2026 年 6 月 15 日　全 10 頁

# エージェント化が迫る AI コストの二極化

## 高度な業務を任せられる AI ほど、高いコストが求められる時代に

経済調査部　AI アナリティックリサーチ室　主任研究員　新田 尭之
AI アナリティックリサーチ室　主任研究員　田邉 美穂

[要約]

■ 生成 AI の料金体系は、足元で二極化が進みつつある。無料・低価格のチャット機能は残る一方、高性能モデルや業務自動化機能では、上位プランや従量課金を求める動きが広がっている。直近では、米 Anthropic 社が 6 月 9 日に提供を開始した「Claude Fable 5」について従量課金への移行を予定していた（ただし 6 月 12 日、米国政府の命令により提供停止中）。軽い用途は低コストに抑え、高負荷・高付加価値な用途には相応の対価を求める構造への転換が進んでいる。

■ この背景の一つとして、AI エージェントの利用の急拡大が指摘できる。AI エージェントは、手順の検討、検索、プログラムコード実行、やり直しなどを内部で繰り返すため、利用者から見える成果物は一つでも、AI が処理する情報量が大きく膨らみやすい。加えて、AI サービス提供企業は AI インフラへの巨額投資を回収する必要に迫られており、今回の料金体系の変更に踏み切ったと考えられる。

■ 一方、ユーザー側の企業も AI サービスに支払ってよいと考える金額が高まっているとみられる。AI が担える仕事の幅が広がるにつれ、企業にとっての AI の比較対象は「ツール」から「人間の労働者」へ移行し、高性能モデルやエージェント機能への高い支払いが合理的と判断されやすくなっている。

■ 日本企業は生成 AI を単なるチャットツールではなく、業務プロセスを支える外部計算資源として認識する必要がある。そのうえで、利用量の可視化、用途に応じたモデルの使い分け、外部 AI サービスと自社管理基盤の最適な組み合わせを通じて、費用対効果を踏まえた管理体制を構築することが求められる。

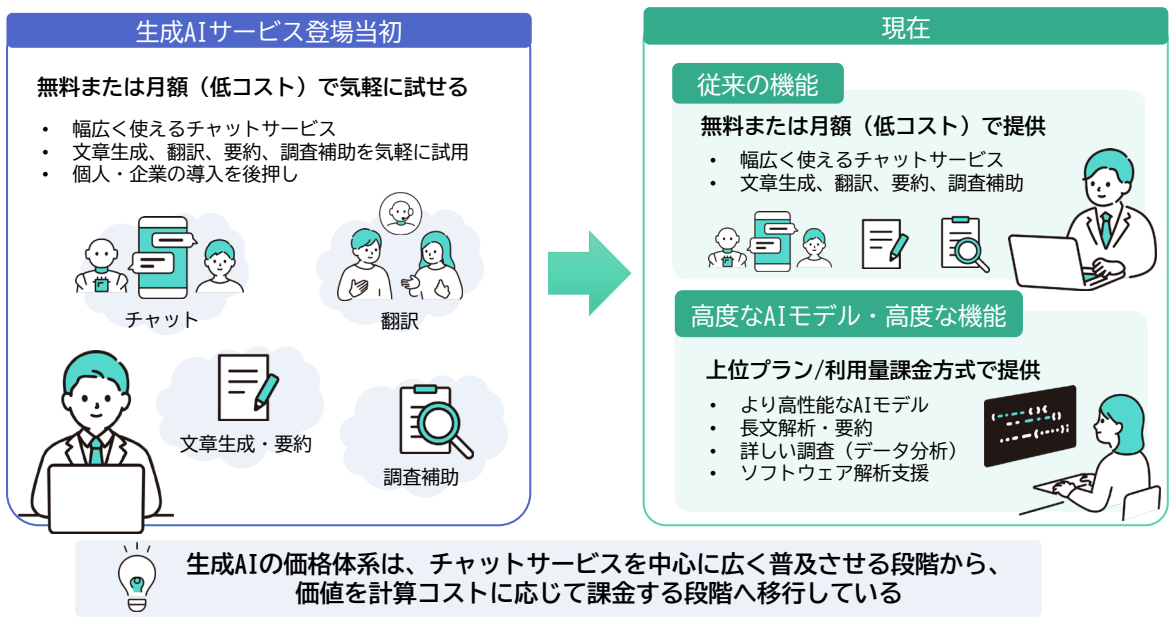
## 1. 料金体系は「使い放題」から二極化・従量化へ

### 「軽いチャット」と「重い作業」の扱いが分かれ始めた

生成 AI は当初、米 OpenAI 社の ChatGPT の無料版や月額 20 ドル前後の個人向け定額プランを通じて急速に普及した。その後登場した米 Anthropic 社の Claude や、米 Google 社の Gemini も同様に、無料プランと同価格帯の有料プランを併設し、一時期には主要サービス間で手ごろな料金体系が定着したようにみえた。こうした料金体系のもと、文章作成、翻訳、調査の補助、プログラミングなどに生成 AI を気軽に試せる環境が整い、利用は着実に広がっていった。

しかし足元ではこうした料金体系に変化が生じつつある。各社は無料・低価格のチャット機能を残しつつ、高性能モデルの利用、長文処理、高度な調査、コーディングエージェントを用いたソフトウェア開発支援などについては、上位の料金プランへの加入や利用量に応じた従量課金を求めるようになっており、こうした動きは業界全体に広がっている（**図表 1**）。

**図表 1　直近の生成 AI 料金体系の変化**




（出所）各種資料より大和総研作成（イラストはソコスト https://soco-st.com/）

たとえば ChatGPT では、料金プランごとに日常的な会話や文章作成に使う機能と、高度な調査、長文処理、社内データとの連携、ソフトウェア開発支援などで利用できる機能に差が設けられている [1]。無料・低価格プランでも最低限の AI 機能を試すことができる一方、利用量の拡大や長時間の処理、詳細な調査などに加え、法人で利用する場合に必要となるセキュリティ確保やアクセス管理、サポート体制といった運用面の要件を満たすためには、上位プランの選択が実質的に不可欠となる場合が多い。

[1] ChatGPT ウェブサイト「料金」（2026 年 5 月 28 日最終閲覧）

また Claude でも、利用形態に応じて課金体系を分ける動きがみられる。具体的には、人間が画面上で対話しながら利用する通常の使い方とは別に、業務ツールへの組み込みや自動処理など、AI エージェント [2]のようにユーザーが直接操作せずに自動で処理を進める利用については、2026 年 6 月 15 日以降、既存のサブスクリプション（以下、サブスク）の利用枠とは別の専用月次クレジットから消費される仕組みが導入される [3]。これは、従来のサブスクの範囲内で多くの自動化処理を動かしていた利用者にとっては、これらの処理が別枠のクレジット消費に移行するため、実質的な負担増になり得る。

加えて、同社は 2026 年 6 月 9 日、サイバーセキュリティ分野で突出した能力を示す最上位モデル「Claude Mythos」と共通の基盤技術を用いつつ、同分野の機能に安全制限を加えた一般公開モデル「Claude Fable 5」の提供を開始した。この Fable 5 についても、6 月 22 日までの期間限定で、月額課金プランの契約者向けに追加料金なしで利用できるものの、同月 23 日以降は従量課金でのみ利用可能となる予定であった [4]。ただし、同社は 6 月 12 日、米国政府の命令を受けて「Claude Mythos 5」および「Claude Fable 5」の全顧客への提供を停止したと発表しており、従量課金への移行時期は不透明となっている [5]。

さらに、プログラミング支援 AI である米 GitHub 社の GitHub Copilot では、2026 年 6 月 1 日から従来の Premium Requests（回数ベースの利用枠）[6]を GitHub AI Credits（使用量に応じて消費される方式）へ置き換え、入力・出力・キャッシュ済みトークンとモデル別単価に基づいて利用量を算定する仕組みへ移行した [7]。日常的なコード補完は引き続きライセンス料に含まれ、追加課金なしで利用可能な一方、チャット、CLI、クラウドエージェント、サードパーティ製コーディングエージェント [8]などはクレジット消費の対象となる。報道では、この新たな仕組みへの移行後、一部のソフトウェア開発者などから、従来よりも速いペースで月間利用枠を消費しているとの不満や戸惑いの声が上がっている [9]。

この流れは、「AI を少し使う人」と「AI に仕事を任せる人」を同じ料金体系で扱い続けることが難しくなっていることを示している。

---

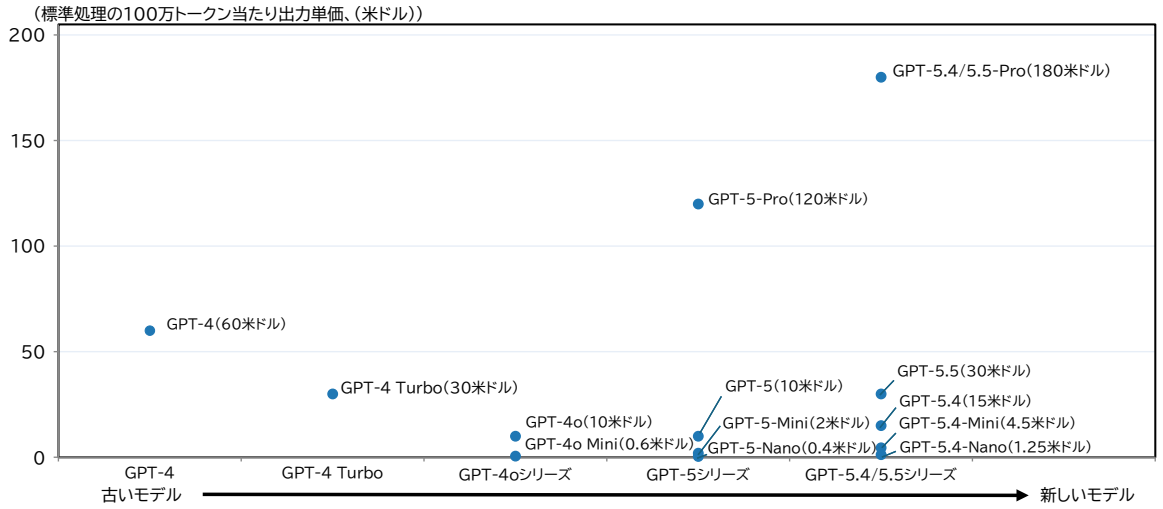
[2] AI エージェントとは、目的や目標の達成に向けて、必要な情報の収集や、手順の検討、実行方法までを、人が都度指示しなくても、AI が自律的に考えて実行する AI を指す。
[3] Claude Support “Use the Claude Agent SDK with your Claude plan”, last updated May 18, 2026.
[4] Anthropic “Claude Fable 5 and Claude Mythos 5”, June 9, 2026
[5] Anthropic “Statement on the US government directive to suspend access to Fable 5 and Mythos 5”, June 12, 2026
[6] Premium Requests とは、月額でリクエスト枠を購入し、高度な AI モデルや高度な AI 機能の利用回数に応じて、その枠を消費していく仕組み。
[7] GitHub Company news “GitHub Copilot is moving to usage-based billing: Starting June 1, your Copilot usage will consume GitHub AI Credits.”, April 27, 2026.
[8] CLI（コマンドラインインターフェース）とは、画面上にコマンド（命令文）を入力して操作する、開発者向けのインターフェースを指す。また、クラウドエージェントとは、クラウド上で AI が自動的にコード生成や修正などを行う機能、サードパーティ製コーディングエージェントとは、外部企業が提供する同様の開発支援 AI を指す。
[9] Business Insider “GitHub Copilot users get a rude awakening as new AI pricing goes into effect”, June 3, 2026

## AI モデルそのものの料金体系も分化

以上のようなサービスプランの階層化に加え、AI モデルそのものの料金体系も、用途ごとに必要となる計算資源や処理負荷の大きさに応じて分化し始めている。具体的には、軽いチャット、短い文章作成、翻訳、要約といった処理負荷が比較的小さい用途は、低単価モデルがカバーし、高度な調査、長文処理、ソフトウェア開発支援、社内データ連携といった処理負荷が大きい用途は、高性能な上位モデルが担うという棲み分けが進んでいる。

実際、米 OpenAI 社の API[10]料金を見ると、トークン当たりの単価は従来と比べて低下しているが、モデル間の価格差は大きく、低価格モデルでは 100 万トークン当たり数ドル程度にとどまる一方、高性能モデルでは 100 ドルを超えるものもある[11]。すなわち、簡易な応答で対応可能な用途には安価なモデルを使いやすくし、高度な推論や長文処理、高い信頼性が求められる用途には高価格なモデルを提供するという二極化が進んでいる（**図表 2**）。

**図表 2　主要 GPT 系モデルの API 単価の推移**




**（注）標準処理の 100 万トークン当たり出力単価。モデル性能や用途が異なるため単純な価格比較は適切ではないが、各モデルの出力単価がどの水準に分布しているかを概観することを目的としている。**
**（出所）米 OpenAI 社より大和総研作成**

こうした価格差は、AI 市場における競争と差別化の構造を反映している。汎用的なチャットや定型的な推論処理は、モデルの効率化、各社の価格競争、オープンウェイトモデル（モデルの内部情報が公開され、自社環境でも利用可能なモデル）の普及により、コモディティ化が進みやすく、モデル間の能力差が結果に現れにくい領域である。これに対し、エージェント的な業務自動化、社内システムとの統合、出力の信頼性、セキュリティ、監査対応といった領域では、高度な処理が求められる中でモデルの性能差が顕在化しやすく、提供側が差別化しやすい。また、利用企業にとっては、出力の質が業務成果やリスクに直結する領域であるため、モデル選択の重要性が高く、相応の料金設定が受け入れられやすい。

---

[10] API とはアプリケーション・プログラミング・インターフェース（Application Programming Interface）のことで、異なるソフトウェアや Web サービスなどをつなぐためのインターフェースを指す。
[11] OpenAI Developers “Pricing” （2026 年 5 月 28 日最終閲覧）

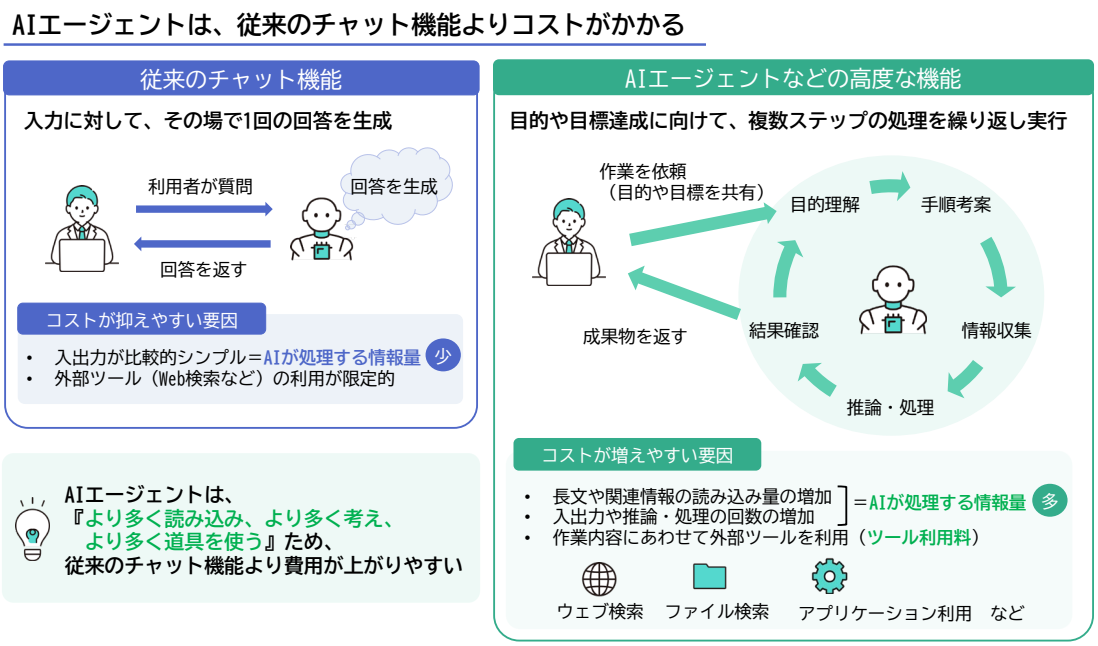
したがって、足元の AI 利用料の変化は、「AI 全体の利用料金が一様に引き上げられる」という単純な値上げではなく、「軽い用途は低価格に抑え、重い用途や高付加価値な用途には高い料金を設定する」という料金体系の二極化として捉えることが適切だろう。今後は、月額課金を基本としながらも、利用量、モデル性能、処理内容、外部ツール連携の有無に応じて料金が変わる、クラウドサービスに近い設計が広がっていくと考えられる。

## 2. AI エージェント化がコストを押し上げる

### エージェント化による AI コストの変化

前章で見た料金体系の二極化は、AI の利用が高度化・多様化する中で、提供側のコスト構造が変わりつつあることを反映している（**図表 3**）。

**図表 3　チャット機能と AI エージェントのコスト構造比較（イメージ図）**




（出所）各種資料より大和総研作成（イラストはソコスト https://soco-st.com/）

AI エージェント化でまず増えるのはトークン量である。ここで、生成 AI の料金でよく用いられる「トークン」とは、AI が読む・書く情報量の単位を指す。

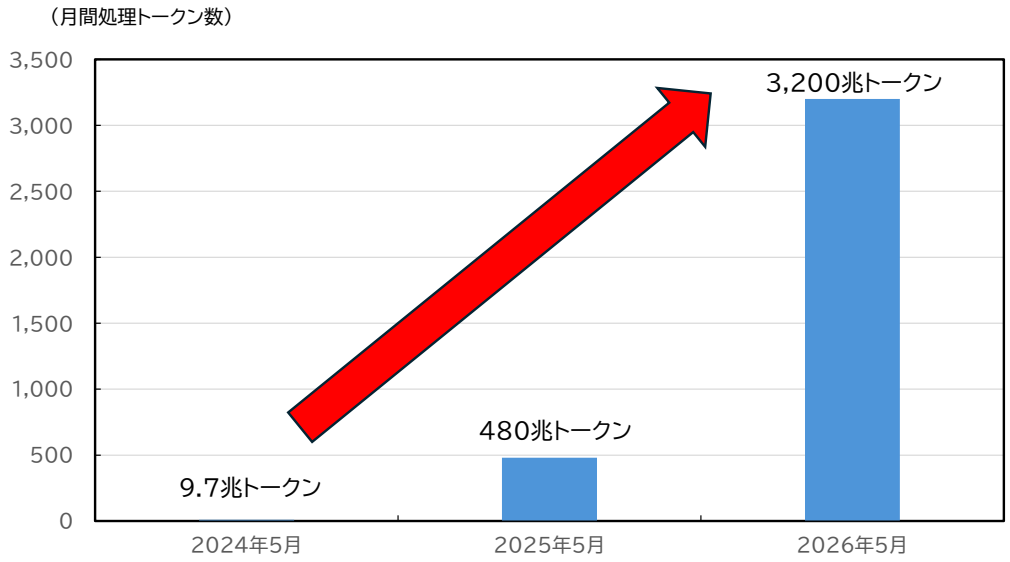
従来のチャットボット形式の利用では、利用者が質問し、AI が答えるという流れが中心だった。これに対して AI エージェントでは、AI が目的を理解し、手順を考え、必要な情報を探し、途中結果を確認し、うまくいかなければやり直す。利用者から見える成果物は一つでも、その内部では複数回にわたる入力や出力、推論といった処理が繰り返し発生しやすい。

また、コストを押し上げる要因は、トークン量の増加だけではない。AI が外部検索、ファイル検索、プログラムコード実行、社内システム連携などの「道具」を使う場合、道具そのものの利用料が発生するほか、検索結果やファイル内容を AI が読み込むための追加トークンも発生

する。米 OpenAI 社の料金表でも、ウェブ検索、ファイル検索、コンテナ利用[12]などは、通常のモデル利用料とは別の料金項目として示されている[13]。

AI が処理するトークン量の拡大は、米 Google 社の公表値からも見て取れる。同社は 2026 年 5 月に開催された開発者会議で、自社 AI 製品が処理するトークン数が月間 3,200 兆トークンに達し、前年の 480 兆トークンから約 7 倍になったと説明した[14]。これはあくまで同社のみの数値であり、生成 AI 市場全体を示すものではないが、AI 利用が急速に拡大していることを示す一例である（**図表 4**）。

**図表 4　AI 利用の高度化に伴う月間処理トークン数の増加**




**（注）米 Google 社の AI 製品に関する公表値であり、生成 AI 市場全体の統計ではない。**
**（出所）米 Google 社より大和総研作成**

もっとも、同等の性能水準でみれば、トークン単価は技術の進展により低下傾向にある。しかし、このトークン単価の低下が、そのまま AI 利用料金の低下を意味するとは限らない。調査会社 Gartner は、2030 年までに 1 兆パラメータ級 LLM の推論コストが 2025 年比で 90%超低下すると予測する一方、エージェント型モデルでは 1 タスク当たりのトークン消費が増えるため、全体の推論コストは上昇し得ると指摘している[15]。つまり、トークンは安くなっても、処理回数や読み書き量、外部ツール利用が増えれば、トークン消費量の増加や、ツール利用に伴うコストの増加が、単価の低下分を上回り、結果として総コストが増加する可能性がある。加えて、AI サービス提供企業がコストの低下を料金にすべて転嫁するとは限らないため、「トークン単価は下がるが請求額は膨らむ」という構図が生じる公算が大きい。

[12] コンテナとは、AI がプログラムコード実行やデータ処理を行うための隔離された計算実行環境を指す。
[13] OpenAI Developers “Pricing” （2026 年 5 月 28 日最終閲覧）
[14] Google “I/O 2026: Welcome to the agentic Gemini era”, May 19, 2026.
[15] Gartner “Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025”, March 25, 2026

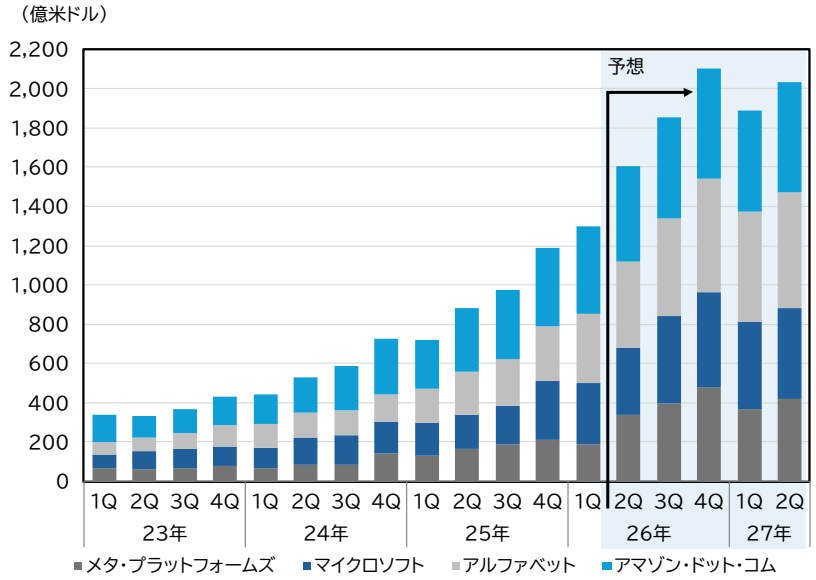
## AI インフラ投資の拡大と価格への波及

こうした AI が処理するトークン量の拡大などを背景に、AI 関連企業各社はデータセンターの建設を相次いで発表するなど、AI インフラへの大規模な投資を続けている。四半期ベースで見ても設備投資額は右肩上がりとなっており、2026 年後半もこの傾向は続くとみられる（**図表 5**）。一方で、収益化までの時間差や資金負担を懸念する声もあり、市場では過剰投資ではないかという議論も徐々に高まっている。

こうした議論が生まれる背景には、AI インフラが長期的な需要拡大を見据えた先行投資になりやすい構造がある。データセンターには着工から稼働までのリードタイムがあり、立地・電力・水資源の確保といった物理的制約も多い。そのため、需要の立ち上がりを待つのではなく、あらかじめ基盤を整える投資が先行しやすい。実際、米 Anthropic 社のダリオ・アモデイ CEO が、需要の急増に対して計算資源の確保が追いついていないと述べるなど [16]、計算資源の制約は既に表面化している。

したがって、今回の AI 利用料の引き上げの背景には、AI インフラへの巨額投資を回収しようとする、AI モデルの開発企業の意図があると考えられる。もっとも、投資回収が必ずしも AI 利用料の一律の値上げとして現れるわけではない。高度なモデルや機能はトークン消費が大きい一方で、他社との差別化を図りやすく、付加価値の高い領域として収益源にもなりやすい。そのため、各社はこうした領域への課金を重点的に強化する一方、基本的な機能については価格を据え置くなど、メリハリのある価格戦略を通じて投資回収を図っていると考えられる。

**図表 5　AI 関連企業各社の設備投資額の推移**




（注）2026 年 2Q 以降は Bloomberg の予測値
（出所）Bloomberg より大和総研作成

[16] CNBC “Anthropic CEO says 80-fold growth in first quarter explains ‘difficulties with compute’”, published May 6, 2026, last updated May 7, 2026.

## 3. 企業はなぜ高い AI 利用料を受け入れるのか

一方、AI の能力が「対話」から「作業の遂行」へと拡大したことで、ユーザー側の企業も AI に支払ってよいと考える金額が高まっているとみられる。

これまでの AI は、質問への回答や文章の要約、翻訳などを担う「賢い対話相手」としての役割が中心であった。こうした用途では、AI の比較対象は検索エンジンやチャットツールであり、企業が支払う金額もそれらに近い水準に収まりやすかった。つまり、AI がいかに賢くとも、業務を自律的に遂行できなければ、労働力としての経済的価値は限られていた。

しかし近年、AI は単に賢いだけでなく、長文処理、高度な調査、コーディングエージェントを用いたソフトウェア開発支援といった作業もこなせるようになり、一部の定型的で検証しやすい業務では、人間より速く処理できる場面が出てきた。この変化により、AI の比較対象は「ツール」から「人間の労働者」へと移行し、企業にとっての AI の経済的価値は大きく高まった。その結果、企業は AI に対してより高い料金を支払うことを合理的に判断しやすくなっている。

米国の AI 評価研究機関である METR が公表している Time Horizon を見ると、この変化を定量的に捉えることができる[17]。Time Horizon とは、AI が一定の成功確率でこなせるタスクの難度を、人間の専門家の作業時間に換算して測る指標である。たとえばある AI モデルの 50%の Time Horizon が 1 時間だった場合、人間の専門家が 1 時間程度かかる作業[18]を AI が 50%の確率で完了できることを意味する。

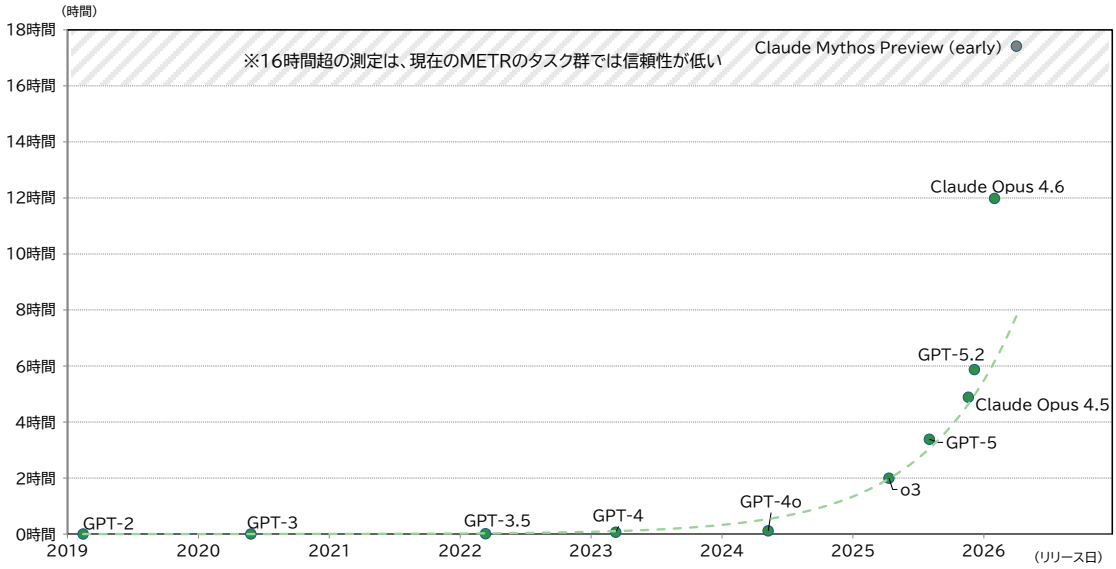
近年の AI モデルではこの Time Horizon が指数関数的に高まっている（**図表 6**）。2019 年発表の GPT-2 では約 3 秒、2020 年の GPT-3 でも約 9 秒にとどまっていたが、2023 年 3 月の GPT-4 で約 4 分に伸び、2025 年 8 月の GPT-5 では 3 時間 23 分へと跳ね上がった。さらに 2026 年 2 月発表の Claude Opus 4.6 は 11 時間 59 分、そしてサイバーセキュリティ分野における能力の高さで最近注目されている Claude Mythos Preview は最低でも 16 時間に達したという。なお、米 METR は 16 時間超の測定について現在のタスク群では信頼性が低いと注記しており、AI の能力向上がベンチマーク自体の測定限界に迫りつつあることを示唆している。

Time Horizon は正解や成功が客観的に判定可能なタスクを中心とした指標であり、あらゆる業務にそのまま当てはめられるわけではない。とはいえ、AI が担える作業時間が長くなるほど、料金の比較対象はチャットツールではなく人間の作業時間に近づく。企業における高度なモデルやエージェント機能の利用拡大は、AI モデルの利用コストの増加につながり得る一方、企業はその費用に加え、確認、権限管理、監査、教育などを含めた総運用コストと業務効率化の効果を比較する必要がある。

---

[17] METR “Task-Completion Time Horizons of Frontier AI Models”, last updated May 8, 2026.
[18] METR のウェブサイトには、人間の専門家が 1 時間強かかるタスクとして、Python の小さなライブラリのバグを修正することが挙げられている。

**図表 6　AI が 50%の成功確率で完了できるタスクの難度（人間の専門家の所要時間換算）**




（注 1）数値は、METR-Horizon-v1.1 における「50%成功時のタスク長」の中心推定値。これは、AI が実際にその時間だけ作業することを意味するものではなく、人間専門家が同じタスクを完了するのに要する時間で測ったタスク難度について、AI の成功確率が 50%となる水準を示す。

（注 2）米 METR は、Claude Mythos Preview の早期測定を 2026 年 5 月 8 日に追加した一方、16 時間超の推定値は現在のタスク群では信頼性が低いと注記している。また、Claude Opus 4.7、Grok 4.3、GPT-5.5 など一部の新モデルについては、2026 年 5 月末時点で同指標を未公表としている。

（出所）米 METR より大和総研作成

## 4. 日本企業への示唆

日本企業は、生成 AI を「便利で安価なチャットツール」として扱う段階にとどまらず、「業務プロセスを支える外部計算資源」として認識したうえで、管理する段階へと重点を移していく必要がある。クラウドサービスと同じように、利用が拡大するほど、データの蓄積や他システムとの連携、運用の最適化が進むことで利便性が高まる一方、利用量が把握できなければ費用対効果も見極めにくい。

まず重要なのは、AI 利用量の可視化である。どの部門が、どの業務で、どの程度 AI を使っているのかを把握する必要がある。そのうえで、用途に応じて AI モデルを使い分けることが重要となる。たとえば、軽い要約や定型文作成には低価格モデルを使い、複雑な判断や重要顧客対応には高性能モデルを適用するといった運用が考えられる。

この点で、米 GitHub 社の GitHub Copilot の AI Credits 方式は、企業向け管理の方向性を示している。利用枠を組織単位で一元管理し、部門・コストセンター・ユーザー単位で予算や上限を設定する仕組みである。上限を超えた場合に追加課金を認めるか、その時点で利用を止めるかを管理者が決めることができる。日本企業においても、全社共通の利用枠と、重要業務・試験導入・個人利用といった用途ごとに上限を設定するなど、利用の粒度に応じた管理設計を行うことが、費用の急増を防ぐうえで重要になる。

もう一つの論点は、外部 AI サービスと、自社で管理・運用する AI 基盤の使い分けである。外部 AI サービスは、最新モデルをすぐに使える利点がある。もっとも、足元では高性能な AI モデルの多くは海外企業によって提供されているため、これらのサービスの利用は海外の基盤に依存する構造を伴いやすい。その結果、利用量が増えれば、従量課金による費用の増加に加え、為替変動の影響を受けやすくなるほか、提供国の政策・制度・規制によって利用条件や提供形態が左右される点も課題となる。そこで、企業によっては、米 Meta 社の Llama、中国 Alibaba 社の Qwen、仏 Mistral AI 社の Mistral、米 Google 社の Gemma など、モデルの中身が公開され、自社のサーバー環境でも動かすことができる「オープンウェイトモデル」を併用する選択肢が出てくる。これらは外部のクラウドサービスを介さずに自社環境で運用できるため、コストやデータ管理の面で柔軟性を確保しやすい。

ただし、自社運用にはコストがかかる。外部 AI サービスでは提供側が担っているインフラ運用（サーバー、GPU などの整備・運用）やセキュリティ管理、モデルの更新・保守といった機能に加え、それらを支える人材の確保も自社で対応する必要がある。したがって、すべてを外部サービスに委ねるのでも、すべてを自社運用するのでもなく、用途ごとに最適な組み合わせを選ぶことが現実的である。

こうした企業の選択を支える政策としては、海外の AI サービスの利用を抑えることではなく、国内企業が AI を業務に組み込み、業務プロセスの設計や社内データとの連携・運用・評価といった付加価値を国内に蓄積できる形にすることが重要になる。海外の AI モデルを使う場合でも、国内での実装力を高めれば、AI 利用の便益を取り込みつつ、コストを抑制する余地が拡大するだろう。

生成 AI の普及局面は、「無料で試す」段階から、「業務価値に応じて支払い、管理する」段階へ移りつつある。こうした環境下では、AI の利用量や用途を可視化したうえで、用途に応じたモデルの使い分けやコスト管理を行い、費用対効果を踏まえて活用していく体制の構築が重要となる。料金上昇を過度に恐れるよりも、どの業務に使えば効果が出るのかを見極め、費用と便益を管理しながら活用することが求められる。

以上