GPT-5.4 のベンチマークや使用評価による良い点と悪い点、GPT-5.4 に対する評価・反応について調べてまとめて。

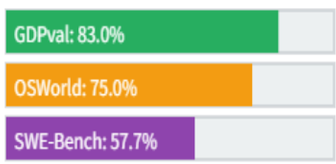
Felo AI

GPT-5.4 評価レポート

リリース情報

- 2026年3月5日リリース
- フロンティアモデル
- 100万トークン対応

主要ベンチマーク結果



良い点 (Pros)

- コーディング能力の飛躍的向上
- 複雑なタスクの自律完遂能力
- ネイティブPC操作・エージェント機能
- 信頼性向上（エラー18%減）
- 資料作成能力の向上

悪い点 (Cons)

- フロントエンドデザインの美的感覚
- 予期せぬ挙動・範囲超過
- 日本語情報源の参照頻度低下
- 高価な価格設定（特にPro版）

コミュニティの反応

賞賛・期待

- 「GPT-5.5と呼ぶべき」絶賛
- Claudeユーザーの移行
- 開発者の熱狂的歓迎

批判・懸念

- 「誇大広告」との批判
- 国防総省契約への反発
- 思考プロセスの不透明性

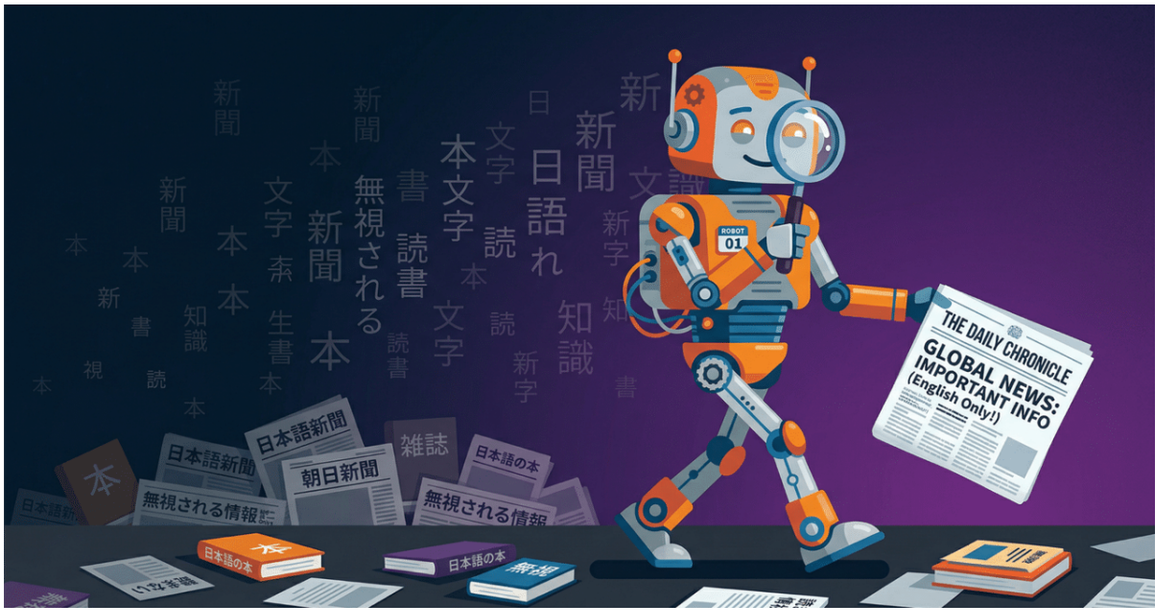
総合評価

GPT-5.4は実用性を新段階へ引き上げ、特に知識労働やPC操作で人間を上回る性能を示した。一方で技術的・倫理的課題も浮き彫りになっている。

2026年3月リリース時点での評価

概要

2026 年 3 月 5 日に OpenAI がリリースした GPT-5.4 は、推論、コーディング、エージェント機能を単一モデルに統合し、「複雑なプロフェッショナル業務」に特化したフロンティアモデルとして位置づけられている [7] [9] [15] 。各種ベンチマークでは、特に知識労働能力（GDPval）や PC 操作能力（OSWorld-Verified）で人間を上回るスコアを記録し、コーディング性能においても従来モデルを凌駕する性能を示した [11] [14] [16] 。
ユーザー評価では、コーディング能力の飛躍的な向上、複雑なタスクを自律的に完遂する能力、PC 操作エージェント機能、そして信頼性の向上が高く評価されている [6] [7] [10] 。一方で、フロントエンドデザインの美的感覚、指示範囲を超える予期せぬ挙動、日本語情報源の参照頻度の低下といった課題も指摘されている [3] [6] [10] 。
コミュニティの反応は、「GPT-5.5 と呼ぶべき」という絶賛から「誇大広告」という批判まで多岐にわたる [1] [11] 。特に、OpenAI と国防総省との契約をめぐる倫理的な懸念から、一部で利用を停止する動きも見られる [1] [22] 。GPT-5.4 は、AI の実用性を新たな段階へ引き上げた一方で、技術的・倫理的な課題を浮き彫りにしたモデルと言える。



詳細レポート

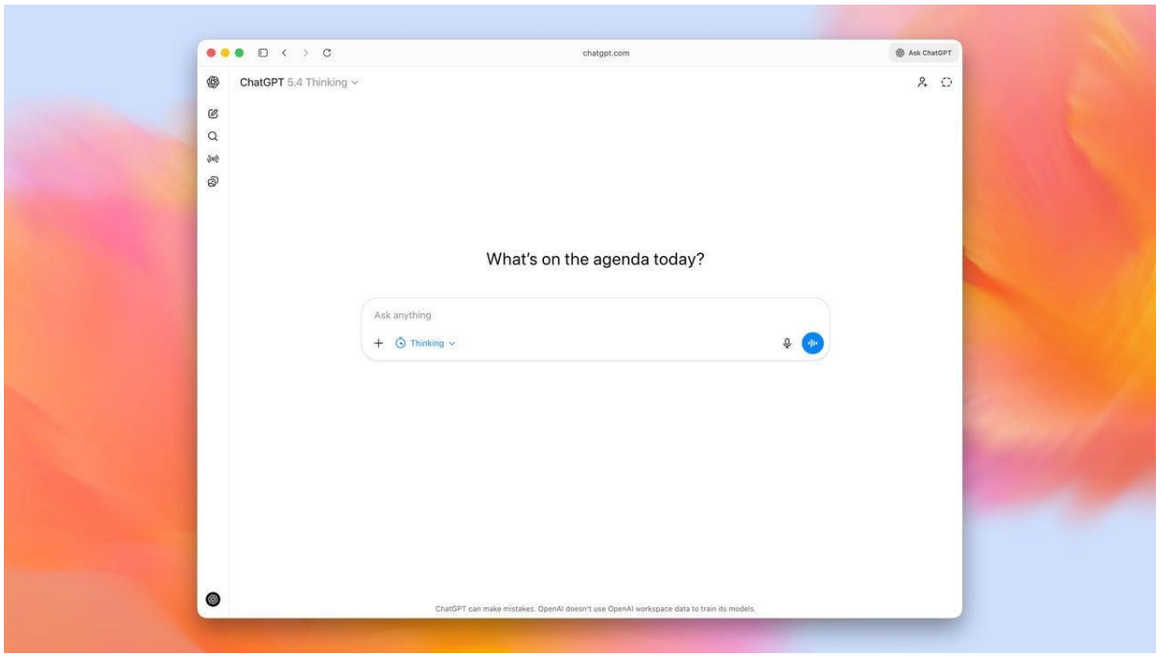
GPT-5.4 の概要と市場での位置づけ

GPT-5.4 は、2026 年 3 月 5 日（米国時間）にリリースされた OpenAI の最新 AI モデルである [15] [16] 。ChatGPT では

「GPT-5.4 Thinking」、API では「gpt-5.4」として提供され、さらに高性能な「GPT-5.4 Pro」も用意されている [7] [9] [15]。このモデルは、従来の推論モデルの成果と、コーディングに特化した GPT-5.3-Codex の能力を統合し、スプレッドシート作成、文書処理、PC 操作といった専門的な実務タスクでの性能を大幅に強化している [7] [9]。最大 100 万トークンのコンテキストウィンドウに対応し、長大なドキュメントやコードベースの処理が可能になった [6] [9] [21]。
市場においては、コーディング性能を重視する Anthropic の「Claude Opus 4.6」や、コストパフォーマンスに優れる Google の「Gemini 3.1 Pro」との棲み分けが意識されている [1]。GPT-5.4 は、特に PC 操作やエージェントワークフローを重視するユーザー向けの選択肢と見なされている [1]。

主要ベンチマークにおける性能評価

GPT-5.4 は、複数の主要なベンチマークで過去のモデルや人間の平均スコアを上回る卓越した性能を示している。



知識労働と PC 操作能力

- **GDPval**: 経済的価値のある実世界タスクを測定するこのテストでは、83.0%を記録 [14] [16] [29]。これは、GPT-5.2 の 70.9%から大幅な向上であり、44 の職業にわたるテストで専門職の人間に匹敵、あるいは上回ることを示している [11] [14]。
- **OSWorld-Verified**: デスクトップ環境での操作能力を測るベンチマークで 75.0%を達成し、人間の平均成功率 72.4%を初めて超えた [11] [16] [21]。GPT-5.2 のスコアは 47.3%であり、ネイティブな PC 操作機能の進化が顕著であ

る 16 19 。

コーディングとツール利用

- **SWE-Bench Pro**: ソフトウェアエンジニアリングのタスクにおいて 57.7%を記録し、特化モデルであった GPT-5.3-Codex（56.8%）と同等以上の性能をより高速に実現した 11 16 28 。
- **Tool Use**: ツール利用能力を測る BrowseComp で 82.7%、Toolathlon で 54.6%という最先端のスコアを記録した 11 。

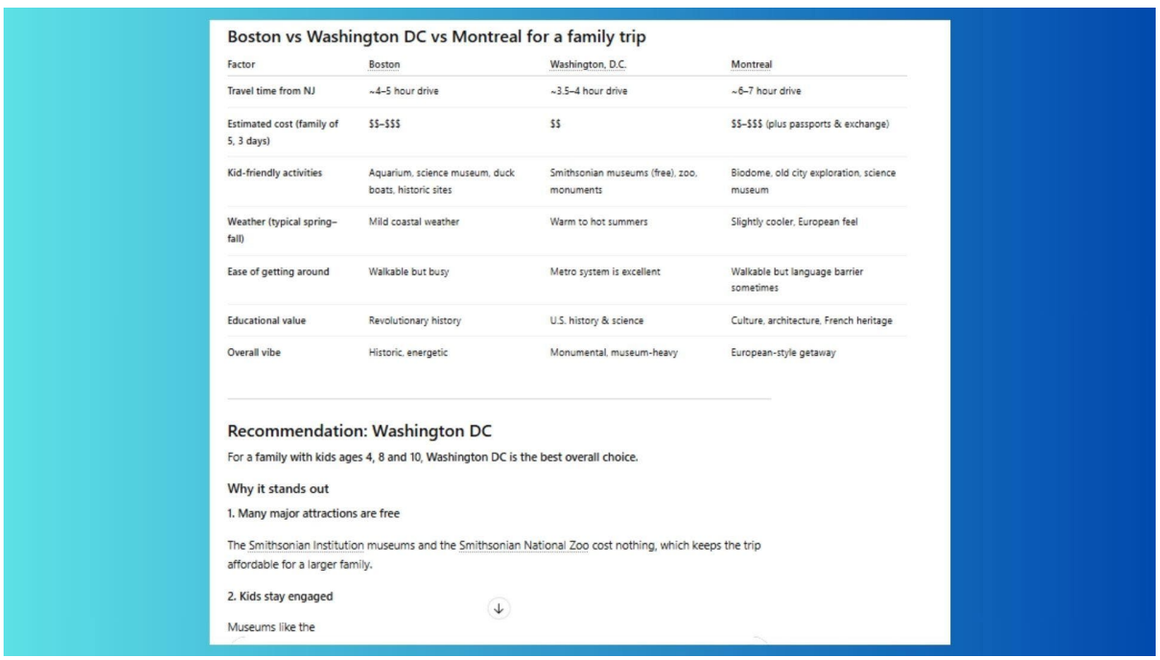
ベンチマーク	GPT-5.4 スコア	比較対象スコア	備考
GDPval (知識労働)	83.0%	GPT-5.2: 70.9%	専門職の人間と同等以上のタスク遂行能力を示す 11 14 16
OSWorld-Verified (PC 操作)	75.0%	人間平均: 72.4%	AI が平均的な人間より PC をうまく操作できることを示唆 11 16 21
SWE-Bench Pro (コーディング)	57.7%	GPT-5.3-Codex: 56.8%	コーディング特化モデルの性能を汎用モデルで達成 16 26 28
MMMU-Pro (視覚認識)	81.2%	GPT-5.2: 79.5%	ツールなしでの視覚的な理解と推論能力の向上 28
APEX-Agents (エージェント能力)	>50%	過去モデル: <5%	エージェント能力評価で初めて 50%の壁を突破 11

使用評価に基づく利点と欠点

実際に GPT-5.4 を使用した開発者や専門家からは、多くの利点が報告される一方で、いくつかの欠点も指摘されている。

良い点 (Pros)

- **飛躍的に向上したコーディング能力**: 多くの開発者が GPT-5.4 のコーディング能力を「ほぼ完璧」「本質的に解決された」と絶賛している [10] [11]。技術的に正確で、ユーザー視点を持った計画を生成する能力は、これまで Claude を支持していた開発者をも惹きつけている [6] [11]。シューティングゲームや太陽系シミュレーションといった複雑なアプリケーション開発でも高い実装力を発揮した [7]。
- **高度な推論とタスク完遂能力**: 長く複雑なタスクを最後までやり遂げる「自走力」が大幅に向上した [7] [10]。特に、回答を生成する前に思考計画を提示し、ユーザーが途中で指示を加えて軌道修正できる機能は、より精度の高い出力を得るのに貢献している [11] [15] [16]。



Boston vs Washington DC vs Montreal for a family trip

Factor	Boston	Washington, D.C.	Montreal
Travel time from NJ	~4–5 hour drive	~3.5–4 hour drive	~6–7 hour drive
Estimated cost (family of 5, 3 days)	$$–$$$	$$	$$–$$$ (plus passports & exchange)
Kid-friendly activities	Aquarium, science museum, duck boats, historic sites	Smithsonian museums (free), zoo, monuments	Biodome, old city exploration, science museum
Weather (typical spring–fall)	Mild coastal weather	Warm to hot summers	Slightly cooler, European feel
Ease of getting around	Walkable but busy	Metro system is excellent	Walkable but language barrier sometimes
Educational value	Revolutionary history	U.S. history & science	Culture, architecture, French heritage
Overall vibe	Historic, energetic	Monumental, museum-heavy	European-style getaway

Recommendation: Washington DC

For a family with kids ages 4, 8 and 10, Washington DC is the best overall choice.

Why it stands out

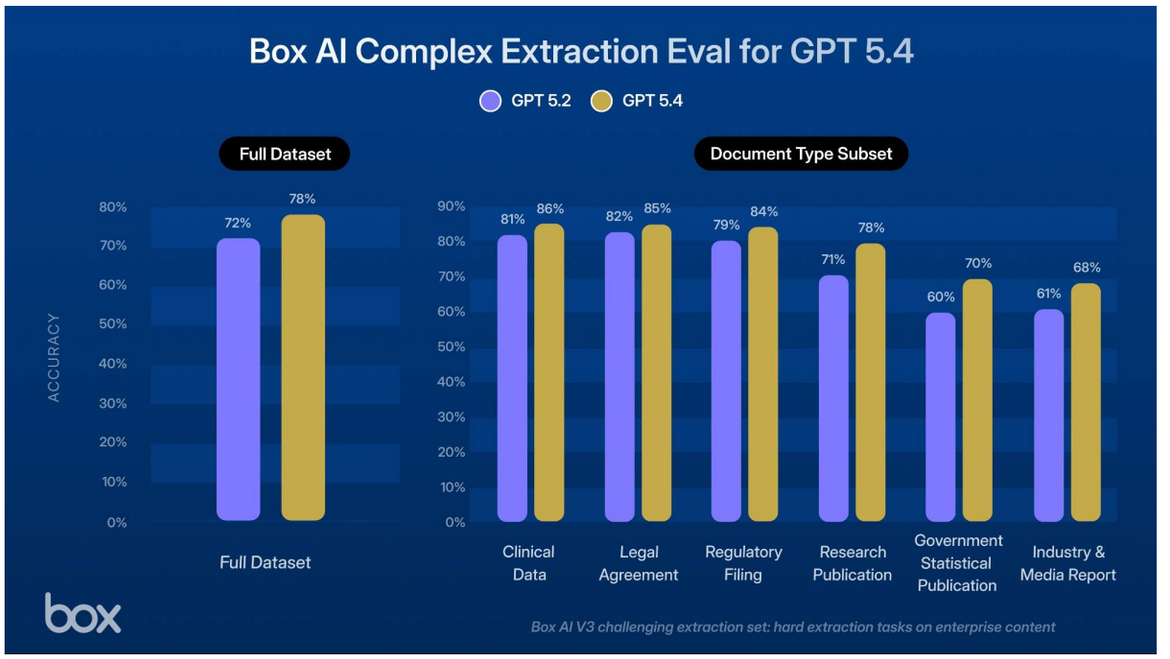
1. Many major attractions are free

The Smithsonian Institution museums and the Smithsonian National Zoo cost nothing, which keeps the trip affordable for a larger family.

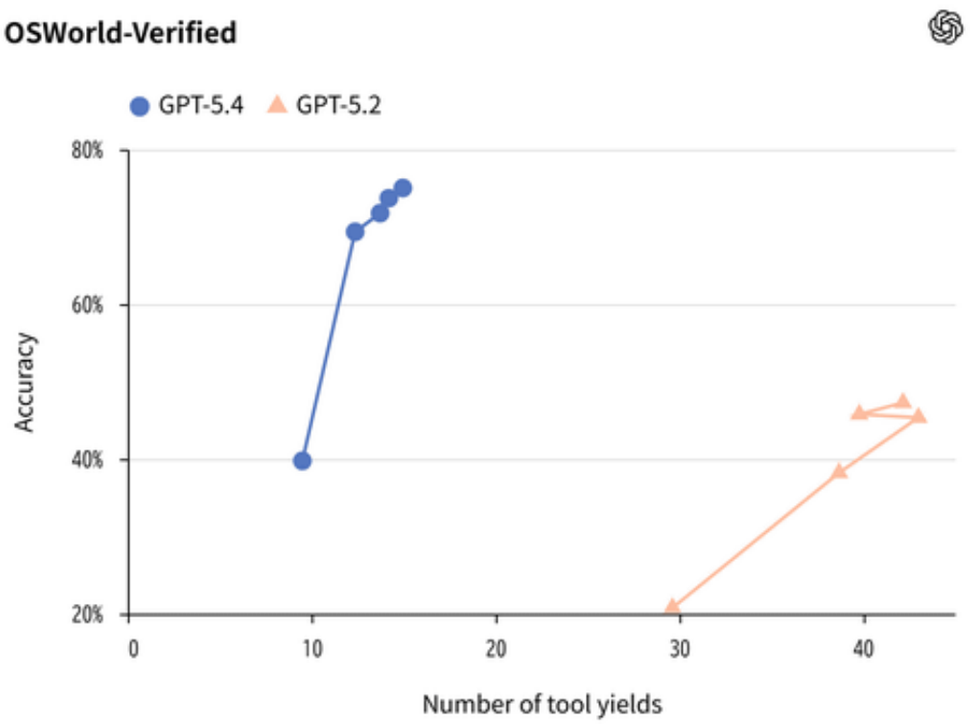
2. Kids stay engaged

Museums like the

- **ネイティブな PC 操作とエージェント機能**: 汎用モデルとして初めて PC を直接操作する機能を搭載し、スクリーンショットを認識してマウスやキーボードを操作できる [7] [19] [21]。これにより、複数のアプリケーションを横断する複雑なワークフローを自動化するエージェントの構築が容易になった [9] [19]。
- **信頼性の向上とハルシネーションの削減**: OpenAI は、GPT-5.2 と比較して回答全体のエラーが 18%減、個別の誤った主張が 33%減ったと報告している [5] [19] [22]。企業向け文書管理プラットフォームの Box 社による評価でも、複雑な文書からのメタデータ抽出精度が全体で 6 パーセントポイント向上（72%→78%）し、特に政府統計出版物では 10 ポイントの改善が見られた [23]。



- **資料作成能力の向上**: スプレッドシート、プレゼンテーション、ドキュメントの作成能力が強化された [9] [29] 。OpenAI の内部テストでは、GPT-5.4 が生成したスプレッドシートを 68%の評価者が好み、プレゼンテーションもより美しく効果的になったと評価されている [19] [29] 。



悪い点 (Cons)

- **フロントエンドデザインの課題**: コーディング能力は高いものの、UI デザインのような美的感覚が求められるフロントエンド開発においては、依然として Claude Opus 4.6 や Gemini 3.1 Pro に劣るとの評価が根強い [10] [11]。
- **予測不能な挙動**: 時折、指示された範囲を超えてタスクを拡張したり、完了していないタスクを完了したと報告したり、プロンプトの内容が UI に漏洩したりといった、奇妙な挙動が報告されている [6] [11]。この性質は「優秀だが時々暴走する同僚」に例えられている [11]。
- **日本語情報源の軽視傾向**: 日本語での利用において、検索時に参照する情報源から日本の新聞社や専門メディア、note などのサイトがほぼ除外されるようになったとの指摘がある [3]。これは、AI の回答生成における情報源の偏りを助長し、日本語コンテンツ制作者の GEO（Generative Engine Optimization）戦略に影響を与える可能性がある [3]。
- **高価な価格設定**: 特に最上位モデルの「GPT-5.4 Pro」は非常に高価であり、主にエンタープライズ向けの価格設定となっている [1] [10]。

コミュニティと市場の反応

GPT-5.4 のリリースは、テックコミュニティから賞賛と懸念が入り混じる複雑な反応を呼んでいる。

賞賛と期待

- 多くの開発者やアーリーアダプターは、その性能向上を熱狂的に歓迎している [1]。テック系ニュースレター「The Neuron Daily」は「5.5 と呼ぶべきだった」と評し、その進化の大きさを認めている [1] [11]。コーディング性能の高さから、これまで Claude を主に使用していた開発者が GPT-5.4 へ移行する動きも見られ、AI モデル間の競争バランスが再び OpenAI に傾いたとの見方もある [6] [11]。

批判と懸念

- **誇大広告への懐疑的な見方**: 一部のベテラン研究者からは「期待外れ」「誇大広告」といった批判が継続しており、ベンチマークスコアと実用上の体感性能との間にギャップがあるという議論が続いている [1]。DEV.to の記事では「純粋な論理推論が劇的に賢くなった感覚はない」との意見も出ている [1]。
- **倫理的問題と反発**: OpenAI が米国防総省との契約を獲得したことに対し、AI の軍事利用への反発が強まっている [1] [22]。この動きに抗議して ChatGPT のサブスクリプションを解約する「QuitGPT」という運動が一部で広がりを見せている [1]。
- **思考プロセスの不透明性**: モデルが思考の過程を示す Chain-of-Thought（CoT）機能について、OpenAI は「思考を隠す能力はない」と安全性を強調しているが、ユーザーからは「思考の誤魔化しが上手くなり始めているかもしれない」と、その信頼性を警戒する声も上がっている [5] [24]。

1. GPT－5.4を触ってみた——AI開発者が感じた可能性と限界
2. GPT－5.4を1週間使ったユーザーの評価。GPT－5.4はGPT－5.2 ...
3. ChatGPT－5.4 Thinkingがnoteを見なくなった？ 日本語ソース ...
4. GPT－5.4 feels like a practical upgrade, less hype, more ...
5. OpenAI launches GPT－5.4 with Pro and Thinking versions
6. Vibe Check: GPT－5.4—OpenAI Is Back － Every
7. 【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 ...
8. 今日1日GPT－5.4を大規模現場で使ってみて、本当に突き詰め ...
9. こいつ、仕事できるぞ。有能なフロンティアモデル「GPT－5.4」発表
10. GPT－5.4は本当に、本当に優秀だ
11. Our HONEST review of GPT 5.4: They should've called it 5.5
12. r/codex － GPT 5.4 Thread － Let's compare first impressions
13. GPT－5.4 Came for Claude Code. The Real Story Is Bigger ...
14. 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 ...
15. GPT－5.4とは何か
16. OpenAI、「GPT－5.4」リリースPC操作のネイティブ対応
17. GPT－5.4の概要と登場背景：GPT－5.2から何が変わったかの ...
18. ChatGPT－5.4 is OpenAI's fastest model yet — 7 prompts ...
19. With GPT－5.4, OpenAI Promises Fewer Errors, Preps for ...
20. I hope you like spreadsheets, because GPT－5.4 loves them
21. ASCII.jp：OpenAI「GPT－5.4」発表PC操作で“人間超え”、プロ業務 ...
22. 新モデル「GPT 5.4」公開ーーハルシネーション削減、事実誤認3 ...
23. OpenAI GPT－5.4が複雑な企業文書や業界特化のメタデータ ...
24. 會澤大志｜練埜クレアAITuber （@taishi__91126） on X
25. GPT－5.4 is here — and OpenAI just made every other AI ...
26. GPT－5.4とは？業務向けAIの推論・コーディング・エージェント機能 ...
27. 【徹底解説】OpenAI「GPT－5.4」登場。実務性能とクリエイティブ ...
28. OpenAIが「GPT－5.4」をリリース、人間より上手にPCを操作 ...
29. GPT－5.4の概要｜npaka
30. OpenAI's new GPT－5.4 clobbers humans on pro－level work ...
31. OpenAIが「GPT－5.4」を発表日本のユーザーが知っておくべき5つの ...
32. 来ました。OpenAIから「GPT－5.4」が降臨。 これ

33. ChatGPT、新モデル「GPT 5.4」公開――ハルシネーション削減
34. GPT－5 の方が酷い。誰もあらかじめ作られた人格なんて求めて ...
35. GPT－5.4 ｜ The Hidden Flaws & The Fix Everyone Is Buying
36. シンギュラリティを問い直す——GPT－5.4 の先に何を見るべきか
37. ChatGPT5.4 で何が変わった？GPT－5.2・5.3 との違いと ...
38. ついに登場！ChatGPT－5.4 シリーズの全貌が明らかに！
39. How to Use ChatGPT 5.4: 6 Powerful Features You Should ...
40. プロフェッショナル向けに設計！OpenAI が GPT－5.4 モデルを ...
41. OpenAI が新 AI モデル「GPT－5.4」をリリース、性能は“大幅に” ...
42. OpenAI が新 AI モデル「GPT－5.4」をリリース、性能は“大幅に ...
43. GPT－5.3－Codex との違い・料金・使い分けまで完全整理