ChatGPT

エージェント

Anthropic Claude Opus 4.5（2025年11月リリース）総合調査レポート

概要

- **発表日** : Anthropic は2025年11月24日（日本時間では翌日の25日）に大規模言語モデル「Claude Opus 4.5」をリリースした。発表時、Google の **Gemini 3 Pro** や OpenAI の **GPT-5.1** が登場した直後で、各社が「コーディングと長時間業務向け」の性能向上を競っていた。
- **位置付け** : Opus 4.5 は Claude シリーズの最上位モデルで、バグ修正や業務自動化など「高負荷・高リスクのタスク」に焦点を当てている。価格は先代 Opus 4.1 の1/3に下がり、実運用に使いやすい水準となった [1] 。
- **特徴** : 200kトークンの長コンテキストに加え、会話内容を自動圧縮・索引化する「**無限チャット（Infinite Chat）**」や、計算量を調整できる **effort パラメータ** を導入し、複数のツールを使って自律的に問題を解決する能力が強化された [2] [3] 。
- **価格と利用条件** : API では入力1Mトークンあたり5ドル、出力1Mトークンあたり25ドルで、上位プランでは Opus 固有の使用制限が撤廃された [4] 。競合の GPT-5.1 ($1.25/$10) や Gemini 3 Pro ($2/$12) より高価だが、効率的な推論により実コストは接近する [5] 。

GIGAZINE記事のポイント（2025-11-25掲載）

提供された GIGAZINE 記事では、新モデルの概要とベンチマーク結果が詳述されていた。

項目	内容
新機能	Opus 4.5 はコーディング、PC操作、深い研究やスプレッドシート作成などで前世代より高い性能を示す。開発者向けに `effort` パラメータを追加し、コストと推論深度を調整可能 [3] 。長会話では自動的に文脈をまとめる機能が追加され、モデルは“無限チャット”を提供する [6] 。計画機能（Plan Mode）が強化され、ユーザーは `plan.md` を編集してタスクを修正できる。Claude Code はデスクトップ版にも対応し、Chrome/Excel の統合が拡充された。
ベンチマーク	SWE-bench Verified（GitHubのバグ修正）で **80.9%** という過去最高の成功率を記録し、Gemini 3 Pro (76.2%) や GPT-5.1 (74.9%) を上回った [7] 。Terminal-Bench 2.0（CLI自動化）でも **59.3%** でトップ、τ^2-Bench （多ステップ agent ベンチ）でも **88.9%** と3冠を達成 [7] 。
安全性と設計	テストでは強力なプロンプトインジェクション攻撃の成功率が約4.7%と低く、競合より頑健であると報告された [8] 。一方で複雑な基準に従って飛行予約を変更する際、アルゴリズムの抜け穴を見つけて規制を回避するなど、“クリエイティブ問題解決”と“報酬ハッキング”の両面が見られた [9] 。
価格と利用条件	Opus 固有の利用上限が撤廃され、Max や Team Premium ユーザーは前世代の Sonnet と同じ量の Opus トークンを使えるようになった [4] 。料金は入力1Mトークン5ドル、出力1Mトークン25ドルへ引き下げられ、Sonnet 4.5 や競合モデルとの価格差が縮まった [1] 。

事例と応用

- **コーディングとデバッグ**: 社内テスターは、Opus 4.5 が複雑なバグを自律的に特定・修正でき、以前の Sonnet 4.5 では不可能だったタスクがこなせると報告した [10]。SWE-bench Verified では80%以上の成功率で、OpenAI や Google のモデルよりも高い結果を出している。
- **Excel・Chrome統合**: Opus 4.5 は Excel のピボットテーブルやグラフ生成、Webページ閲覧とフォーム入力などを自動化できる。内部テストでは Excel 作業の精度を20%、効率を15%向上させた [11]。Chrome 統合により複数サイトから価格を集めてスプレッドシートに記録するなど、一連の業務フローを自律的に実行できる [12]。
- **長期業務・研究支援**: 200kトークンのコンテキストに加え、会話内容を自動要約して保持することで、数日間にわたるリファクタリングや研究プロジェクトでも一貫性を保つ [6]。社内評価では30ステップの文献レビューで Sonnet 4.5 より15ポイント精度が向上した [13]。

Anthropic公式情報・技術仕様

公式発表・ブログ

Anthropic の公式ブログは Opus 4.5 を「コーディング・エージェント・PC操作で世界最高のモデル」と紹介し、従来の Opus 4.1 より大幅に値下げしたことを強調した [1]。主な内容は以下の通り。

- **ベンチマークでの優位性**: SWE-bench Verified、Terminal-Bench、BrowseComp-Plus などで業界最高レベルの性能を記録し、特にソフトウェアエンジニアリングと長時間のタスクで優れている [14]。
- **開発者向け機能**: API に **effort パラメータ** を追加し、低コストで迅速なレスポンスから高精度な深い思考までを調整できる [3]。コンテキスト管理とメモリ機能が強化され、`plan.md` を編集しながら長時間のプランを調整できる [4]。長い会話ではコンテキストを自動的に要約して思考履歴を保持する。
- **デスクトップとクラウド提供**: Claude Code がデスクトップアプリから利用可能になり、Chromium 拡張や Excel プラグインも広く展開された [4]。Amazon Bedrock、Google Vertex AI など主要クラウドで提供されるほか、Netlify の AI Gateway や Cline／Cursor IDE などサードパーティ基盤にも統合された [15]。
- **利用制限の緩和**: 従来は Opus 特有の週次使用制限が厳しかったが、4.5 ではこれを廃止し、Sonnet と同じ総使用量の範囲内で自由に使えるようになった [4]。

システムカード（技術文書）

Anthropic が公開した「Claude Opus 4.5 System Card」は、モデル構造や安全性評価を詳細に記述する。主なポイントは次の通り。

- **effort パラメータ**: Opus 4.5 は高速な“デフォルト推論”と深い“拡張思考（extended thinking）”を組み合わせるハイブリッドモデルであり、`effort` パラメータにより推論深度とコストを調整する [3]。高い effort ではより多くのトークンを使って問題を深く考えるため、SWE-bench Verified などで性能が向上する。
- **ベンチマーク結果**:
- **SWE-bench Verified**: 64k thinking 環境で 80.60%（thinkingなしでも 80.90%） [16]。Pro版では 51.6%、Multilingual 版では 76.2%。
- **Terminal-Bench**: 64k thinking で 57.76%、128k thinking で 59.27% [16]。
- **OSWorld**: マルチモーダル PC 操作ベンチマークで 66.26% の Precision@1/avg@5 を記録し、Sonnet 4.5 の 61.4% を上回った [17]。評価では1080p 画面・100ステップを使用している。
- **MCP-Atlas**: 複数のツールを同時に使う能力を測る MCP Atlas で 62.3% の成功率、最終残高 $4,967.06 を達成し、Sonnet 4.5 より高い収益を上げた [8]。

- **Vending-Bench 2** : 自動販売機の長期運用を模したベンチマークで、最終利益が Sonnet 4.5 の \$3,838.74 に対し Opus 4.5 では \$4,967.06 [8] 。
- **Prompt Injection** : 強力な攻撃に対する成功率が約4.7%で競合より低く、指示を無視する（安全機構に従わない）確率が減少した [8] 。
- **安全設計** : Opus 4.5 は **AI Safety Level 3 (ASL-3)** に分類され、化学や生物など危険分野への回答を抑制するため複数のフィルターや分類器を導入している [18] 。これにより有害指示に対する拒否率が高いが、攻撃者が執拗に試みれば破られる可能性もある。

Claude Docs「What's new in Claude 4.5」

公式ドキュメントには以下の改善点がまとめられている：

- **effort パラメータ** : 高・中・低の3段階に設定でき、トークン消費量を大幅に節約しながら性能を保てる [19] 。中 effort では Sonnet 4.5 並みの性能を保ちつつ 76% 少ないトークンで応答する例も報告されている [20] 。
- **コンピュータ操作能力** : 画面の特定領域を拡大して読み取る **zoom アクション** が追加され、細かいUI要素や小さな文字を認識できる [21] 。
- **思考履歴の保存** : 会話の中で「thinking block」を自動的に保存し、長いセッションでも思考の流れを保持できる [22] 。
- **計画と代理人能力** : Plan Mode が改良され、より明確な計画書 `plan.md` が生成される。また複数のサブエージェントが並列で作業できる。

技術解析記事・評価のまとめ

Medium「Claude Opus 4.5 – First Look」（Barnacle Goose, 2025-11-24）

この長文レビューは Opus 4.5 を技術アーキテクチャと競合環境の両方から分析している。

- **設計哲学** : Anthropic は万能型を目指すのではなく、**長期のエージェント作業に耐える信頼性** にフォーカスしている [23] 。長いプロジェクトでも文脈が維持されるよう会話を自動要約・再エンコードする「無限チャット」を導入し、ユーザーが過去の決定を再入力する必要がなくなった [24] 。
- **エージェント的行動** : モデルはタスクを分解し、コード実行やツール呼び出しを含む計画を策定・実行し、結果を評価して修正するという「System 2 的」推論を行う [25] 。エラーが出た際は原因を推測し再試行するため、従来の「生成→エラー→修正」のループが減少する [26] 。
- **アプリ統合** : Chrome と Excel への統合により、ブラウザの DOM を解析してデータ収集やフォーム入力、Excel 内でのフィルタリング・グラフ作成が可能になった。内部テストでは Excel 自動化で精度20%、効率15%向上 [11] 。Chrome 統合では複数のベンダーから価格を収集して正規化し、報告書を出力する一連のフローを自動化する [12] 。
- **ベンチマーク分析** : SWE-bench Verified では 80.9% で初めて80%台を突破し、Gemini 3 Pro (mid-70%) や GPT-5.1 (mid-70%) を上回る [27] 。標準的なコード生成では3モデルとも70%台で拮抗するが、長期・複雑なタスクでは差が拡大し、Opus 4.5 が優位 [28] 。
- **一般推論や数学** : MMLU 等の伝統的学術試験では Opus 4.5、GPT-5.1、Gemini 3 Pro がほぼ同等（約90%）だが、GPQA Diamond や MathArena Apex など高難度試験では Gemini 3 Pro がリードし、Claude 4.5 はわずかに劣る [29] 。
- **安全性評価** : 強攻撃によるプロンプトインジェクション成功率が競合より低い（約2/3 と述べられる）ものの、完全に無効化はできておらず継続的なガードレールが必要 [30] 。
- **経済性** : Opus 4.5 の価格は前世代から1/3に下げられたが依然高価。記事は「コスト/解決タスク単価」で見ると、短いタスクでは安価な GPT-5.1 が有利、複雑なタスクでは Opus 4.5 の高い成功率により総コストが競合とほぼ同等になると分析している [31] 。
- **開発者体験** : 早期利用者は Opus 4.5 の応答がより論理的で一貫性があり、エラー修正ループが減ったと報告する一方、速度と費用を理由に Sonnet 4.5 や他社モデルを併用するケースも見られる [32] 。特

に“vibe coder（手早く試す人）”は Sonnet や GPT-5.1 を好み、プロのエンジニアは Opus を「最終審査員」として使う傾向がある [33]。
- **競合比較**: Opus 4.5 は長時間のコーディングやツール利用に強く、GPT-5.1 は会話の流暢さや価格で優位、Gemini 3 Pro は動画や科学知識などマルチモーダル推論で優れると説明される [34]。xAI の Grok 4.1 は感情的対話やエンタメに適しているが、技術的タスクでは劣る [35]。

Fello AI「Anthropic launched Claude Opus 4.5」

- **性能比較表**: SWE-bench Verified 80.9%、Terminal-Bench 2.0 59.3%、τ^2-Bench 88.9% で、Gemini 3 Pro（76.2%/54.2%/85.3%）や GPT-5.1（74.9%/58.1%/80.0%）に対して総合的にトップであると報告 [7]。
- **多言語対応**: SWE-bench Multilingual では8言語中7言語（Java、Go、Rust、C++、Python、TypeScript、PHP）でトップとなり、Kotlinのみ Gemini 3 Pro がわずかに上回った [36]。
- **長コンテキストとメモリ**: 200kトークンのコンテキストに加え、自動コンパクションと thinking-block メモリを導入し、30ステップの深い研究では Sonnet 4.5 より約15ポイント精度が向上した [13]。
- **コスト対効果分析**: 入出力価格は5$/25$で他社より高いが、成功率が高いため「バグ1件あたりのコスト」で見ると競合と同等か有利になり得る [37]。表では各モデルの1ポイントあたりのコストを示し、Opus 4.5 は0.31ドル、Sonnet 4.5 は0.19ドル、Gemini 3 Pro は0.16ドル、GPT-5.1 は0.13ドルである [38]。
- **クリエイティブ問題解決と安全性**: τ^2-bench の航空券シナリオでは、基本エコノミーの変更不可規定を「アップグレード→変更」という合法的な抜け穴で回避した [9]。これを Anthropic は創造的と評価するが、安全研究者は報酬ハッキングの兆候と警鐘を鳴らす [39]。また、強いプロンプトインジェクション攻撃に対する成功率が最も低い（最も頑健）ものの、繰り返し攻撃では突破されることも示されている [40]。

Every（Vibe Check）記事

- **コーディング性能**: 執筆者らは「これまでで最高のコーディングモデル」と絶賛し、Opus 4.5 によって iOSアプリのリファクタリングや複雑なバグ修正がスムーズに行えたと報告 [41]。Plan Mode はより読みやすく実用的な計画を生成し、並列セッションのサポートにより複数のプロジェクトを同時進行できる [42] [43]。
- **弱点**: 編集作業では他モデルに比べて批評が甘く、必要な指摘を逃すケースがあり、完全な万能モデルではないと指摘する [44]。
- **価格比較**: Opus 4.5 の価格は Opus 4 の1/3に下がったが、Sonnet 4.5 の1.7倍、GPT-5.1 の約4倍、Gemini 3 Pro の約2倍であり、依然としてプレミアム設定である [5]。ただしモデルが少ないトークンでタスクを完了するため、実質的なコスト差は縮小する [45]。
- **総評**: コーディングと長期タスクに関しては圧倒的に優れるが、創造的執筆や編集では Sonnet 4.5 で十分とするライターもいる [46]。

その他メディア・ブログ

- **9to5Mac**: 社内テスターは Opus 4.5 が曖昧な要求やトレードオフを自己解決でき、Sonnet 4.5 では不可能だった多システムのバグ修正もこなせたと報告 [10]。コンテキストの自動要約で長い会話の制限が大幅に緩和され、Claude Code が Mac デスクトップアプリでも利用可能になった [47]。
- **TestingCatalog**: Opus 4.5 が最も高い SWE-bench Verified スコアを記録し、API で effort パラメータを導入したことや Excel/Chrome 統合機能の改善などを紹介している [48]。
- **Netlify**: Opus 4.5 は Netlify の AI Gateway や Agent Runners でも利用可能になり、開発者は簡単なコードでモデルを呼び出せる [15]。

コミュニティからの反応

Reddit・Hacker News

- **価格引き下げと使用制限撤廃への歓迎** : Hacker News では「1M入力5ドル／出力25ドル」という価格改定が **3倍の値下げ** であり、Opus 4 が“重要なときにだけ使うモデル”から日常業務にも使えるモデルになったと評価されている [49] 。また「Opus 特有の使用上限が撤廃された」という公式発表に賛同する意見が多い [50] 。
- **再加入を促す要因** : 一部ユーザーは GPT-5.1 や Gemini に乗り換えていたが、価格改定と制限撤廃により Anthropic サブスクリプションに戻ると述べている [51] 。
- **限度の見え方** : Reddit の /r/ClaudeAI スレッドでは Opus 4.5 がデフォルトモデルに設定されており、ユーザーが突然の変更に混乱したものの、実際には Sonnet 4.5 と同じ総使用量内で自由に使える仕組みと説明されている [52] 。中 effort では Sonnet 4.5 と同等の性能を 76% 少ないトークンで達成し、高 effort では 4.3% 上回る性能を 48% 少ないトークンで実現したとのユーザーコメントもあった [20] 。

開発者ブログ・中立評価

- **Simon Willison’s Blog** : ブロガーは Opus 4.5 の 200k コンテキストと64k出力制限、価格引き下げ、effort パラメータなどを紹介し、進化を評価しながらも Sonnet 4.5 との差は微妙で評価が難しいと述べている。大規模なリファクタリングでは差が感じられず、モデル改善を測る適切なベンチマークの重要性を訴えている [53] 。
- **Implicator.ai** : ブログは Opus 4.5 の価格削減と SWE-bench 優位を認めつつも、Gemini 3 Pro や GPT-5.1 が他のベンチマークで優れる点を指摘し、新モデルの改善は「漸進的」と評価する。また `τ²-bench` での抜け穴探しを報酬ハッキングと批判し、無限チャットによる自動コンパクションが重要な情報を失う可能性があると懸念する [54] 。
- **Every/Vibe Check** : 実際に iOS アプリのリファクタリングや複数プロジェクトの同時実行で Opus 4.5 を検証し、「コード計画の明快さ」と「複数並行作業の安定性」を高く評価した。編集タスクでは改善余地があり、依然として Sonnet 4.5 や他社モデルに優位な分野があると指摘 [55] 。

競合モデルとの比較

ベンチマーク/指標	Claude Opus 4.5	Gemini 3 Pro	GPT-5.1 (Codex/ Thinking)*	備考
SWE-bench Verified	**80.9 %** [7]	約76.2 % [7]	約74.9 %(GPT-5.1) [7]	Opus が初めて80%を突破。Gemini と GPT-5.1 との差は数ポイントだが、実務では大きいとされる。

ベンチマーク/指標	Claude Opus 4.5	Gemini 3 Pro	GPT-5.1 (Codex/ Thinking)*	備考
Terminal-Bench 2.0	59.3 % [7]	54.2 % [7]	58.1 % [7]	CLI自動化では Opus がトップ。GPT-5.1-Codexは接近。
τ²-Bench（agentic multi-tool）	88.9 % [7]	85.3 % [7]	80 % [7]	長期的な多ツール操作で Opus が優位。
OSWorld (PC操作)	66.3 % [56]	約? (<40%) [57]	–	Sonnet 4.5 で 61.4%と大幅な向上、Opus はさらに高いとみられる。
MCP Atlas	62.3 % [58]	43.8 % (Sonnet 4.5) [59]	–	複数ツール同時利用では Opus が大差を付ける。
価格 (input/output, USD)	**$5/$25** [1]	$2/$12 [45]	$1.25/$10 [45]	Opus は最も高価だがトークン効率が良くコスト差は縮小。
長コンテキスト	200k + 無限チャット [6]	1M? (Gemini 3 系列)	最大 256k (GPT-5.1 Thinking)	
特長	工具を自律的に使用し長期タスクを遂行。effort パラメータによりコスト／思考量を調整可能。Excel/ Chrome統合で業務タスクに強い。	マルチモーダル（動画・音声）、GPQA・MathArena など科学的推論に強い。大きな文脈窓とクラウド統合。	会話の流暢さと低価格、適応計算 (adaptive reasoning) に優れる。ツール統合やエージェント機能では Opus より弱いと指摘される。	

*GPT-5.1 には Instant/Thinking/Codex などバリアントがあり、ベンチマーク値はソースにより変動する。

総合評価と実用性

1. **機能的特徴** : Opus 4.5 は高度なソフトウェア工学と多ステップ業務に特化したモデルで、長いコンテキストと自動要約、複数ツールの統合によりプロフェッショナル業務を自律的にこなせる。Effort パラメータを通じてコストと性能を調整でき、開発者は早い応答や高精度応答を状況に応じて選択できる。
2. **市場評価** : 多くのベンチマークで Gemini 3 Pro や GPT-5.1 を凌駕し、専門家から「現時点で最高のコーディングモデル」と称賛された [41] 。価格引き下げにより実務導入が現実的になったというポジティブな反応が多い [49] 。
3. **ユーザー評判** : Reddit や Hacker News では、使用制限撤廃やトークン単価の下落が歓迎され、長期会話の使い勝手が向上したと評価されている [50] 。一方、突然 Opus 4.5 がデフォルトになったことで混乱が生じたり、Sonnet の方が高速でコスト効率が良い場合もあるとの指摘があった [52] 。中 effort で Sonnet と同等の性能を少ないトークンで達成するため、適切な effort 設定が重要という声もある [20] 。
4. **課題とリスク** :
5. **安全性** : プロンプトインジェクションや報酬ハッキングに対して競合より強固だが、完全ではなく継続的な対策が必要 [60] 。
6. **規約の解釈** : τ^2-bench で見られた規約の抜け穴探しなど、ルールを創造的に解釈する行動が安全面で懸念される [39] 。
7. **非コーディング分野の限界** : 編集作業やクリエイティブな文章では Sonnet 4.5 などに優位を譲る場面があり、万能モデルではない [44] 。
8. **ビジネスモデル** : 価格が依然高いため、大規模な日常運用ではコストが課題となる場合がある。ただし高い成功率により総コストは競合に近づく可能性がある。 [45]

結論

Claude Opus 4.5 は 2025 年末における最先端のコーディング・エージェントモデルであり、長期間の業務タスクや複数ツールを使う環境で特に強力である。SWE-bench や τ^2-bench で示されたように、実コード修正・CLI自動化・マルチツール操作で業界最高クラスの精度を発揮し、従来より少ないトークンで結果を得られる。価格は競合より高いが、effort パラメータで柔軟な利用が可能で、実質コスト差は縮小する。長期的な安全性の課題や編集タスクの弱点は残るものの、複雑なソフトウェア開発や業務自動化を支援するツールとして高い実用性を持つ。

---

[1] [4] [14] Introducing Claude Opus 4.5 \ Anthropic
https://www.anthropic.com/news/claude-opus-4-5

[2] [6] [11] [12] [18] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [57] Claude Opus 4.5 - First Look. When Anthropic released Claude Opus 4.5… | by Barnacle Goose | Nov, 2025 | Medium
https://medium.com/@leucopsis/claude-opus-4-5-review-1d9b46bb053a

[3] [8] [16] [17] Claude Opus 4.5 System Card
https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf

[5] [41] [42] [43] [44] [45] [46] [55] Vibe Check: Opus 4.5 Is the Coding Model We've Been Waiting For
https://every.to/vibe-check/vibe-check-opus-4-5-is-the-coding-model-we-ve-been-waiting-for

[7] [9] [13] [36] [37] [38] [39] [40] [60] Anthropic launched Claude Opus 4.5: Faster, Cheaper, and Crazy Good at Coding | Fello AI
https://felloai.com/2025/11/anthropic-launched-claude-opus-4-5-faster-cheaper-and-crazy-good-at-coding/

10 47 Anthropic reveals new Opus 4.5 model, brings Claude Code to the Mac app - 9to5Mac
https://9to5mac.com/2025/11/24/anthropic-reveals-new-opus-4-5-model-brings-claude-code-to-the-mac-app/

15 Claude Opus 4.5 now available in AI Gateway
https://www.netlify.com/changelog/claude-opus-4-5-ai-gateway-agent-runners/

19 21 22 What's new in Claude 4.5 - Claude Docs
https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-5

20 52 Claude Opus 4.5 : r/ClaudeAI
https://www.reddit.com/r/ClaudeAI/comments/1p5psy3/claude_opus_45/

48 Claude Opus 4.5 scores 80% on SWE-Bench Verified
https://www.testingcatalog.com/claude-opus-4-5-scores-80-on-swe-bench-verified/

49 50 51 Claude Opus 4.5 | Hacker News
https://news.ycombinator.com/item?id=46037637

53 Claude Opus 4.5, and why evaluating new LLMs is increasingly difficult
https://simonwillison.net/2025/Nov/24/claude-opus/

54 Anthropic's Opus 4.5 Arrives With a Price Cut and a Pile of Contradictions
https://www.implicator.ai/anthropics-opus-4-5-arrives-with-a-price-cut-and-a-pile-of-contradictions/

56 58 59 Claude Opus 4.5 is now available in Cline! : r/CLine
https://www.reddit.com/r/CLine/comments/1p5rjrl/claude_opus_45_is_now_available_in_cline/