ChatGPT

GPT-5.4のベンチマーク・使用評価・反応に関する公開情報調査（2026年3月7日まで）

エグゼクティブサマリー

本調査の結論は、「GPT-5.4は“プロ向け知識業務＋エージェント実行”に最適化された統合フロンティアモデルとして、OpenAIが公式に複数の“実務・エージェント系”評価で改善を示した一方、日本語の標準ベンチ（JGLUE/JMMLU等）での**公式スコアは未公開（情報不足）**であり、導入判断は“自社タスクでの再現性テスト”を前提に行うべき」です。[1]

OpenAI公式の主要主張は、(a) 知識業務ベンチGDPvalで「専門職と同等以上の比較が83.0%」(b) OSWorld-Verifiedで「75.0%」(c) BrowseCompで「82.7%（Proは89.3%）」(d) Toolathlonで「54.6%」(e) 事実性について「個々の主張の誤り確率がGPT-5.2比で相対33%低下、回答全体の誤り含有も相対18%低下」(f) 最大約105万トークン級のコンテキスト（API）――などです。[2]

一方で弱点・注意点は、(a) GPT-5.4 Proは高コスト（APIで入力$30/1M tokens、出力$180/1M tokens）かつ遅い（リクエストにより「数分」かかり得る）(b) 長コンテキストは“できる”が“常に高精度”ではなく、OpenAI自身が長文用の内部評価（Graphwalks/MRCR等）を提示しており、1M帯で精度が落ちる指標もある (c) 安全性面では、GPT-5.4 ThinkingがPreparedness Framework上で生物・化学、およびサイバーで「High capability」として扱われ、相応のセーフガードが前提 (d) SNSやコミュニティでは、価格・速度・ベンチの意味（汚染/条件差）への懐疑が目立つ、という点です。[3]

導入を検討する組織向けには、①“高付加価値な長手順・文書/表計算・UI操作”に限定して投入し、②モデルをスナップショットで固定して回帰テスト、③コンピュータ操作は確認ポリシー（ヒューマン・イン・ザ・ループ）必須、④日本語品質は自社データ＋日本語ベンチ（Nejumi/llm-jp-eval系）で実測、⑤安全性はSystem Cardの想定リスク（jailbreak/プロンプトインジェクション等）を踏まえてガードレールと監査ログを整備――が実務的です。[4]

調査範囲と方法

本レポートは、ユーザー指定の優先順位（公式→一次資料→日本語情報）に沿い、主に以下を収集・突合しました：OpenAI公式リリース（GPT-5.4、関連するGPT-5.3 Instant/GPT-5.3-Codex、SWE-bench評価方針等）、OpenAI開発者向け技術文書（モデル仕様・価格・機能制約・プロンプトガイド）、OpenAI Deployment Safety HubのSystem Card（GPT-5.4 Thinking）です。[5]

学術一次資料としては、ユーザーが例示した一般ベンチ（MMLU/HELM/GLUE相当）と、日本語ベンチ（JGLUE等）および、GPT-5.4公式が採用した“エージェント/マルチモーダル系”評価（OSWorld、WebArena、Mind2Web/Online-Mind2Web、MMMU、OmniDocBench、SWE-Bench Pro、Terminal-Bench 2.0）について、原論文・公式サイトを参照しました。[6]

独立ベンチマーク・導入事例・反応については、主要テックメディア（例：The Verge[7]、TechRadar[8]、Gizmodo Japan[9]、PC Watch[10]、gihyo.jp[11]）と、SNS/コミュニティ（X[12]、Reddit[13]、Hacker News[14]）のスレッドを“代表例”として採録しました。いずれも、主張の性質（公式/一次/二次/体験談）を明示し、推測は避けています。[15]

公式情報とモデルの位置づけ

OpenAI [16] の公式発表では、GPT-5.4はChatGPT（GPT-5.4 Thinkingとして）、API、Codexで提供され、「推論・コーディング・エージェント的ワークフローの進歩を単一モデルに統合」し、GPT-5.3-Codex相当のコーディング性能を取り込みつつ、ツール・ソフトウェア環境・スプレッドシート/プレゼン/ドキュメント等のプロ業務を強化した“プロ向け統合モデル”として位置づけられています。 [17]

ChatGPT側の「Thinking」体験としては、複雑な問いに対し“最初にプラン（作業方針）を提示し、応答の途中で方向修正できる”旨が説明されています。これは、長手順の成果物作成（資料、分析、調査）で“往復ターン”を減らすことを狙った設計と読めます（公式主張）。 [18]

API/仕様面では、モデルカード（開発者ドキュメント）上、GPT-5.4はテキスト＋画像入力、テキスト出力で、Reasoningの強さは最上位、速度は“Medium”、価格は入力$2.50/1M tokens・出力$15/1M tokens、コンテキストは約1,050,000、最大出力は128,000 tokenとされています。 [19]

上位版のGPT-5.4 Proは「より計算を使って考える」ことで精度を上げる設計で、Responses APIのみで提供、Reasoning effortはmedium/high/xhigh、価格は入力$30/1M・出力$180/1M tokensと明示されています。難問だと「数分かかる」場合があるためbackground modeの利用が推奨されています（公式）。 [20]

ベンチマーク結果一覧

OpenAI公式が公表した主要結果（スコア・比較対象・条件）

“重要度の高い比較（実務導入の判断材料になりやすいもの）”を中心に、OpenAI公式ページに掲示された数値を表に整理します。評価は研究環境で行われ、原則reasoning effort=xhigh（例外あり）と明記されています。 [21]

ベンチマーク名	GPT-5.4スコア	比較対象モデル（例）	テスト条件・注記（公開範囲）	出典
GDPval（wins or ties）	83.0%	GPT-5.2：70.9%、GPT-5.3-Codex：70.9%	44職種の知識業務成果物。GPT-5.4はreasoning effort=xhigh、GPT-5.2はheavy。	[22]
投資銀行モデリング（Internal）	87.3%	GPT-5.2：68.4%（他：GPT-5.3-Codex 79.3%など）	内部評価（詳細データ/再現手順は未公開）。	[22]
OfficeQA	68.1%	GPT-5.2：63.1%（GPT-5.3-Codex 65.1%）	社内/業務系評価（詳細タスク公開は限定的）。	[22]
SWE-Bench Pro（Public）	57.7%	GPT-5.3-Codex：56.8%、GPT-5.2：55.6%	実務寄りソフトウェア工学ベンチ（Proの論文/サイトは別途）。	[23]
Terminal-Bench 2.0	75.1%	GPT-5.3-Codex：77.3%、GPT-5.2：62.2%	端末タスク（89タスク）を測るベンチ。	[24]

ベンチマーク名	GPT-5.4スコア	比較対象モデル（例）	テスト条件・注記（公開範囲）	出典
OSWorld-Verified	75.0%	GPT-5.2：47.3%、人間：72.4%	スクショ＋マウス/キーボードでデスクトップタスク。GPT-5.3-Codexは新APIパラメータで74.0%と注記。	25
WebArena-Verified	67.3%	GPT-5.2：65.4%	DOM＋スクショの両方で操作。Verified版は評価の堅牢化を狙う系統。	26
Online-Mind2Web	92.8%	ChatGPT Atlas（Agent Mode）：70.9%（英語版）／84.0%（日本語版）	スクショ観測のみでブラウザ操作。英語/日本語ページで比較対象値に差分があるため注意。	27
MMMU-Pro（no tools）	81.2%	GPT-5.2：79.5%	マルチモーダル理解・推論。MMMU/Proは大学・専門知識と画像理解を含む。	28
OmniDocBench（normalized edit distance）	0.109（低いほど良い）	GPT-5.2：0.140	文書解析（PDF等）で、GPT-5.4はreasoning effort=noneで実行と明記。	29
BrowseComp	82.7%	GPT-5.2：65.8%（Proは77.9%）	ChatGPT search tool使用。ブロックリストで汚染対策。測定日差（インターネット状態/検索系更新）に注意。	30
Toolathlon	54.6%	GPT-5.2：45.7%/46.3%（表で差分）	“実ツール/APIを使う多段タスク”で、メール読取→添付処理→採点→スプレッドシート記録等の例が説明される。	31
ARC-AGI-2（Verified）	73.3%	GPT-5.2：52.9%	抽象推論系（Verified）。Proは83.3%と記載。	32
Humanity’s Last Exam（no tools）	39.8%	GPT-5.2：34.5%	学術系難問。tools有りは52.1%（Proは58.7%）。	33
Graphwalks BFS 256K–1M	21.4%	（比較値なし）	1M級長文でのグラフ探索系。0–128Kでは93.0%（GPT-5.2は94.0%）。	33

表から読み取れる傾向（推測ではなく、数値が示す範囲の要約）として、GPT-5.4は「知識業務（GDPval）」「コンピュータ操作（OSWorld/WebArena/Online-Mind2Web）」「ツール活用（BrowseComp/Toolathlon）」でGPT-5.2からの改善幅が大きく、コーディング（SWE-Bench Pro）では改善はあるが差分が小さい――という“強化領域の偏り”が見えます。 34

主要な汎用ベンチ（MMLU/HELM/GLUE相当）と日本語ベンチの公開状況

ユーザーが指定した“MMLU/HELM/GLUE相当”は、学術界で広く用いられる一般能力・透明性の主要ベンチです（例：MMLUは57科目の多分野知識テスト、HELMは多軸評価の枠組み、GLUEはNLU複数タスクの統合ベンチ）。 35

ただし、OpenAIの「GPT-5.4公式リリース本文」には、少なくとも“MMLU”や“HELM”という項目は含まれていません（ページ内検索結果としてヒットなし）。したがって、**GPT-5.4のMMLU/HELMスコアを公式に一覧表へ記載するための公開情報は不足**しています。 36

日本語領域については、JGLUEが日本語の一般言語理解を測るベンチとして構築されていること、また“日本語LLMを評価する枠組み”としてW&BのNejumi Leaderboard 4がJMMLUやMMLU_Prox_JA、HLE-JA等の日本語評価を採用していることは一次資料で確認できます。 37
一方で、**GPT-5.4の日本語ベンチ（JGLUE/JMMLU等）の公式スコアはOpenAIのGPT-5.4評価一覧に掲示が見当たらず**（少なくとも同ページにJMMLU等の記載はなく）、本レポートでは「未公開/情報不足」として扱います。 38

使用評価の利点と欠点（機能別）

本節は、公式文書（リリース、Prompt guidance、モデルカード、System Card）と、二次資料・利用者反応を分離しながら整理します。

生成品質（文章・成果物の完成度）：OpenAIは、GPT-5.4を“プロの成果物”（スプレッドシート、プレゼン、ドキュメント）でより一貫して洗練された出力を出すと位置づけ、GDPvalや内部の投資銀行モデリング系評価、プレゼン評価（人手比較）を根拠として挙げています。 39
一方でコミュニティ側では、（特にPro系の）“考えすぎ/長い/遅い”への不満が出やすい構造で、公式にも難問では「数分」かかり得る旨が明記されています。 40

事実性（ハルシネーション/誤り）：公式には、ユーザーが事実誤りを指摘した匿名化プロンプト集合で、GPT-5.2比で「個々の主張が相対33%誤りにくい」「回答全体が相対18%誤りを含みにくい」と主張しています（ただし評価集合はOpenAI内部）。 27
またPrompt guidanceでは、研究・引用の дисцип律（ソース収集と引用規則、完了条件の明示）を強く推奨しており、“放っておけば常に正しい”というより“契約（output contract）で堅牢化する”設計思想が見えます。 41

安全性・有害出力：GPT-5.4 ThinkingのSystem Cardでは、メンタルヘルス、エモーショナル・リライアンス、自傷などを含む“動的（多ターン）評価”を導入し、not_unsafe（規約違反しない応答率）で評価する説明があります。 42
同文書は、脱獄（jailbreak）評価をより難しい多ターン手法へ更新し、GPT-5.1→5.2で大きく改善、5.2→5.4で追加改善と述べています。 43
ただしPreparedness Frameworkでは、生物・化学、およびサイバーで“High capability”として扱い、相応のセーフガードを適用する前提が明記されています。これは「能力が上がるほど、誤用・悪用リスクの管理が重要になる」ことを示す公式シグナルです。 44

速度・レイテンシ：OpenAIは、GPT-5.4を“トークン効率が高い推論モデル”として、GPT-5.2より少ないトークンで問題解決しやすい（＝速度にも寄与）と説明しています。 45
モデルカード上も、GPT-5.4はSpeed=Mediumで、Proは“Slowest”で、遅延が構造的トレードオフであることが明示されています。 46
さらにChatGPT/Codex文脈では、/fastモード等で処理速度向上を謳っています（ただしこれは“能力そのまま速度のみ改善”という運用機能の話で、純モデル比較とは別軸です）。 47

コスト：API価格は、GPT-5.4が入力$2.50/1M・出力$15/1M、GPT-5.4 Proが入力$30/1M・出力$180/1Mと明記され、長コンテキスト（>272K入力）には追加倍率がかかる条件もあります。 48
SNS上では価格上昇を問題視する反応が典型で、特に“Proが高い/遅い”系の体験談が散見されます（ただし個別投稿は再現性が保証されない点に注意）。 49

マルチモーダル性能（視覚・文書・UI）：GPT-5.4は“汎用モデルとして初めてネイティブのコンピュータ操作能力”を備えるとされ、スクリーンショットに応じたクリック/キー操作やPlaywright等での自動化コード生成を強調しています。[27]
ベンチ面でも、OSWorld-Verified、WebArena-Verified、Online-Mind2Web、MMMU-Pro、OmniDocBenchでの数値が公表されています。[50]
加えて、画像入力の“original detail”レベル導入（最大1,000万画素または6K相当）など、UI/文書で効く高解像度処理の方向性が示されています。[47]

日本語対応・多言語性能：OpenAIは日本語ページでGPT-5.4を公式告知し、日本語での機能説明とベンチ結果（OSWorld等）を掲載しています。[47]
しかし、JGLUE/JMMLU等の日本語中心ベンチのGPT-5.4スコアを公式に示す資料は本調査範囲では確認できず、Nejumi/llm-jp-evalのような枠組み自体は公開されているものの、GPT-5.4の“日本語ベンチの点数”を意思決定にそのまま使える形では不足しています。[51]

カスタマイズ性・API/ツール連携：開発者向けには、reasoning.effortで計算量を可変にでき、ツール使用（web search、file search、computer use、MCP等）を含むエージェント構築を前提にしたガイドが整備されています。[52]
コンピュータ操作では、開発者メッセージで挙動を調整でき、リスク許容度に合わせた確認ポリシーも設定できると説明されています。これは“実行系エージェント”を業務導入する際のガバナンス要素です。[27]
一方、GPT-5.4 Proは“Structured outputsは未対応”、Fine-tuningやDistillationも未対応と明記されており、カスタマイズの主戦場は（Proの場合）プロンプト設計＋ツール設計＋ガードレールに寄ります。[53]

推論時の制約（長文・長手順・評価の揺らぎ）：OpenAIは長コンテキスト評価（Graphwalks/MRCR v2等）を提示し、0–128Kでは高いが、256K–1M帯で落ちる指標（例：Graphwalks BFS 256K–1Mが21.4%）も示しています。従って“1Mコンテキスト＝常に強い”とは読み替えないのが安全です。[33]
またBrowseCompでは、ブロックリスト更新や測定日差により「モデルだけでなく検索システムとインターネット状態もスコアに影響する」と明示され、時点差・条件差の解釈が必要です。[54]

ユーザー・研究者・企業の評価と反応の傾向分析

企業・導入側の肯定的評価として、公式ページには、Mercor[55]のCEO Brendan Foody[56]が「APEX-Agentsでトップ」「スライド/財務モデル/法務分析など長期成果物に強い」「速く低コスト」と述べる引用が掲載されています。[18]
またHarvey[57]のApplied Research責任者 Niko Grupen[58]のコメントとして、“文書量の多い法務作業で新しい基準”“BigLaw Benchで91%”などが掲示されています（ただしこのベンチは同社評価で詳細手順は本文では限定的）。[59]

技術系メディアの論調は、概ね「エージェント化（PC操作）へ踏み込んだ」「長コンテキスト」「仕事向け統合」「誤り低下」を主要ポイントとして報じています。[60]

一方、否定的・懸念の代表例は次の三系統に収れんします。第一に、価格と速度（特にPro）で、コミュニティでは“高い/遅い/計算し過ぎ”の体験談が出やすい点です（ただし投稿の一般化は不可）。[61]
第二に、ベンチマークへの懐疑で、SWE-Bench Proの伸びが小さいことを根拠に“最適化された見せ方では”と疑うコメント、あるいはSWEの絶対値がまだ低いという指摘（例：日本語X投稿で「57%程度、まだ低い」）が見られます。[62]
第三に、安全性・ポリシー設計を巡る議論で、System Cardの扱い（例：“emotional reliance”を自傷等と並べる分類）に反応するスレッドが立つなど、強いモデルほどアラインメントや運用基準が注目される傾向があります。[63]

総じて、肯定的反応は「実行（computer use）」「調査/探索（BrowseComp）」「成果物生成（GDPval系）」「統合（推論×コーディング×エージェント）」に集中し、否定的反応は「コスト」「レイテンシ」「ベンチの解釈」「安全性/透明性」に集中する——という構図です。 [64]

総合評価と導入判断の基準

総合評価

公開情報ベースでの総合評価は、「GPT-5.4は“実務エージェント（ツール・ブラウザ・UI操作）”を強く意識した評価セットで改善が示され、少なくともOpenAIが重視する“仕事の完遂”指標で優位性を主張できる状態にある」一方、「高コスト構成（Pro）や長コンテキスト帯の性能劣化、そして日本語標準ベンチの不足により、“自社条件での受入試験なしに全面置換”は危険」です。 [65]

組織向け判断基準

判断基準は、(a) **成果物の価値がトークンコストを上回るか**（例：法務レビュー、定例レポート、資本市場分析、RPA代替）、(b) **“多手順”が本当に必要か**（単発QAなら安価モデルで十分）、(c) **UI操作を許容できるガバナンスがあるか**（確認ポリシー、操作ログ、権限分離）、(d) **日本語品質を自社データで検証したか**（JGLUE相当や社内FAQでの誤答率）——で整理すると実務的です。 [66]

リスク緩和策

コンピュータ操作は、OpenAIが“確認ポリシー”でリスク許容度に応じた挙動を設定できると述べているため、**不可逆操作（送信/購入/削除/本番反映）は常に人間確認**、可逆操作のみ自動、のように段階化するのが基本です。 [67]
安全性面は、System Cardがjailbreakや多ターン安全評価を重視しているため、導入側でも“脱獄・プロンプトインジェクション・ツール悪用”を含む赤チームテストを実施し、最悪ケースの挙動（拒否・中断・エスカレーション）を仕様化するのが合理的です。 [68]

テスト手順の提案（短期PoC→段階導入）

最小限の現実的手順は、(1) **スナップショット固定**（例：gpt-5.4-pro-2026-03-05等）で再現性確保、(2) **社内タスクを“成功/失敗”で採点できる形に変換**（可能なら自動評価＋人手監査の二層）、(3) **品質×コスト×レイテンシ×安全性**を同時に測定、(4) **モデル切替は“一度に一変更”**（Prompt guidanceの推奨）で回帰検知、(5) **日本語はNejumi/llm-jp-eval系のタスク分類（NLI/QA/翻訳/真実性/安全性/コード等）を参考に社内版ベンチを作る**——です。 [69]

参考URL（主要一次資料・公式）

```
OpenAI公式リリース（英）: https://openai.com/index/introducing-gpt-5-4/
OpenAI公式リリース（日）: https://openai.com/ja-JP/index/introducing-gpt-5-4/
GPT-5.4 Thinking System Card（Deployment Safety Hub）: https://deploymentsafety.openai.com/
gpt-5-4-thinking
OpenAI API docs（gpt-5.4）: https://developers.openai.com/api/docs/models/gpt-5.4
OpenAI API docs（gpt-5.4 pro）: https://developers.openai.com/api/docs/models/gpt-5.4-pro
Prompt guidance for GPT-5.4: https://developers.openai.com/api/docs/guides/prompt-guidance/

主要ベンチ一次資料（例）
```

```
MMLU論文: https://arxiv.org/abs/2009.03300
HELM論文: https://arxiv.org/abs/2211.09110
GLUE論文: https://aclanthology.org/W18-5446/
JGLUE論文: https://aclanthology.org/2022.lrec-1.317/
OSWorld論文: https://arxiv.org/abs/2404.07972
WebArena Verified（OpenReview）: https://openreview.net/forum?id=CSIo4D7xBG
SWE-Bench Pro論文: https://arxiv.org/abs/2509.16941
Terminal-Bench 2.0論文: https://arxiv.org/abs/2601.11868
```

---

1 2 4 5 7 9 10 12 14 16 17 18 21 22 23 24 25 26 27 28 29 30 31 32 33 34 36 38 39 45 50 54 56 57 59 64 65 66 67 Introducing GPT-5.4 | OpenAI
https://openai.com/index/introducing-gpt-5-4/

3 11 20 40 53 61 69 GPT-5.4 pro Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-5.4-pro

6 35 Measuring Massive Multitask Language Understanding
https://arxiv.org/abs/2009.03300?utm_source=chatgpt.com

8 63 GPT-5.4'S SYSTEM CARD: OpenAI put "emotional reliance ...
https://www.reddit.com/r/OpenAI/comments/1rlu1ul/gpt54s_system_card_openai_put_emotional_reliance/?utm_source=chatgpt.com

13 42 43 44 55 68 GPT-5.4 Thinking System Card - OpenAI Deployment Safety Hub
https://deploymentsafety.openai.com/gpt-5-4-thinking

15 58 60 OpenAI's new GPT-5.4 model is a big step toward autonomous agents
https://www.theverge.com/ai-artificial-intelligence/889926/openai-gpt-5-4-model-release-ai-agents?utm_source=chatgpt.com

19 46 52 GPT-5.4 Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-5.4?utm_source=chatgpt.com

37 JGLUE: Japanese General Language Understanding ...
https://aclanthology.org/2022.lrec-1.317/?utm_source=chatgpt.com

41 Prompt guidance for GPT-5.4 | OpenAI API
https://developers.openai.com/api/docs/guides/prompt-guidance/

47 GPT-5.4 が登場 | OpenAI
https://openai.com/ja-JP/index/introducing-gpt-5-4/

48 Introducing GPT-5.4
https://openai.com/index/introducing-gpt-5-4/?utm_source=chatgpt.com

49 GPT-5.4 Pricing MORE EXPENSIVE THAN GPT-5.2
https://x.com/scaling01/status/2029619520860565648?utm_source=chatgpt.com

51 GitHub - wandb/llm-leaderboard: Project of llm evaluation to Japanese tasks · GitHub
https://github.com/wandb/llm-leaderboard

62 GPT-5.4 Benchmarks : r/OpenAI
https://www.reddit.com/r/OpenAI/comments/1rlp1m0/gpt54_benchmarks/?utm_source=chatgpt.com