GPT-5.4 ベンチマーク・評価・反応 総合分析レポート

エグゼクティブサマリー

2026 年 3 月 4 日に OpenAI がリリースした GPT-5.4 は、推論・コーディング・エージェント型ワークフローを統合した最新フロンティアモデルである。GDPval で 83.0%（人間の専門職と同等以上）、OSWorld-Verified で 75.0%（人間のベースライン 72.4%を超過）という注目すべきスコアを記録し、ネイティブ Computer Use 機能、最大 1M トークンのコンテキストウィンドウ、Tool Search 機能による 47%のトークン削減など、実務向けの機能強化が施されている。一方で、API 価格の上昇、フロントエンドデザイン能力の不足、272K トークン超過時の料金倍増といった課題も指摘されている。本レポートでは、ベンチマーク結果、良い点・悪い点、および業界・ユーザーからの評価・反応を総合的に分析する。[1][2][3][4]

主要ベンチマーク結果

GPT-5.4 は、前世代の GPT-5.2 および GPT-5.3-Codex と比較して、多くのベンチマークで顕著な改善を示している。以下に主要ベンチマークの比較を示す。

実務能力・知識ベンチマーク

ベンチマーク	GPT-5.4	GPT-5.4 Pro	GPT-5.2	GPT-5.3-Codex
GDPval（人間専門職との比較 勝率/引分）	83.0%	—	70.9%	70.9%
投資銀行モデリング（内部）	87.3%	—	68.4%	79.3%
OfficeQA	68.1%	—	63.1%	65.1%
Harvey BigLaw Bench（法務）	91%	—	—	—

GDPval は米 GDP 上位 9 業界にまたがる 44 職種の実務タスクを評価するベンチマークであり、GPT-5.4 は 83.0%の比較で人間の業界実務者に匹敵または上回った。ペンシルベニア大

学ウォートン・スクールの Ethan Mollick 准教授は、GDPval を「経済の観点から見ておそらく最も適切な AI 能力の指標」と評している。GPT-5.1 の 38%、GPT-5.2 の 70.9%から一貫して上昇しており、進歩の速度は驚異的である。[2][5][6]

推論・科学ベンチマーク

ベンチマーク	GPT-5.4	GPT-5.4 Pro	GPT-5.2
GPQA Diamond（PhD 科学）	92.8%	94.4%	92.4%
HLE（ツールなし）	39.8%	42.7%	34.5%
HLE（ツールあり）	52.1%	58.7%	45.5%
FrontierMath Tier 1-3	47.6%	50.0%	40.7%
ARC-AGI-1（Verified）	93.7%	94.5%	86.2%
ARC-AGI-2	73.3%	—	52.9%

ARC-AGI-2 では 73.3%を達成し、GPT-5.2 の 52.9%から約 20 ポイント向上した。ただし、人間は 100%を達成するベンチマークであり、依然として「真の流動的知性」との間にはギャップがある。[7][8]

コーディング・エンジニアリングベンチマーク

ベンチマーク	GPT-5.4	GPT-5.3-Codex	GPT-5.2
SWE-Bench Pro	57.7%	56.8%	55.6%
Terminal-Bench 2.0	75.1%	77.3%	62.2%
BrowseComp	82.7%	—	65.6%
Toolathlon	68.9%	—	—

SWE-Bench Pro ではコーディング性能が 57.7%と微増にとどまっているが、Terminal-Bench 2.0（75.1%）や BrowseComp（82.7%）では大幅な改善が確認される。ただし、Claude Opus 4.6（SWE-bench 80.8%）や Gemini 3.1 Pro（SWE-bench 80.6%）と比較すると、SWE-bench スコアでは GPT-5.4 が大きく後れを取っている点には注意が必要である。[9][1][7]

Computer Use（PC 操作能力）

ベンチマーク	GPT-5.4	GPT-5.2	人間ベースライン
OSWorld-Verified	75.0%	47.3%	72.4%
WebArena-Verified	67.3%	—	—
Online-Mind2Web	92.8%	—	—

OSWorld-Verified で 75.0%を達成し、人間のベースライン（72.4%）を初めて超えた汎用モデルとなった。GPT-5.2 の 47.3%から約 28 ポイントの飛躍であり、世代間で質的な変化が起きている。[10][3][11][12]

GPT-5.4 の良い点（Strengths）

人間超えの Computer Use 能力

GPT-5.4 は OpenAI 初の汎用ネイティブ Computer Use モデルであり、Playwright コードによるコンピュータ操作と、スクリーンショットを見ながらのマウス/キーボード操作の両方に対応する。OSWorld-Verified で人間を上回る 75.0%を記録したことは、エージェント型 AI の実用化における大きなマイルストーンである。[13][4][14]

実務能力の大幅向上

GDPval で 83%のスコアは、44 職種のプロフェッショナルと同等以上の実務パフォーマンスを示す。投資銀行モデリングでは 87.3%（GPT-5.2 の 68.4%から 19 ポイント向上）、法務の BigLaw Bench では 91%を達成しており、知識労働分野での適用可能性が飛躍的に広がった。[5][2]

事実精度（ファクチュアリティ）の改善

GPT-5.2 と比較して、個別の主張レベルで偽りの情報が 33%減少し、エラーを含むレスポンス全体では 18%減少した。GPT シリーズ全体としてのハルシネーション低減の流れ（GPT-4→GPT-5 で 65〜84%改善）の延長線上にあり、最も事実に忠実な OpenAI モデルと位置づけられている。[6][15][16][2]

Tool Search 機能によるトークン効率化

MCP Atlas の 250 タスク評価（36 の MCP サーバー使用）で、精度を維持しながら総トークン使用量を 47%削減できることが確認された。大規模エージェントシステムにとって、コストとレイテンシの直接的な低減に繋がる重要な機能である。[2][10][13]

1M トークンのコンテキストウィンドウ

Codex では実験的に 100 万トークンに対応し、GPT シリーズ最大のコンテキストウィンドウとなった。大規模なコードベースや長大なドキュメントの一括処理が可能になった。
[17][13]

高い自律性（自走力）と処理速度

開発者からの実体験レポートでは、「指示を出すと最後まで完遂しようとする姿勢が強い」と評価されている。Codex の/fast モードにより、token velocity が最大 1.5 倍に向上し、長時間のコーディング作業でも快適な動作が報告されている。[18][1][13]

対話品質の向上

GPT-5.2 と比較して「数字数字している」「頭が固い」感じが減り、Claude のような自然な対話感に近づいたとされる。抽象的な議論への対応力や説明力が向上し、批判的フィードバックを含む納得感のある回答が可能になっている。[13]

GPT-5.4 の悪い点（Weaknesses）

フロントエンドデザイン能力の不足

GPT-5.4 の最も一貫して指摘される弱点は、UI デザインのセンスである。Claude Opus 4.6 や Gemini 3.1 Pro と比較して「かなり劣る」と評価されている。プログラムのロジック実装は優秀だが、SVG イラスト作成、レイアウト調整、スライド作成ではデザインが崩れることがあるとの報告が複数ある。ランディングページの生成テストでは、Claude 4 の方がデザイン的に洗練されたアウトプットを生成した。[19][20][21][22][13]

API 価格の上昇

モデル	入力（/1M トークン）	キャッシュ入力	出力（/1M トークン）
GPT-5.2	$1.75	$0.175	$14.00
GPT-5.4（<272K）	$2.50	$0.25	$15.00
GPT-5.4（>272K）	$5.00	$0.50	$22.50
GPT-5.4 Pro（<272K）	$30.00	—	$180.00

入力トークン単価は GPT-5.2 の$1.75 から$2.50 へ 43%上昇した。272K トークンを超えると入力コストが 2 倍の$5.00/M になる「ロングコンテキストサーチャージ」が設けられており、大規模コードベースや法務文書の処理コストが予想以上に高額になる可能性がある。OpenAI はトークン効率の改善により総コストは抑えられると主張しているが、実際のワークロードでの検証が必要である。[23][24][25]

SWE-Bench での競合劣位

SWE-Bench Pro/Verified において、GPT-5.4 のスコア（57.7%）は Claude Opus 4.6（80.8%）や Gemini 3.1 Pro（80.6%）に大きく及ばない。コーディング品質を重視する開

発者にとって、この差は無視できない。Terminal-Bench 2.0 では GPT-5.4 が優位（75.1% vs Claude 65.4%）であり、ターミナル操作中心のワークフローでは利点がある。[26][9]

自律性が高すぎることの問題

自走力の高さは利点である一方、仕様が曖昧な場合に意図しない方向に実装が進んでしまう問題がある。「こうあるべきだ」という判断で独自に作り込んでしまい、UI が散らかったり、1 時間かけた実装がごちゃごちゃした見た目になることがある。[21][13]

機能面の制約

GPT-5.4 にはいくつかの機能的制約がある：

- 音声・動画入力に未対応（モデル単体）[27]
- GPT-5.4 Pro は Structured Outputs や Code Interpreter を使えない[27]
- ヘルスケア関連のベンチマークでは GPT-5.2 より若干悪化しているとの報告がある[17]
- 現実世界の文脈を見落とすケースがある（例：旅行計画で季節的な混雑を考慮しない）[22]

レスポンスの長さ

GPT-5.4 は GPT-5.2 と比較して平均レスポンスが 24%長くなっているとの指摘があり、出力トークンの増加がコスト増に繋がる可能性がある。[17]

競合モデルとの比較

2026 年 3 月時点での 3 大フロンティアモデルの位置づけは以下の通りである。

評価軸	GPT-5.4	Claude Opus 4.6	Gemini 3.1 Pro
SWE-Bench Verified	57.7%（Pro）	80.8%[9]	80.6%[9]
GPQA Diamond	92.8%	—	94.3%[9]
ARC-AGI-2	73.3%	—	77.1%[9]
Computer Use (OSWorld)	**75.0%**[4]	—	—
コンテキスト窓	1M（Codex 実験的）	1M（β）	1M
入力価格(/1M)	$2.50	$3.00[28]	$2.00-$4.00[28]

日本語品質	◎[29]	◎[29]	△（直訳調）[29]
UI デザイン	△[22]	◎[20]	○
エージェント能力	◎ (Codex)	◎ (Claude Code)[29]	○

GPT-5.4 は Computer Use 能力と GDPval（実務能力）で圧倒的な優位性を持つが、コーディング品質では Claude、推論の GPQA Diamond・ARC-AGI-2 では Gemini 3.1 Pro に後れを取る。価格対性能比では Gemini 3.1 Pro が「price-performance king」と評されている。[9]

ユーザー・業界からの評価と反応

日本国内の反応

日本の SNS では「キタァァァ！！！ GPT-5.4」と歓声が上がり、「GPT-5.4 が明確に Claude 潰す方向性の性能向上を観測できる」など、期待感の高い投稿が多く見られた。技術メディアの ZDNet Japan は、83%の実務能力について「専門職で生涯にわたって磨いてきたスキルによって家族を養っている私たち人間は、当惑しながら不安な気持ちで深呼吸をし、最善を祈るしかない」と、AI の進歩に対する複雑な心境を表現している。[30][5]

GIZMODO Japan のレビューでは、Codex CLI での使用感について「精度もスピードも上がってる感じで、すごい快適」「体感、1.5 倍くらいのスピードで生成してる」と好意的な評価が報告されている。[31]

日本人開発者による詳細レビューでは、「仕事をちゃんと終わらせる力が強化された」という公式の説明を「体験からも実感できる進化」と評し、特に自走力の高さと処理速度の向上を高く評価している。一方で、UI デザインの弱点や自律性の高さに起因する問題点も率直に指摘されている。[13]

海外の反応

Reddit r/codex では「GPT 5.4 Thread - Let's compare first impressions」に 128 票・109 コメントが集まり、コーディングの実務利用における活発な議論が展開されている。Tom's Guide は「OpenAI just made every other AI model look slow」と題して、OSWorld スコアの革新性を強調した。[32][18]

一方、GPT-5 シリーズ全般に対する批判的なユーザー意見も根強い。Reddit 上では「creativity seems to have vanished（創造性が消えた）」「feels sterile, simplistic, and lacks substance（無菌的で単純、実質がない）」との声があり、「品質低下はスケーラビリティのためのトレードオフか技術的限界か」という根本的な疑問を呈するスレッドも存在する。[33][34]

専門家・レビューサイトの評価

Is It Good AI は GPT-5.4 を 75.1/100（7.5/10）と採点し、「企業やエージェント型ワークロードにとって明白な新デフォルト」と結論づけている。Awesome Agents のレビューは、OSWorld スコアについて「汎用モデルが標準化されたデスクトップナビゲーションベンチマークで人間を上回ったのを見たのは初めて」と高く評価している。[12][2]

ある詳細なレビューでは「ほぼすべての測定可能な指標において、これまでに作られた中で最高の AI モデル」と評されつつも、「ベンチマーク最適化されている感じ」というユーザーの声も紹介されている。[19][17]

System Card における自己評価

OpenAI の Deployment Safety Hub に公開された GPT-5.4 Thinking System Card では、Chain of Thought（CoT）の制御可能性が低いことが認められている。また、「情報が不足している可能性がある場合のコンテキスト探索が弱い」点が主な弱点として挙げられている。[35]

総合評価

GPT-5.4 は、「仕事を完遂する AI」としてのコンセプトを具現化した意欲的なモデルである。Computer Use 能力の人間超え、GDPval における専門職水準の実務能力、Tool Search によるエージェント効率化は、企業の DX 推進や AI エージェント開発にとって大きな意義を持つ。

しかし、SWE-Bench における競合（Claude・Gemini）との差、フロントエンドデザインの弱さ、価格上昇は無視できない課題であり、「万能モデル」とは言い難い。2026 年 3 月時点の最適戦略は、タスク特性に応じた複数モデルの使い分け——コーディング品質重視なら Claude、実務・エージェント能力重視なら GPT-5.4、コストパフォーマンス重視なら Gemini——という組み合わせ型アプローチと言える。[29][9]

GPT-5.2 のリリースからわずか 3 ヶ月という異例の速さでの投入は、OpenAI がフロンティアモデル競争で巻き返しを図る強い意志を示している。今後のユーザー実環境での検証とフィードバックの蓄積が、このモデルの真の評価を決定づけることになる。

References

1. OpenAI launches GPT-5.4 Thinking and Pro combining coding ... - On the coding front, GPT-5.4 scores 57.7 percent on SWE-Bench Pro, just slightly above GPT-5.3 Codex...
2. GPT-5.4 Review (2026) - Is It Good AI - Honest GPT-5.4 review for 2026. Pricing ($2.5/$15 per 1M tokens), context window, strengths, weaknes...
3. GPT-5.4 Review: Is This OpenAI's Most Powerful Model Ever? - Is GPT-5.4 truly the best? Test 1M context, Thinking mode, agentic coding & native computer use. 75%...

4. Introducing GPT-5.4 - OpenAI - We've designed GPT-5.4 to be performant across a wide range of computer-use workloads. It is excelle...

5. OpenAI の新モデル「GPT-5.4」、8 割以上のタスクで人間の専門職に ... - 専門職の人間に対する勝率は 83% 今回、3 月初旬という GPT-5.2 から 3 カ月も経たない時期にリリースされた GPT-5.4 は、83%の確率で専門職の人間に匹敵する、あ ...

6. OpenAI's new GPT-5.4 clobbers humans on pro-level work in tests - OpenAI's new GPT-5.4 clobbers humans on pro-level work in tests - by 83%. GPT-5.4 is also more relia...

7. GPT-5.4 | OpenAI's Most Capable AI Model - Gemini 3 - GPT-5.4 benchmarks: 75.0% OSWorld (surpasses human 72.4%), 92.8% GPQA Diamond, 73.3% ARC-AGI-2, 83.0...

8. GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoning - An in-depth analysis of OpenAI's GPT-5.2 achieving a 54% score on the ARC-AGI-2 benchmark for abstra...

9. GPT-5.4 vs Claude Opus 4.6 vs Gemini 3.1 Pro - EvoLink.AI - A practical 2026 comparison of GPT-5.4, Claude Opus 4.6, and Gemini 3.1 Pro for cost, context, and c...

10. OpenAI's GPT-5.4 sets new records on professional benchmarks - GPT-5.4 topping the leaderboard means it is the best-performing model in a field where no model is y...

11. GPT-5.4: Computer Use, Tool Search, Benchmarks, Pricing - OpenAI releases GPT-5.4 with native computer use, 1M context, and tool search reducing tokens by 47%...

12. GPT-5.4 Review: The Computer-Use Frontier | Awesome Agents - GPT-5.4 brings native computer use, a 1M token context window, and serious coding muscle to OpenAI's...

13. 【徹底解説】GPT-5.4 が遂に公開！PC 操作や検索性能＆1M トークン ... - GPT-5.4 の性能を示す各種ベンチマークにおいて、GPT-5.2 や GPT-5.3-Codex と比較して著しい向上が見られます。特に注目すべきは実務能力を測る GDPval で、米 ...

14. GPT-5.4 Uses a Computer Better Than Most Humans : r/OpenAI - In this video, I break down what GPT-5.4 actually brings to the table: a 75% score on OSWorld — a de...

15. GPT-5 vs GPT-4: 80% Reduction in AI Hallucinations

16. Marked reduction in hallucination rates with GPT-5 - PMC - NIH

17. GPT-5.4 Review — Pricing, Ratings & Use Cases | LLMPick - OpenAI's most capable and efficient frontier model for professional work. Combines industry-leading ...

18. GPT-5.4 is here — and OpenAI just made every other AI model look ... - On an internal benchmark of spreadsheet modeling tasks designed for junior investment banking analys...

19. GPT-5.4 は本当に、本当に優秀だ - 5.4 Pro は極めて高価だが、従来解けなかった暗号パズルを数分で解くなど特定の難問に対して圧倒的な性能を示す。総じて、コーディングタスクにおいては「 ...

20. GPT-5 vs Claude 4：AI コーディング戦争の勝者は？価格 - note - エグゼクティブ・キーメッセージ 性能は一進一退: GPT-5 はベンチマークで Claude を上回る一方、実際のフロントエンド開発テストでは、UI デザインの美しさで Claude が勝る場面もあり、タスクによ...

21. GPT-5.4 ついに公開！ChatGPT の新機能や Google Gemini との違い - OpenAI の最新 AI モデル「GPT-5.4」の機能や使い方、Google の Gemini 3.1 Pro、Anthropic の Claude Opus 4.6 との違いをわかりやすく解説します。

22. 【速報】GPT-5.4 登場｜何が変わった？概要・料金からビジネス活用 ... - GPT-5.4 とは何かを、特徴・料金・使えるプラン・ベンチマーク・Claude/Gemini との違いまで整理。ChatGPT・API・Codex での使い方や、企業導入時の注意点を公式情報ベースで分かり ...

23. GPT-5.4 deep dive: pricing, context limits, and tool search explained - The per-token price is higher, but GPT-5.4 is meaningfully more token-efficient — in OpenAI's MCP At...

24. GPT-5.4 Pricing (2026): API Costs, Benchmarks & Worth the Upgrade? - Get the official GPT-5.4 pricing for 2026. Compare API rates, the 272K surcharge, and see why GPT-5....

25. GPT-5.4 Pricing, Benchmarks & API Costs (2026) | TokenCost - GPT-5.4 costs $2.50/1M input tokens and $15/1M output. Full breakdown of pricing tiers, benchmark re...

26. Claude Code を解約すべきか？6 つの視点からの実測分析 - Claude Code vs GPT-5.4 の核心的な結論は以下の通りです： コーディングベンチマークでは依然として Claude がリード: SWE-Bench で 80.8% vs 77.2% ...

27. GPT-5.4 とは何か - Zenn

28. GPT-5.4 API Pricing 2026: Latest Forecast, Scenarios & Cost ... - A practical GPT-5.4 pricing analysis with current OpenRouter listing data, scenario planning, and co...

29. 「ChatGPT vs Gemini vs Claude を 100 時間使って分かった本当の ... - 2026 年 3 月時点で、ChatGPT の最上位モデルは GPT-5.4 Thinking だ。推論・コーディング・エージェント型ワークフローを統合したフラッグシップモデルで、作業 ...

30. OpenAI が GPT-5.4 をリリース、ユーザーは「キタァァァ！」と歓喜 ... - さらに「GPT-5.4 が明確に Claude 潰す方向性の性能向上を観測できる」など、性能比較に熱く語る投稿も見られ、全体的に期待感が高まっている様子だ。 AI ...

31. OpenAI が新 AI モデル「GPT-5.4」をリリース、性能は“大幅に”向上 ... - Gemini の性能と主義を貫く Anthropic に押されていましたが...。2026 年 3 月 6 日、OpenAI が新しい AI モデル「GPT-5.4」を発表しました。GPT-5.4 は知的な実務作業（具体的には、...

32. GPT 5.4 Thread - Let's compare first impressions : r/codex - Reddit - 128 votes, 109 comments. I am pushing with Fast+XHIGH doing every day coding tasks. Now first time I...

33. Many people are saying GPT 5 is horrible. What pros and cons have you experienced? - Many people are saying GPT 5 is horrible. What pros and cons have you experienced?

34. GPT-5 Quality Decline: Scalability Trade-offs or Technical Limitations? - GPT-5 Quality Decline: Scalability Trade-offs or Technical Limitations?

35. GPT-5.4 Thinking System Card - Deployment Safety Hub - OpenAI - Its main weaknesses are poorer context-seeking when information may be missing. 3.7 Avoid Accidental...