GPT-5.4 の総合評価：ベンチマーク、使用評価、専門家の反応に基づく分析

作成日: 2026 年 3 月 7 日 **作成者:** Manus AI

はじめに

OpenAI が 2026 年 3 月 5 日（米国時間）にリリースした最新フロンティアモデル「GPT-5.4」は、「プロフェッショナルワーク向けの最も高性能かつ効率的なモデル」として位置づけられています [1]。本レポートでは、公開されているベンチマークデータ、専門家や一般ユーザーによる使用評価、各種メディアの報道内容を多角的に分析し、GPT-5.4 の性能、新機能、実用性における長所と短所、そして今後の課題について包括的に評価します。

主な新機能と技術的進化

GPT-5.4 は、単なる対話能力の向上に留まらず、実世界でのタスク実行能力を大幅に強化する新機能を複数搭載しています。特に注目すべきは、**ネイティブコンピューター操作**、**100 万トークンのコンテキストウィンドウ**、そして**ツール検索（Tool Search）** の三点です。

ネイティブコンピューター操作は、GPT-5.4 が汎用モデルとして初めて搭載した画期的な機能です。これにより、モデルはスクリーンショットを認識し、マウスやキーボードを直接操作してデスクトップアプリケーションやウェブサイトを自律的に利用できます [1]。この機能は、これまで RPA（Robotic Process Automation）ツールなどが担ってきた定型業務の自動化を、より柔軟かつ高度なレベルで実現する可能性を秘めています。

100 万トークンのコンテキストウィンドウは、API および Codex を通じて提供され、大規模なコードベースの読解や長文の学術論文の分析など、従来モデルでは困難だった長大な文脈を要するタスクの処理能力を飛躍的に向上させます [2]。ただし、この長文コンテキスト処理能力は、特に 256K トークンを超えるあたりから性能が大

幅に低下するとのベンチマーク結果も報告されており、今後の改善が期待されます [1]。

ツール検索は、API 利用時の効率を劇的に改善する新機能です。従来、モデルは利用可能な全てのツールの定義をプロンプトに含める必要がありましたが、ツール検索機能により、必要なツール定義を動的に検索・読み込みすることが可能になりました。これにより、特に多数のツールを連携させる複雑なワークフローにおいて、トークン消費量とコストを大幅に削減できます [3]。

ベンチマークによる性能評価

GPT-5.4 は、多岐にわたる業界標準ベンチマークにおいて、先行モデルである GPT-5.2 や GPT-5.3-Codex、さらには競合モデルを凌駕する性能を示しています。特に、専門的な知識業務やコンピューター操作の領域でその能力が際立っています。

ベンチマークカテゴリ	主要な評価指標	GPT-5.4	GPT-5.2	GPT-5.3-Codex	備考
知識業務	GDPval（専門職業務）	**83.0%**	70.9%	70.9%	44 職種の専門家を上回る確率 [1]
	BigLaw Bench（法律）	**91%**	-	-	Harvey による評価 [1]
PC 操作	OSWorld-Verified	**75.0%**	47.3%	74.0%	人間のスコア（72.4%）を超える [1]
コーディング	SWE-Bench Pro	**57.7%**	55.6%	56.8%	わずかな向上 [1]
	Terminal-Bench 2.0	75.1%	62.2%	**77.3%**	エージェント的タスクでは Codex に劣後 [1]
ツール使用	BrowseComp（Web 検索）	**82.7%**	65.8%	77.3%	Pro 版は 89.3%を記録 [1]
学術・推論	ARC-AGI-2（抽象推論）	**73.3%**	52.9%	-	大幅な性能向上 [1]

ベンチマークカテゴリ	主要な評価指標	GPT-5.4	GPT-5.2	GPT-5.3-Codex	備考
医療	HealthBench	62.6%	**63.3%**	-	わずかに性能が低下 [4]

特筆すべきは、PC 操作能力を測る **OSWorld-Verified** において、人間のパフォーマンスを上回るスコアを記録した点です。これは、AI がデジタル環境における人間の作業を代替・支援する能力を具体的に示した重要な成果と言えます。一方で、エージェント的なコーディングタスク（Terminal-Bench 2.0）や一部の医療関連ベンチマーク（HealthBench）では、特化モデルである GPT-5.3-Codex や旧モデルの GPT-5.2 にわずかに及ばない結果となっており、汎用化に伴うトレードオフの可能性が示唆されます [1, 4]。

使用評価とユーザーの反応

GPT-5.4 のリリース後、開発者コミュニティや一般ユーザーからは、その性能を称賛する声と、新たな問題点を指摘する声が同時に上がっています。

肯定的な評価

多くのユーザーが、GPT-5.4 を「GPT-5.2 の高度な分析・設計能力と、GPT-5.3-Codex の優れたコーディング能力を統合した、真の汎用モデル」と評価しています [5]。特に、大規模なコードベースのリファクタリングや、複雑なアプリケーションの新規実装といったタスクにおいて、従来モデルを大幅に上回る性能を発揮したとの報告が多数寄せられています。あるユーザーは、「4 万行のコードベースの構造変更を、ほとんど修正なしに一発で完了させた」とその能力を絶賛しています [5]。また、応答速度についても「5.3-Codex よりも明らかに速い」と感じるユーザーが多く、生産性の向上に直結する点が高く評価されています [5]。

否定的な評価と課題

一方で、GPT-5.4 にはいくつかの深刻な課題も指摘されています。最も多く報告されている問題の一つが、**推論の一貫性の欠如**と**基本的な常識の欠落**です。Reddit では、「簡単なバグ修正を依頼したところ、データベースのテーブルを一度削除して

再作成するという、非常に危険なコードを提案された」という事例が報告されており、モデルの判断を鵜呑みにすることの危険性が浮き彫りになりました [5]。また、「車で 300m 先の洗車場に行くべきか、歩くべきか」という問いに対し、非現実的な前提（車が既に洗車場にあると仮定するなど）を立てて回答するなど、常識的な判断ができないケースも散見されます [6]。

さらに、日本語ユーザーにとっては、**日本語の情報源をほとんど参照しなくなる**という重大な問題が報告されています [7]。これは、GPT-5.4 で強化された「エージェント型 Web 検索」機能が、元々存在した英語ソースを優先するバイアスをさらに強めた結果と推測されています。この「GEO（Generative Engine Optimization）問題」は、日本語での情報収集能力の著しい低下を意味し、多くの日本人ユーザーから「劣化」であるとの厳しい批判を招いています [7]。

価格とコストパフォーマンス

GPT-5.4 は、性能向上に伴い価格も引き上げられました。API 価格は、入力が 100 万トークンあたり$2.50、出力が$15.00 となっており、それぞれ GPT-5.2 の$1.75、$14.00 から上昇しています [8]。

OpenAI は「トークン効率が向上したため、実質的なコストは場面によって抑えられる」と説明していますが [3]、独立系 AI 分析機関である Artificial Analysis の評価では、「同等性能のモデルと比較してやや高価」であり、特に「非常に冗長（verbose）で、不要なトークンを多く生成する傾向がある」と指摘されています [9]。実際に、あるユーザーが実施した比較テストでは、GPT-5.4 Pro は GPT-5.2 Pro と同等の正答率でありながら、コストは 2.2 倍、速度は 50%遅かったという結果も報告されており、「価格に見合った性能向上ではない」との批判的な見方も存在します [10]。

結論

GPT-5.4 は、特に専門的な知識業務の自動化とコンピューターの自律操作という領域において、AI の能力を新たな段階へと引き上げた画期的なモデルです。ベンチマークスコア、特に人間を超える PC 操作能力は、そのポテンシャルの高さを明確に

示しています。多くの開発者がその統合された高い能力を評価しており、複雑なタスクにおける生産性を大幅に向上させるツールとなり得ます。

しかし、その一方で、価格の上昇、推論の一貫性の欠如、基本的な常識の欠落、そして日本語環境における検索能力の著しい劣化といった、看過できない多くの課題も露呈しました。特に、危険なコードを生成するリスクや、特定の言語圏の情報を無視するバイアスは、AI を社会インフラとして広く展開していく上での大きな障壁となります。

結論として、GPT-5.4 は、特定の専門領域において驚異的な能力を発揮する一方で、その信頼性、安全性、公平性においては未だ発展途上にあると言えます。ユーザーは、その強力な能力を享受しつつも、モデルの出力を常に批判的に検証し、その限界を十分に理解した上で活用していく必要があります。今後の OpenAI による改善、特に安全性とバイアスに関する取り組みが、GPT シリーズの真価を決定づける重要な鍵となるでしょう。

参考文献

[1] OpenAI. (2026, March 5). *Introducing GPT-5.4*. https://openai.com/index/introducing-gpt-5-4/ [2] DataCamp. (2026, March 6). *GPT-5.4: Native Computer Use, 1M Context Window, Tool Search*. https://www.datacamp.com/blog/gpt-5-4 [3] TechCrunch. (2026, March 5). *OpenAI launches GPT-5.4 with Pro and Thinking versions*. https://techcrunch.com/2026/03/05/openai-launches-gpt-5-4-with-pro-and-thinking-versions/ [4] Trending Topics. (2026, March 5). *GPT-5.4 Targets Anthropic's Claude With Premium Pricing and Coding Muscle*. https://www.trendingtopics.eu/gpt-5-4-targets-anthropics-claude-with-premium-pricing-and-coding-muscle/ [5] Reddit. (2026, March 5). *GPT 5.4 Thread - Let's compare first impressions*. r/codex. https://www.reddit.com/r/codex/comments/1rlpyhv/gpt_54_thread_lets_compare_first_impressions/ [6] Reddit. (2026, March 6). *ChatGPT 5.4 still lacks basic common sense (...and its reasoning is inconsistent)*. r/OpenAI. https://www.reddit.com/r/OpenAI/comments/1rm7rt1/chatgpt_54_still_lacks_basic_common_sense_and_its/ [7] たぬ｜生成 AI を使い倒す人. (2026, March 6). *ChatGPT-5.4 Thinking が note を見なくなった？ 日本語ソース軽視の実態【GEO 問題】*. note. https://note.com/tank_ai/n/n08eccf0a39d8 [8] OpenAI. (2026). *API Pricing*. https://openai.com/api/pricing/ [9] Artificial Analysis. (2026). *GPT-5.4 (xhigh) - Intelligence, Performance & Price Analysis*. https://artificialanalysis.ai/models/gpt-5-4 [10] Reddit. (2026, March 5). *GPT-5.4 is more expensive than GPT-5.2*. r/singularity.

https://www.reddit.com/r/singularity/comments/1rlosvo/gpt54_is_more_expensive_than_gpt52/