToolsへ:** モデルが外部ツールやAPIを呼び出すことを可能にするfunctionsパラメータが導入されました（Azureでは2023年7月1日プレビューで追加されたことが示唆されています [1]）。しかし、このパラメータは2023年11月に、より汎用的なtoolsパラメータに置き換えられる形で廃止されました [2]。この進化は、単なる関数呼び出しから、より広範なツール連携へと概念が拡張されたことを示唆しています。functionsからtoolsへの移行は、モデルが外部システムと対話する方法について、より抽象的で拡張可能なアプローチを採用したことを意味します。
- **強化されたツール制御:** 特定のツールの使用を強制したり、ツール呼び出しを必須にしたりできるtool_choiceパラメータが追加されました（初出はAssistants APIの文脈で2024年4月 [4] ですが、概念的には適用可能です。[2]はtool_choice: "required"が2024年4月に追加されたと言及）。parallel_tool_calls（並列関数呼び出し）のサポートが追加され（Azure API 2024-09-01プレビューで言及 [1]）、それを無効にするオプション(parallel_tool_calls=false) も2024年6月に導入されました [2]。これらのパラメータは、開発者が複雑なツールインタラクションを管理するための具体的なニーズに応えるものです。実用的な使用を通じて発見された、複数のツールを効率的に管理する必要性がparallel_tool_calls [1] を生み、ツールの使用を強制または防止する必要性がtool_choice [4] をもたらしました。
- **Log Probabilities:** logprobs（ブール値）およびtop_logprobs（整数）パラメータが追加

され（2023年12月 [2]、Azure APIでは2024年3月に言及 [11]）、モデルのトークン選択とその確信度に関するより詳細な洞察を提供できるようになりました [35]。これは、モデルの内部動作を理解し、出力を分析する上で役立ちます。

- 出力フォーマット: 特定の出力構造を強制するためのresponse_formatパラメータが導入されました。当初はJSONモード用の{"type": "json_object"}をサポートしていましたが [32]、後にStructured Outputs機能により{"type": "json_schema", "json_schema": {...}}をサポートするよう強化されました（2024年8月ローンチ [2]、gpt-4o-2024-08-06 [6] およびo3-mini [10] がサポート）。これにより、信頼性の高い構造化データの抽出が大幅に改善されました。信頼性の低いJSON出力の問題は、response_format [32] の導入、そして最終的にはJSONスキーマをサポートする構造化出力 [2] へと進化することで解決されました。
- トークン制限: max_tokensパラメータ [2] は、生成されるテキストの長さを制限する標準的な方法でした。しかし、oシリーズモデルの導入に伴い、これらのモデルではmax_tokensが機能しないため [1]、max_completion_tokensが追加されました（Azure API 2024-09-01プレビューで言及 [1]）。max_completion_tokensは、表示される出力トークンと内部的な推論トークンの両方を考慮します [7]。max_tokensとmax_completion_tokens [1] の分離は、新しいモデルアーキテクチャ（oシリーズ）が内部的な「推論」ステップに対して異なる計算方法を必要とする直接的な結果です。
- ストリーミング時の使用量: stream_options: {"include_usage": true}パラメータが追加され（2024年5月 [2]、Azure API 2024-09-01プレビューで言及 [1]）、開発者はストリーミング応答中にトークン使用量の統計情報を受け取ることができるようになり、リアルタイムのコスト監視が改善されました。
- 再現性: seedパラメータが導入され（2023年11月のDevDayの文脈で言及 [36]、ファインチューニングでは2024年4月 [2]、Azureファインチューニングでは2024年5月 [1]）、より再現性の高い出力が可能になり、デバッグやテストが容易になりました。非決定的な挙動のデバッグは、seed [36] の導入によって支援されました。

これらのパラメータの進化は、開発者により細かい制御（ツール選択、並列呼び出し、構造化出力）、透明性の向上（logprobs、ストリーミング使用量）、そして新しいモデル能力（推論トークン、特定の最大トークンパラメータ）への対応という明確な傾向を示しています。

2.3. モデルの進化と能力（Chat Completions）

Chat Completions APIで利用可能なモデルは、この期間中に目覚ましい進化を遂げました。

- **GPT-3.5**シリーズ: gpt-3.5-turboが主力モデルとなり、様々なスナップショット（0301, 0613, 1106, 0125）がリリースされました [11]。コンテキストウィンドウは拡大し、後のバージョンでは16kに標準化されました [27]。古いスナップショットは廃止の対象となりました [19]。
- **GPT-4**シリーズ: GPT-4 APIは2023年7月に一般提供が開始されました [20]。初期バージョン（0314, 0613）は8k/32kのコンテキストウィンドウを持っていました [11]。GPT-4 Turbo（1106-preview, 0125-preview, turbo-2024-04-09）は、より大きなコンテキスト（128k）と低価格を提供しました [2]。gpt-4-vision-previewにより視覚能力が追加されました [2]。古

いGPT-4スナップショットとvisionプレビューは廃止されました[11]。

- **GPT-4o:** 2024年5月にリリースされ[2]、フラッグシップモデルとなりました。テキスト、視覚、音声をネイティブに扱うマルチモーダルであり、GPT-4 Turboよりも高速かつ手頃な価格です[2]。コンテキストウィンドウは128kトークン、最大出力は16,384トークンです[11]。様々なスナップショット（2024-05-13, 2024-08-06, 2024-11-20）がリリースされました[2]。費用対効果の高い代替としてgpt-4o-miniが2024年7月にリリースされました[2]。Structured Outputs[2]や、後のバージョンでの最大出力トークンの増加[11]がサポートされました。GPT-4o[2]は、ネイティブなマルチモーダル性と効率性への大きなアーキテクチャシフトを代表するモデルです。
- **GPT-4.5 Preview:** 2025年2月にリサーチプレビューとしてリリースされました[2]。2025年7月14日に廃止が発表されました[9]。より大規模で、潜在的によりニュアンスのある応答が可能でしたが、GPT-4.1よりも費用対効果は低いと位置づけられました。
- **GPT-4.1シリーズ:** 2024年4月にローンチされました[3]。gpt-4.1、gpt-4.1-mini、gpt-4.1-nanoが含まれます。主な特徴は、最大100万トークンのコンテキストウィンドウ、改善された長文脈理解、増加した最大出力トークン（gpt-4.1で32,768トークン）、向上したコーディング/指示追従能力です[2]。GPT-4oの後継であり、GPT-4.5 Previewの代替と位置づけられました[9]。GPT-4.1[9]は、コンテキスト長の限界とコーディング能力を押し広げ、以前のイテレーションからの学びを取り込み、新しいフラッグシップ汎用モデルとしての地位を確立しました。これが実験的なGPT-4.5[9]の廃止計画につながりました。
- **oシリーズ（推論モデル）:** 2024年9月にo1-previewとo1-miniで導入が開始されました[1]。複雑な推論（数学、科学、コーディング）向けに設計されています。reasoning_effortパラメータ[1]とdeveloperメッセージロール[2]が導入されました。max_tokensの代わりにmax_completion_tokensが必要でした[1]。その後、o1-pro（2025年3月[2]）、o3-mini（2025年1月[6]）、そしてo3 / o4-mini（2025年4月[2]）が登場しました。oシリーズ[1]は、特定の推論の弱点に対処するために分岐しました。

モデルランドスケープは、加速するイテレーション、マルチモーダルへの移行（gpt-4o）、コンテキスト処理能力の大幅な向上（gpt-4.1）、特化（推論用のoシリーズ）、そしてプレビュー/スナップショットモデルのリリースとその後の旧バージョンの安定化・廃止という明確なパターンを示しています。これは、開発者が変化に対応し、コスト、能力、必要なコンテキスト長に基づいてモデルを慎重に選択する必要があることを意味します。この進化は、コア製品の改善、要求の厳しいタスクへの特化、そしてパフォーマンス/効率の限界を常に押し上げるという多面的な戦略を反映しています。

3. 特化型およびエージェント指向APIの出現と発展

テキスト生成の中核機能に加え、OpenAIはより複雑なタスクやエージェント的な動作を可能にするための特化型APIを導入・発展させました。

3.1. Assistants API (v1 -> v2)

Assistants APIは、状態を持つエージェントのような体験を構築するために設計されました。

- **導入:** 2023年11月にローンチされ [2]、会話スレッドを管理し、Code InterpreterやRetrievalといったツールを利用できるステートフルなAPIとして登場しました [36]。これは、OpenAIによるステートフルなツール利用エージェントフレームワークの最初の主要な試みでした。
- **v1 Beta:** 初期のリリースはコア機能を提供しましたが、開発者からは複雑さに関するフィードバックがありました [4]。
- **v2 Beta (2024年4月):** 大幅な機能強化が行われました [4]。
  - Retrievalに代わる新しいfile_searchツール（最大10,000ファイル対応、旧版は20）。
  - ファイル管理のためのvector_storeオブジェクトの導入。
  - 実行ごとの最大トークン数を制御するパラメータ（max_completion_tokens, max_prompt_tokens）。
  - tool_choiceパラメータのサポート。
  - assistantロールメッセージの追加機能。
  - コアパラメータ（temperature, top_p, response_format）のサポート。
  - ファインチューニング済みモデルのサポート（当初はgpt-3.5-turbo-0125）。
  - ストリーミングサポートの追加（2024年3月にも言及あり [2]）。
  - ストリーミング/ポーリング用のSDKヘルパー。ベータフィードバックに基づくv2 [4] の迅速なリリースは、応答性を示す一方で、初期設計の複雑さを示唆している可能性があります。
- **その後のアップデート:** 画像入力のサポート（2024年5月 [2]）、ファイル検索のカスタマイズ/ランキング（2024年6月/8月 [2]）。
- **廃止計画:** 2025年3月に発表され [2]、v1は2024年12月18日にシャットダウン [21]、v2はResponses APIが機能的に同等になった後、2026年前半に廃止予定です。v1からv2への急速な進化と、その後のResponses APIを優先するための廃止計画は、複雑なAPI設計に対するOpenAIの反復的なアプローチを浮き彫りにしています。ベクトルストアと強化されたファイル検索の導入は、外部知識でモデルを補強することの重要性を示しています。

3.2. Responses API

Responses APIは、Assistants APIからの学びを活かし、よりシンプルで柔軟なエージェント構築を目指して導入されました。

- **導入:** 2025年3月にローンチされ [1]、Assistantsよりもシンプルで柔軟なエージェントワークフローのための新しいプリミティブとして登場しました。これは、ステートレスなChat Completions（手動での履歴管理が必要 [42]）とステートフルなAssistants API（フィードバックあり [4]）の両方の制限と潜在的な複雑さから生じたニーズを満たすことを目指しています。

- コアコンセプト: Chat Completionsのシンプルさと、組み込みツールの利用および状態管理（previous_response_id経由[13]）を組み合わせます。オプションの状態管理(previous_response_id) と共通ツールの直接統合を提供することで、開発者の定型コードを削減します。
- 組み込みツール: Web Search、File Search、Computer Use（computer-use-previewモデル/ツール）と共にローンチされました[1]。Code Interpreterも計画されています[4]。
- **Assistants APIとの関係:** 将来の方向性として位置づけられ[4]、Assistants API廃止前に機能的な同等性を目指しています[4]。システムメッセージと同様のinstructionsパラメータを使用します[13]。Assistantsの廃止計画[4]と同時に導入されたことは、この新しいインタラクションモデルへの明確な戦略的方向性を示しています。

これは、エージェントAPI設計に関するOpenAIの最新の考え方を表しており、使いやすさを優先し、強力な組み込みツールを直接統合しています。Chat Completions固有の状態管理の問題を、Assistants APIのThread/Run構造のオーバーヘッドなしに解決しようとしています。

3.3. Audio API（Speech-to-Text & Text-to-Speech）

高品質な音声処理機能がAPIに統合されました。

- 導入: Speech-to-Text（Whisper）は2023年9月[1]、Text-to-Speech（TTS）は2024年2月に追加されました[1]。OpenAIは統合音声処理の需要を認識し、既存のWhisper研究をSTTに活用しました[1]。
- モデル: 初期のWhisperモデル（whisper-1[2]）。TTSモデルも導入されました。後に、gpt-4oを活用した音声モデルが追加されました：gpt-4o-audio-preview（2024年10月[2]）、gpt-4o-transcribe、gpt-4o-mini-transcribe、gpt-4o-mini-tts（2025年3月[2]）。重要な進化は、これらの機能をフラッグシップのマルチモーダルモデルであるGPT-4o[2]と統合したことで、よりシームレスな音声入出力アプリケーションを可能にし、潜在的により良い文字起こし/音声生成のためにGPT-4oの基盤となる知能を活用しました。
- パラメータ:
  - STT: timestamp_granularitiesが追加されました（2024年2月[2]、Azureでは2024年4月[1]）。AzureではaudioWordオブジェクトが追加されました[1]。
  - TTS: wav、pcmなどの追加のresponse_formatsが追加されました（Azureでは2024年4月[1]）。新しい音声が追加されました（2024年10月[2]）。timestamp_granularities[2]や新しいフォーマット[1]のようなパラメータ追加は、よりリッチな音声処理ワークフローに対する開発者のニーズに基づく標準的な改良です。

高品質な音声機能の迅速な統合は、特化モデル（Whisper）とGPT-4oのマルチモーダル能力の両方を活用しています。パラメータの追加は、よりリッチな出力データ（タイムスタンプ）とフォーマットの柔軟性を提供することに焦点を当てています。

3.4. Image Generation API

テキストだけでなく、画像の生成もAPIを通じて可能になりました。

- **導入:** DALL-E 3のサポートが追加されました（Azureでは2023年12月[1]、OpenAIでは2023年11月[2]）。DALL-Eの成功を受けて、画像生成を開発者プラットフォームに統合することは自然な流れでした[1]。
- **モデルアップデート:** gpt-image-1モデルが2025年4月に追加され[2]、APIがChatGPTの能力と整合しました[6]。このモデル固有の新しいパラメータも導入されました[2]。Azureプレビュー APIには2024年5月にDALL-E 2のサポートが追加されました[1]。最新のChatGPT画像モデルを使用するように更新されたこと[2]は、APIユーザーが利用可能な最高の技術にアクセスできるようにし、消費者向け製品との同等性を維持します。

これにより、テキストのみのインタラクションを超え、最先端の画像生成がAPIプラットフォームに直接統合されました。

3.5. Embeddings API

セマンティック検索やRAG（Retrieval-Augmented Generation）に不可欠なEmbeddings APIも進化しました。

- **レガシーモデル:** 古い埋め込みモデル（類似性、検索、コード検索のバリアント）は2024年1月4日に廃止されました[20]。OpenAIは当初、多くの特化型埋め込みモデルを持っていました[20]。
- **text-embedding-ada-002:** 標準モデルとなりました[15]。text-embedding-ada-002[20]に統合することで、提供が簡素化されました。
- **V3モデル:** 新しく、より高性能な埋め込みモデル（text-embedding-3-small, text-embedding-3-large）が2024年1月にリリースされました[2]。V3モデル[2]のリリースは、パフォーマンス/コストの改善を提供しました。
- **パラメータ:** encoding_format（2023年10月[2]）およびdimensions（2024年1月[2]、Azureでは2024年3月[1]）パラメータが追加され、出力フォーマットとベクトル次元（V3モデル用）の制御が可能になりました。dimensions[2]の追加により、開発者は埋め込みサイズ/コストと精度のトレードオフを制御できるようになり、これは本番システムにおける重要な要素です。encoding_format[2]は、異なる下流システムに対する柔軟性を追加します。

埋め込みモデルのパフォーマンスと費用対効果が継続的に改善され、新しいパラメータによって柔軟性が追加されました。特にdimensionsパラメータは、ストレージと検索パフォーマンスの最適化に重要です。

3.6. Fine-Tuning API

モデルを特定のユースケースに合わせて調整するFine-Tuning APIも機能が拡張されました。

- **進化:** 元の/v1/fine-tunesエンドポイントは2024年1月4日に廃止され[21]、/v1/fine_tuning/jobs[23]に置き換えられました。新しいAPI[23]は、より良いインフラストラク

チャと機能を提供した可能性があります。

- モデルサポート: 当初はベースGPT-3モデル（ada-002, babbage-002, curie-002, davinci-002）に焦点を当てていましたが [19]、サポートはgpt-3.5-turbo [20]、Function Calling（2023年10月 [2]）、gpt-4o（GA 2024年8月 [2]）、gpt-4o-mini（2024年7月 [2]）、および画像/視覚ファインチューニング（2024年10月 [2]、GA 2024年11月 [11]）へと拡張されました。babbage-002/davinci-002でのファインチューニングトレーニングは2024年10月に廃止されました [23]。Direct Preference Optimization (DPO) が2024年12月に追加されました [2]。gpt-3.5-turbo [20] および後のgpt-4o [2] へのサポート拡張は、高性能モデルのカスタマイズを可能にするために重要でした。Function Calling [2] およびVision [2] のサポート追加は、これらの高度な機能をファインチューニングする必要性を反映しています。DPO [2] は、新しいアライメント技術の採用を表しています。
- パラメータ: seed（2024年4月 [2]、Azure 2024年5月 [1]）、checkpoints（2024年4月 [2]、Azure 2024年5月 [1]）、metadata（2025年3月 [1]）のサポートが追加されました。seed、checkpoints、metadata [1] のようなパラメータは、本格的な開発ワークフローにおけるプロセスの管理性と再現性を向上させます。

ファインチューニング能力は大幅に拡張され、基本的なモデル適応から、新しいアーキテクチャ（GPT-3.5, GPT-4o）、高度な機能（Function Calling, Vision）のサポートへと移行し、より多くの制御/可観測性（seed, checkpoints, metadata）を提供しています。

3.7. Moderations API

コンテンツの安全性を確保するためのModerations APIも更新されました。

- アップデート: omni-moderation-latestモデルが2024年10月にリリースされ、画像とテキストのモデレーションをサポートし、精度が向上しました [2]。OpenAIモデルがマルチモーダル（GPT-4 Vision, DALL-E, GPT-4o）になるにつれて、異なる入力タイプにわたる安全性を確保するためにマルチモーダルモデレーション（omni-moderation-latest [2]）の必要性が明らかになりました。
- パラメータ変更: max_tokensパラメータが追加されました（2023年10月 [2]）。注意：[2]はこのパラメータを*2023*年*10*月に追加されたとしていますが、コンテンツを分類し、トークン制限付きで生成するわけではない*Moderations*エンドポイントにとっては非常に珍しいように思われます。これは[2]の要約における誤解釈またはエラーである可能性があります。主要な機能は分類であり、生成ではありません。このmax_tokens [2] の追加は不可解であり、注意が必要です。これは生成制限ではなく処理制限に関連する可能性があります。

モデレーション能力は、プラットフォーム全体のマルチモーダルシフトを反映して、マルチモーダルへと進化しています。

3.8. Realtime API

リアルタイムでの音声対話を実現するAPIが導入されました。

- 導入: 2024年10月にローンチされ [2]、WebSocketsを介した低遅延の音声対音声通信を提供します。標準的なREST APIには、リアルタイム音声に適さない遅延制限があります。Realtime API [2] はWebSockets [2] を使用してこの問題に直接対処します。
- 機能強化: WebRTC接続メソッドが追加されました（2024年12月 [2]）。新しい音声が追加されました（2024年10月 [2]、2024年12月 [11]）。プロンプトキャッシングのサポートが追加されました（2024年12月 [11]）。gpt-4o-mini-realtime-previewが追加されました [11]。レート制限が接続数/分からRPM/TPMに変更されました [11]。WebRTC [2] の追加は、リアルタイム通信のための別の一般的なプロトコルを提供します。より多くの音声 [2]、キャッシングサポート [11]、ミニバージョン [11] の提供は、魅力を広げ、パフォーマンス/コストを最適化するための典型的な改良です。レート制限の変更 [11] は、課金/スロットリングを標準化します。

これにより、音声アシスタントやリアルタイム翻訳などのアプリケーションに不可欠な、応答性の高い音声インタラクションのニーズに対応します。

3.9. Batch API

非同期処理のためのBatch APIが導入されました。

- 導入: 2024年5月～7月に追加され（Azure [1]）、2024年5月にはファインチューニング済みモデルのサポートが追加されました（OpenAI [2]）。大規模なワークロードを低コストで非同期に処理できます [37]。すべてのAPI使用がリアルタイムインタラクションを必要とするわけではありません。大規模なデータ分類や要約などのタスクには、非同期バッチ処理 [1] がより効率的で費用対効果が高いです。
- 特徴: 即時応答が不要な非対話型の大規模タスクに対応し、大幅なコスト削減（50%割引 [37]）を提供します。ファインチューニング済みモデル [2] のサポートにより、汎用性が向上しています。

4. パラメータ範囲と能力の重要なシフト

APIパラメータの範囲やモデルの基本的な能力にも、注目すべき変化が見られました。

4.1. コンテキストウィンドウの拡張

モデルが一度に処理できる情報量（コンテキストウィンドウ）は、劇的に増加しました。

- 傾向: より大きなコンテキストウィンドウへの明確かつ重要な傾向が見られます。初期のモデルは限られたコンテキスト（例：GPT-3で2k-4k [43]）しか持っていませんでした。
- 例: GPT-4 Turboは128kコンテキストを導入しました [36]。GPT-4oは128kを維持しました [34]。GPT-4.1シリーズはこの数値を劇的に増加させ、100万トークンに達しました [2]。GPT-4は8k/32kを提供しました [11]。128kへのジャンプ [36] は大きな飛躍であり、その後の1Mトークンへの飛躍 [9] はさらに桁違いの増加です。

- 影響: これにより、はるかに大きなドキュメント、コードベース全体、またはより長い会話履歴の処理が可能になり、要約、分析、複雑な問題解決における新しいユースケースが解放されました。ただし、潜在的な処理時間とコストも増加します。
- 背景: OpenAIは、コンテキスト長を主要な差別化要因であり、より複雑なAIタスクを実現するものとして認識し、その限界を積極的に押し広げています。これは、開発者からのより大きな入力処理に対する高い需要と、OpenAIが達成した技術的な実現可能性を反映しています。

4.2. 出力トークン制限

モデルが一度に生成できる最大トークン数も増加しましたが、コンテキストウィンドウほど劇的ではありませんでした。

- 傾向: 最大出力トークン制限も増加しています。
- 例: GPT-4oは当初4,096の最大出力トークンを持っていましたが、2024-08-06バージョンでは16,384に増加しました [11]。GPT-4.1はこの数値をさらに32,768トークンに増加させました [9]。
- 影響: より冗長で完全な応答が可能になり、長文コンテンツの生成や大規模なコードファイルのリライトなどのタスクに役立ちます [9]。
- 背景: コンテキスト入力が急速に拡大する一方で、出力制限はより緩やかに増加しています。これは、能力と生成コスト/遅延のバランスを反映している可能性があります。GPT-4.1での増加 [9] は、その大きなコンテキストによって可能になったユースケース（完全なファイルリライトなど）を具体的にサポートしています。

4.3. 強化された制御パラメータ

開発者がモデルの動作をより細かく制御するためのパラメータが増加しました。

- 例: tool_choice [2]、parallel_tool_calls [1]、stream_options [1]、response_format（JSONオブジェクト/スキーマ）[2]、seed [1]、logprobs/top_logprobs [1]。
- 影響: 開発者は、モデルの実行、外部ツールとの対話、出力構造、再現性、および可観測性に対して、大幅にきめ細かい制御を得ることができます。
- 背景: これは、APIプラットフォームの成熟を反映しています。基本的な生成を超えて、信頼性、予測可能性、および特定の対話パターンを必要とする複雑な本番グレードのアプリケーションのサポートへと移行しています。開発者がより複雑なアプリケーションを構築するにつれて、彼らは限界に遭遇しました。これらのパラメータは、実世界の利用から生じる問題点や要件に直接対応しています。

4.4. 推論能力とパラメータ

複雑な推論タスクに特化したモデルとパラメータが導入されました。

- モデル: oシリーズ（o1, o1-mini, o1-pro, o3-mini, o3, o4-mini）の導入 [1]。汎用LLMは、

複雑な多段階推論に苦労することがあります。
- パラメータ: reasoning_effort [1]、developerメッセージロール [2]、個別のトークン計算（reasoning_tokens [1]、max_completion_tokens [1]）。これらは、「思考時間」を制御し [1]、内部処理ステップを計算する [1] 新しい方法を必要としました。
- 影響: 多段階思考を必要とするタスク専用のモデルとパラメータ。潜在的により高い遅延やコストがかかる可能性がありますが、複雑な問題（数学、科学、コーディング）でより良い結果を達成します。
- 背景: OpenAIは推論を改善すべき特定の分野として特定し、特化モデルを作成しました。この二分化により、ユーザーはタスクに応じて、より高速/安価な汎用モデルと、より低速/高価な特化モデルを選択でき、リソース割り当てを最適化できます。新しいパラメータは、この特化された推論プロセスに対する制御と可視性を提供します。

5. モデルライフサイクル管理（2023年1月～2025年4月）

この期間中、OpenAIは新しいモデルを継続的に導入し、古いモデルを体系的に廃止してきました。

5.1. 主要モデル導入のタイムライン

- **GPT-3.5 Turbo:** 期間初期に prominence を増し、旧GPT-3モデルを置き換え [20]。
- **GPT-4:** GA 2023年7月 [20]。
- **GPT-4 Turbo (Preview):** 2023年11月 [2]。
- **Embedding V3 Models:** 2024年1月 [2]。
- **GPT-4o:** 2024年5月 [2]。
- **GPT-4o mini:** 2024年7月 [2]。
- **o1-preview / o1-mini:** 2024年9月 [2]。
- **Realtime API / Models:** 2024年10月 [2]。
- **o3-mini:** 2025年1月 [6]。
- **GPT-4.5 Preview:** 2025年2月 [3]。
- **o1-pro:** 2025年3月 [2]。
- **GPT-4.1 Series:** 2025年4月 [3]。
- **o3 / o4-mini:** 2025年4月 [2]。

5.2. 廃止されたモデルとシャットダウン日の概要

- 主要な傾向: OpenAIは新しいモデル/バージョンをリリースし、その後、古いモデルの廃止タイムラインを発表するパターンに従います。通常、数ヶ月の通知期間が設けられます。廃止ページ [23] やAzureの廃止スケジュール [11] を通じて明確なコミュニケーションが行われます。継続的なモデル改善はOpenAIの核となる目標であり、これにより必然的に古いモデルは時代遅れになるか、効率が悪くなります。
- 例: レガシーCompletionsモデル（2024年1月 [20]）、旧Embeddingsモデル（2024年1月 [20]

)、Codexモデル（2023年3月 [23]）、特定のGPT-3.5 Turboスナップショット（2024年9月 [23]）、特定のGPT-4スナップショット（2025年6月 [11]）、GPT-4 Vision Preview（2024年12月 [23]）、GPT-4.5 Preview（2025年7月 [9]）、Assistants API v1（2024年12月 [21]）。

- **影響:** 急速なモデルイテレーションは、構造化された廃止プロセスを必要とします。開発者は、サービスの中断を避けるために、これらのスケジュールを積極的に監視し、移行を計画する必要があります。特定のモデルスナップショットに依存すると安定性が得られますが [40]、最終的には移行が必要です。一方、gpt-4oのようなフローティングエイリアスを使用すると、最新の改善にアクセスできますが、挙動の変化のリスクが伴います [40]。したがって、廃止スケジュールは、プラットフォーム上に構築するすべての人にとって重要な情報です。

5.3. 統合モデル/API廃止スケジュール表（2023年1月～2025年4月）

以下の表は、この期間に発表された主要なモデルおよびAPIエンドポイントの廃止スケジュールをまとめたものです。これは、移行計画に不可欠な参照情報となります。

モデル/APIエンドポイント	バージョン (該当する場合)	発表日 (概算)	シャットダウン日	推奨代替モデル/API	出典例
Codex Models	code-davinci-002, etc.	2023-03-20	2023-03-23	gpt-4o	[23]
Legacy Completions Models	text-davinci-003, etc.	2023-07-06	2024-01-04	gpt-3.5-turbo-instruct	[20]
Legacy Embedding Models	text-similarity-*, etc.	2023-07-06	2024-01-04	text-embedding-3-small	[20]
Edits API & Models	text-davinci-edit-001	2023-07-06	2024-01-04	/v1/chat/completions, gpt-4o	[20]
Fine-Tuning API (Legacy)	/v1/fine-tunes	2023-08-22	2024-01-04	/v1/fine_tuning/jobs	[23]
gpt-3.5-turbo-0301	0301	2023-06-13	2024-09-13	gpt-3.5-turbo	[23]

gpt-3.5-turbo-0613	0613	2023-11-06	2024-09-13	gpt-3.5-turbo	[23]
gpt-3.5-turbo-16k-0613	16k-0613	2023-11-06	2024-09-13	gpt-3.5-turbo	[23]
gpt-4-vision-preview	vision-preview	2024-06-06	2024-12-06	gpt-4o	[23]
Assistants API v1 Beta	v1	2024年4月頃	2024-12-18	Assistants API v2 / Responses API	[21]
Fine-tuning on babbage-002	babbage-002	2024-08-29	2024-10-28	gpt-4o-mini	[23]
Fine-tuning on davinci-002	davinci-002	2024-08-29	2024-10-28	gpt-4o-mini	[23]
gpt-4-32k models	32k, 32k-0613, 32k-0314	2024-06-06	2025-06-06	gpt-4o	[23] (Azure: [11])
gpt-4.5-preview	preview	2025-04-14	2025-07-14	gpt-4.1	[9]

注: シャットダウン日は変更される可能性があります。常に公式の廃止ドキュメントを参照してください。*Azure OpenAI Service*の廃止日は若干異なる場合があります [11]。

6. Azure OpenAI APIに関する注記

Azure OpenAI Serviceは、OpenAIモデルをAzureクラウド環境内で提供し、しばしば追加のエンタープライズ機能（セキュリティ、VNETなど）を備えています [1]。

- **APIバージョニング:** Azureは独自の日付ベースのAPIバージョニングスキーム（例: 2024-05-01-preview, 2024-10-21 GA）を使用しており、これは対応するOpenAI APIバージョンの機能を組み込んでいます [1]。
- **機能ラグ:** 一般的に、新しいOpenAIモデルと機能は、特にGAリリースにおいて、遅れてAzure OpenAIに登場します。プレビュー機能はより密接に連携する可能性があります [1]。

例えば、oシリーズモデル、Responses API、GPT-4.1は、OpenAIでのリリース後にAzureに登場しました [1]。
- パラメータのパリティ: コアパラメータ（temperature, max_tokens/max_completion_tokens, tool_choiceなど）は一般的に整合していますが、Azureはインフラストラクチャや機能に関連する特定のパラメータを追加する場合があります（例：user_security_context [1]、応答内のコンテンツフィルタリング詳細 [7]）。Azure APIバージョンの変更は、しばしば複数のOpenAI機能アップデートをバンドルします [1]。
- 廃止スケジュール: Azureは独自のモデル廃止スケジュールを維持しており [11]、これはOpenAIの直接的なAPIスケジュールと若干異なる場合がありますが、一般的には同じ傾向に従います [22]。

同じコアモデルを活用しているにもかかわらず、Azure OpenAI APIは、独自のリリースサイクル、バージョニング、そしてAzureのエコシステムとエンタープライズフォーカスを反映したわずかに異なるパラメータセットや機能を持つ、別個の製品として存在します。Azureを使用する開発者は、Azure固有のドキュメントとタイムラインを追跡する必要があります。

7. 結論

2023年1月から2025年4月にかけてのOpenAI APIの進化は、急速な技術革新と開発者体験への注目の高まりを特徴としています。分析から浮かび上がった主要な傾向は以下の通りです。

- エージェント指向への注力: 複雑でステートフルな、ツールを使用するエージェントを可能にするAPI（Assistants APIからResponses APIへ）への移行が進んでいます。
- マルチモーダリティ: テキストを超えて、視覚、音声（入出力）、画像生成能力が統合されました。
- コンテキスト拡張: モデルのコンテキストウィンドウが劇的に増加しました（128kから1Mトークンへ）。
- 制御と可観測性の強化: 生成、ツール利用、出力フォーマット、再現性を細かく制御するためのパラメータが急増しました。
- 急速なモデルイテレーション: 能力、効率、または特化を改善した新しいモデル（GPT-4 -> Turbo -> 4o -> 4.1; oシリーズ）が継続的にリリースされました。
- 構造化された廃止: 古いモデルとAPIを廃止するための正式なプロセスが確立されました。

これらの傾向は、開発者にとって以下の意味を持ちます。

- 適応性の必要性: 変化の速さは、廃止スケジュールの監視やアプリケーションの移行を含む、継続的な学習と適応を要求します。
- 複雑性と選択肢の増加: より多くのAPI、モデル、パラメータは、より大きなパワーを提供しますが、ユースケース、コスト、パフォーマンスのニーズに基づいて慎重な選択も必要とし

ます。
- 開発者体験への焦点: Responses APIや改善されたSDKのような取り組みは、OpenAIが複雑なAIアプリケーションの開発を簡素化することにますます焦点を当てていることを示唆しています。
- プラットフォームの成熟: APIエコシステムは大幅に成熟し、堅牢な本番対応アプリケーションを構築するのに適したツールと制御を提供しています。

OpenAI APIは、この期間を通じて目覚ましい変貌を遂げました。パラメータ、範囲、モデルの変更を理解することは、この強力なプラットフォームを効果的に活用し、将来の進化に備えるために不可欠です。

引用文献

1. Azure OpenAI Service API version lifecycle - Learn Microsoft, 4月 25, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/api-version-deprecation
2. Changelog - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/changelog
3. openai-python/CHANGELOG.md at main - GitHub, 4月 25, 2025にアクセス、https://github.com/openai/openai-python/blob/main/CHANGELOG.md
4. What's new in Assistants API - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/assistants/whats-new
5. ChatGPT — Release Notes - OpenAI Help Center, 4月 25, 2025にアクセス、https://help.openai.com/en/articles/6825453-chatgpt-release-notes
6. Announcements - OpenAI Developer Community, 4月 25, 2025にアクセス、https://community.openai.com/c/announcements/6
7. azure-sdk-for-net/sdk/openai/Azure.AI.OpenAI/CHANGELOG.md at main · Azure/azure-sdk-for-net · GitHub, 4月 25, 2025にアクセス、https://github.com/Azure/azure-sdk-for-net/blob/main/sdk/openai/Azure.AI.OpenAI/CHANGELOG.md
8. GPT Release Notes - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/gpts/release-notes
9. Introducing GPT-4.1 in the API | OpenAI, 4月 25, 2025にアクセス、https://openai.com/index/gpt-4-1/
10. Model Release Notes | OpenAI Help Center, 4月 25, 2025にアクセス、https://help.openai.com/en/articles/9624314-model-release-notes
11. What's new in Azure OpenAI Service? - Learn Microsoft, 4月 25, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/whats-new
12. Releases · openai/openai-python - GitHub, 4月 25, 2025にアクセス、https://github.com/openai/openai-python/releases
13. Introducing the Responses API - Announcements - OpenAI Developer Community, 4月 25, 2025にアクセス、https://community.openai.com/t/introducing-the-responses-api/1140929

14. OpenAI News, 4月 25, 2025にアクセス、https://openai.com/news/
15. New embedding models and API updates - Announcements - OpenAI Developer Forum, 4月 25, 2025にアクセス、https://community.openai.com/t/new-embedding-models-and-api-updates/610540
16. Introducing OpenAI o1-preview | New OpenAI Announcement - API, 4月 25, 2025にアクセス、https://community.openai.com/t/introducing-openai-o1-preview-new-openai-announcement/937861
17. GPT-4.5-preview model will be removed from the API on 2025-07-14 - Deprecations, 4月 25, 2025にアクセス、https://community.openai.com/t/gpt-4-5-preview-model-will-be-removed-from-the-api-on-2025-07-14/1230050
18. Deprecations - OpenAI Developer Community, 4月 25, 2025にアクセス、https://community.openai.com/c/api/deprecations/32
19. OpenAI Deprecation Summary, 4月 25, 2025にアクセス、https://community.openai.com/t/openai-deprecation-summary/289539
20. GPT-4 API general availability and deprecation of older models in the Completions API, 4月 25, 2025にアクセス、https://openai.com/index/gpt-4-api-general-availability/
21. OpenAI Model Deprecation Guide - Portkey, 4月 25, 2025にアクセス、https://portkey.ai/blog/openai-model-deprecation-guide/
22. Azure OpenAI Service model deprecations and retirements - Learn Microsoft, 4月 25, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/model-retirements
23. Deprecations - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/deprecations
24. Should GPT4 be depreciated by now? - Deprecations - OpenAI Developer Community, 4月 25, 2025にアクセス、https://community.openai.com/t/should-gpt4-be-depreciated-by-now/1139509
25. Any plans to deprecate the GPT-4 Turbo preview models? - API - OpenAI Developer Forum, 4月 25, 2025にアクセス、https://community.openai.com/t/any-plans-to-deprecate-the-gpt-4-turbo-preview-models/739312
26. How is gpt-3.5-turbo-0613 still available in the API when its shutdown date was set to 2024-09-13? - Deprecations - OpenAI Developer Forum, 4月 25, 2025にアクセス、https://community.openai.com/t/how-is-gpt-3-5-turbo-0613-still-available-in-the-api-when-its-shutdown-date-was-set-to-2024-09-13/1230965
27. Is gpt-3.5-turbo-16k being deprecated? - API - OpenAI Developer Community, 4月 25, 2025にアクセス、https://community.openai.com/t/is-gpt-3-5-turbo-16k-being-deprecated/932563
28. Are all gpt-3.5-turbo versions getting deprecated? - OpenAI Developer Forum, 4月 25, 2025にアクセス、

https://community.openai.com/t/are-all-gpt-3-5-turbo-versions-getting-deprecated/923634
29. deprecation date for the GPT-3.5-turbo - OpenAI Developer Forum, 4月 25, 2025にアクセス、https://community.openai.com/t/deprecation-date-for-the-gpt-3-5-turbo/934002
30. Pleading with Openai Developers to not retire gpt-3.5-turbo-0613 on June 13th, 4月 25, 2025にアクセス、https://community.openai.com/t/pleading-with-openai-developers-to-not-retire-gpt-3-5-turbo-0613-on-june-13th/804830
31. When gpt-3.5-turbo-0613 is shutdown will my fine-tune shutdown also? - Deprecations, 4月 25, 2025にアクセス、https://community.openai.com/t/when-gpt-3-5-turbo-0613-is-shutdown-will-my-fine-tune-shutdown-also/742112
32. API Reference - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/api-reference/chat/create
33. Models - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/models/overview
34. Model - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/models/gpt-4o
35. Azure OpenAI Service REST API reference - Learn Microsoft, 4月 25, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/reference
36. Automate Blog Post writing? - API - OpenAI Developer Community, 4月 25, 2025にアクセス、https://community.openai.com/t/automate-blog-post-writing/410486
37. API Platform - OpenAI, 4月 25, 2025にアクセス、https://openai.com/api/
38. Complete Guide to the OpenAI API 2025 | Zuplo Blog, 4月 25, 2025にアクセス、https://zuplo.com/blog/2025/04/10/openai-api
39. Models - OpenAI API, 4月 25, 2025にアクセス、https://platform.openai.com/docs/models
40. Dated model GPT-4o-2024-05-13 vs updated model GPT-4o - OpenAI Developer Forum, 4月 25, 2025にアクセス、https://community.openai.com/t/dated-model-gpt-4o-2024-05-13-vs-updated-model-gpt-4o/918612
41. OpenAI API: Overview, 4月 25, 2025にアクセス、https://platform.openai.com/
42. How to pass conversation history back to the API - OpenAI Developer Forum, 4月 25, 2025にアクセス、https://community.openai.com/t/how-to-pass-conversation-history-back-to-the-api/697083
43. Azure OpenAI Service deprecated models - Learn Microsoft, 4月 25, 2025にアクセス、https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/legacy-models