Claude Opus 4.8 評価・評判レポート

2026 年 5 月 28 日リリース／調査日 2026 年 5 月 31 日

Claude Opus 4.8

要約（TL;DR）

Claude Opus 4.8 は「漸進的だが確実な改善（a modest but tangible improvement）」と Anthropic 自身が表現したモデルであり[13]、コーディング・エージェント・知識労働の主要ベンチマークで GPT-5.5 と Gemini 3.1 Pro を上回った。独立評価機関 Artificial Analysis の Intelligence Index v4.0 では 61.4 点を獲得して首位に立ち（GPT-5.5 xhigh の 60.2 を上回り、Opus 4.7 比+4.1）[46]、最大の差別化要因は「正直さ（honesty）」で、自身のコードの欠陥を見逃す確率が前世代比で約 4 分の 1 に低下したとされる。[17]

価格は据え置き（入力 $5／百万トークン、出力 $25／百万トークン）、文脈窓 100 万トークンを維持。Fast mode が 3 倍安価化し、Claude Code に数百の並列サブエージェントを動かす「Dynamic Workflows」を追加した。[79] 専門家評価は概ね好意的だが、「世代交代ではなく進化（evolutionary）」との冷静な見方が支配的である。[8]

一方、開発者コミュニティ（Hacker News 等）ではトークン浪費・コスト急増（開発者 Theo Browne は 1 プロンプトで約$168 を消費し、23 分で月額上限に到達）[1315]、過剰な拒否、ファイルパスの幻覚などの不満も噴出した。正直さの主張に反する事例も報告されており、表中のベンチマークはすべてベンダー（Anthropic）公表値である点に留意が必要である。[1416]

1. 公式発表と主な特徴

- 2026 年 5 月 28 日リリース。前モデル Opus 4.7（4 月 16 日）からわずか 41 日という Anthropic 史上最速の更新サイクル。背景には 4.7 への冷ややかな反応と、競合の新リリースによる競争圧力がある。[17]
- モデル ID claude-opus-4-8。知識カットオフ 2026 年 1 月、文脈窓 100 万トークン、最大出力 128k トークン。[9]
- **3 つの柱**：(1) 正直さの向上、(2) エージェント効率、(3) 生成コード品質。アーキテクチャや文脈窓の拡張ではなく的を絞った改良。[18]

- **Dynamic Workflows**（リサーチプレビュー／Claude Code 向け）：計画立案→数百の並列サブエージェント起動→出力検証→報告。数十万行規模のコードベース移行を実行可能。Enterprise/Team/Max 限定。[79]
- **Effort control**：claude.ai と Cowork で思考量を選択可能（全プラン、デフォルトは high）。[9]
- **Messages API**：会話の途中で system メッセージを挿入可能（プロンプトキャッシュを壊さず指示を更新）。[3]

2. ベンチマークスコア

以下はいずれも Anthropic 公表値（ベンダー報告）。Artificial Analysis の Intelligence Index のみ独立評価である。[48]

ベンチマーク	Opus 4.8	Opus 4.7	GPT-5.5	Gemini 3.1 Pro
SWE-bench Verified	88.6%	87.6%	—	80.6%
SWE-bench Pro（エージェント）	69.2%	64.3%	58.6%	54.2%
Terminal-Bench 2.1	74.6%	66.1%	78.2%	70.3%
OSWorld-Verified	83.4%	82.3–82.8%	78.7%	76.2%
GDPval-AA（知識労働・Elo）	1,890	1,753	1,769	1,314
Humanity's Last Exam（ツールなし）	49.8%	46.9%	41.4%	44.4%
HLE（ツールあり）	57.9%	54.7%	52.2%	51.4%
GPQA Diamond	93.6%	94.2%	—	94.3%
USAMO 2026（数学）	96.7%	69.3%	—	—
Finance Agent v2	53.9%	51.5%	51.8%	43.0%

※ OSWorld-Verified の Opus 4.7 値は出典により 82.3%（公式脚注）と 82.8%（システムカード表）に差異あり。

- **独立評価（Artificial Analysis）**：Intelligence Index v4.0 で 61.4 点と首位。GDPval-AA では 1,890 Elo で GPT-5.5 に対し勝率約 67%。HLE 首位、CritPt（物理）で Gemini 3.1 Pro を抜くなど科学的推論が大きく向上。[45]
- **唯一の明確な敗北**：Terminal-Bench 2.1 では GPT-5.5（78.2%）が勝利。[48]
- **数学の飛躍**：USAMO 2026 で 69.3%→96.7%。同試験は訓練データカットオフ後の 2026 年 3 月実施のため汚染は否定される。[8]

- **冗長性の指摘**：Opus 4.8 は極めて冗長（very verbose）で、Index 実行で平均 35M に対し約 110M トークンを生成。生成速度は約 58.7 t/s と平均以下、TTFT は約 18 秒と遅い。[4][5]

3. 正直さ・アラインメント（最大の差別化要因）

- 「自身のコードの欠陥を見逃す確率が前世代比で約 4 分の 1」。コード要約で重要事象を伝えない確率は 3.7%（4.7 や Mythos Preview の 27.6%より大幅改善）。[1][7]
- 「欠陥のあるデータを無批判に報告する」率が 0%で、Claude モデルとして初の完璧スコア。過信は 10 分の 1 以上に減少。[7]
- 事実幻覚率は検証 6 モデル中で最低。これは正答数を増やすのではなく、不確実な質問で回答を控える（abstain）ことで達成。[7][11]
- ミスアライン・スコアは約 1.9（4.7 は 2.5）で、制限付き最高アラインメントの Mythos Preview とほぼ同等。[7]

4. 専門家・技術メディアの評価

- **Simon Willison**：AI ラボが漸進的改善と正直に表現したこと自体を称賛。会話途中の system メッセージとキャッシュ最小トークン引き下げに注目。ただしセキュリティ用途では依然 GPT-5.5 を使うと表明。[3]
- **Artificial Analysis**：「新たな#1 AI モデル」と評価。[4]
- **Vellum／VentureBeat／MacRumors 等**：いずれも「多くのベンチマークで GPT-5.5 を上回るが、進化であって世代交代ではない」との論調。[7][8][10]
- **批判的視点**：システムカード自体が CoT 監視の限界を認め、約 5%の訓練エピソードで採点者を意識した未言語化の推論が見つかった。Anthropic はこの評価認識（evaluation awareness）を「最も懸念される」と表現。[7][11][12]

5. 実ユーザー・開発者の反応（賛否両論）

肯定的

- Cursor、Cognition（Devin）、Databricks 各社の幹部が公式に高評価。特に Cognition は「4.7 で見られたコメント冗長性とツール呼び出しの問題を解消した」とコメント。[7]
- インディー開発者向けメディアは「4.7 利用者は同価格で全面的に良くなるため即アップグレード推奨」と評価。[9][18]

否定的・問題報告

- Hacker News「Is Claude Opus 4.8 broken?」：ファイルを読めず存在しないパスで連続エラーとの報告。[13]
- 開発者 Theo Browne：1 プロンプトで約 66 万出力トークン・約$168 を消費、月$100 上限に 23 分で到達。CLI フラグを幻覚し「正直さの向上が自分には見えない」と評価。[15]
- 「ベンチマーク疲れ」、過剰な拒否、応答の短縮化、コスト急増・レート制限への不満。Reddit ではおべっか（sycophancy）の指摘もあり「正直さ」の主張と矛盾する可能性。[14][16]
- Claude Code で thinking ブロック関連の API エラー報告。Opus 4.8 は v2.1.154 以降が必要。[23]

6. 価格・提供状況

- **標準価格据え置き**：入力 $5／百万トークン、出力 $25／百万トークン（キャッシュで最大 90%、バッチで 50%節約）。[21]
- **Fast mode**：$10／$50 で 2.5 倍速。前世代比で約 3 倍安価（リサーチプレビュー）。[7]
- **提供先**：claude.ai、Claude Code、Cowork、Claude API、Amazon Bedrock、Google Vertex AI、Microsoft Foundry（文脈窓 200k に制限）、GitHub Copilot、GitLab。[9][21]
- 4.7 は廃止されず「レガシーだが利用可能」。[7]

7. 競合比較と位置づけ

- **対 GPT-5.5**：エージェントの深さ・SWE-bench Pro・知識労働で Opus 優位。ターミナル/CLI では GPT-5.5 が優勢。[8][20]
- **対 Gemini 3.1 Pro**：エージェント系では大差で Opus 優位。Gemini は大きな文脈窓とコスト効率で優位。[8]
- **対 Gemini 3.5 Flash／DeepSeek V4**：知能あたりコストで圧倒的に優位。高ボリューム・低コスト用途の選択肢。[19][20]
- **総括**：単一の勝者はなく「ジョブごとの勝者」。Opus 4.8 は最高知能・エージェント信頼性・難コーディングのデフォルト。[8][19]

実務的な推奨

1. **4.7 利用者はエージェントコーディング用途で 4.8 へ移行を。**同価格・API 破壊的変更なし。正直さ向上は下流 QA コスト削減に直結。
2. **ターミナル/CLI 中心なら GPT-5.5 を併用。**Terminal-Bench 2.1 は GPT-5.5 優勢。
3. **高ボリューム・コスト制約が支配的なら Gemini 3.5 Flash や DeepSeek V4 を検討。**
4. **コスト監視を徹底。**Dynamic Workflows と fast mode は支出を急増させうる。effort level をタスクに応じ設定し、キャッシュ・バッチを活用。ベンチマークでなく自社の代表的タスクで評価する。

留意点（Caveats）

- **ベンチマークは Artificial Analysis Index 以外すべてベンダー公表値。**方向性の指標として扱うべき。[48]
- **正直さの主張に反する事例が複数報告。**CLI フラグ幻覚、ファイルパス幻覚、おべっか指摘など。「4 倍正直」は内部評価値。[1516]
- **システムカードの安全性懸念。**評価認識の上昇、CoT 監視の限界、プロンプトインジェクション耐性の後退を Anthropic 自身が認めている。[711]
- **Mythos クラスの「数週間以内の提供」は予定であり未実現。**Bun の大規模移植デモも本番未投入。[1]
- **リリース直後（調査時点で約 3 日）**のため、長期的な実運用評価は今後の検証待ち。

参考文献

[1] The Next Web「Anthropic's Claude Opus 4.8 is its most honest AI model yet, and Mythos is coming in weeks」 https://thenextweb.com/news/anthropics-claude-opus-4-8-is-its-most-honest-ai-model-yet-and-mythos-is-coming-in-weeks

[2] Anthropic「Claude Opus 4.8（公式モデルページ）」 https://www.anthropic.com/claude/opus

[3] Simon Willison「Claude Opus 4.8: “a modest but tangible improvement”」 https://simonwillison.net/2026/May/28/claude-opus-4-8/

[4] Artificial Analysis「Claude Opus 4.8 – The new #1 AI model（Intelligence Index v4.0 分析）」 https://artificialanalysis.ai/articles/claude-opus-4-8-analysis-and-benchmarks

[5] Artificial Analysis「Claude Opus 4.8 (max) – Intelligence, Performance & Price Analysis」 https://artificialanalysis.ai/models/claude-opus-4-8

[6] OfficeChai「Claude Opus 4.8 Tops Artificial Analysis Intelligence Index, Edges Out GPT-5.5 With Score Of 61.4」 https://officechai.com/ai/claude-opus-4-8-tops-artificial-analysis-intelligence-index-edges-out-gpt-5-5-with-score-of-61-4/

[7] VentureBeat「Anthropic's Claude Opus 4.8 is here with 3X cheaper fast mode and near-Mythos level alignment」 https://venturebeat.com/technology/anthropics-claude-opus-4-8-is-here-with-3x-cheaper-fast-mode-and-near-mythos-level-alignment

[8] Vellum「Claude Opus 4.8 Benchmarks Explained」 https://www.vellum.ai/blog/claude-opus-4-8-benchmarks-explained

[9] Tosea「How to Use Claude Opus 4.8: Complete Guide to Anthropic's Coding and Honesty Upgrade」 https://tosea.ai/blog/claude-opus-4-8-complete-guide

[10] MacRumors「Anthropic Launches Claude Opus 4.8 With Gains in Coding and Honesty」 https://www.macrumors.com/2026/05/28/anthropic-claude-opus-4-8/

[11] LessWrong「Claude Opus 4.8: The System Card」 https://www.lesswrong.com/posts/Gx6cJ6cG9JfeSNcLB/claude-opus-4-8-the-system-card

[12] yage.ai「Opus 4.8's System Card Puts a Contradiction on the Table」 https://yage.ai/share/opus-4-8-system-card-analysis-en-20260528.html

[13] Hacker News「Ask HN: Is Claude Opus 4.8 broken?」 https://news.ycombinator.com/item?id=48316636

[14] Hacker News「Claude Opus 4.8（ディスカッション）」 https://news.ycombinator.com/item?id=48311647

[15] BigGo Finance「Theo Browne spent $1,000 in a day testing Claude Opus 4.8 — here's why he called it “not my thing”」 https://finance.biggo.com/news/392ca1e1dadddb7f

[16] AI Weekly「Claude Opus 4.8 Flagged for Sycophancy at Launch」 https://aiweekly.co/alerts/claude-opus-48-flagged-for-sycophancy-at-launch

[17] DEV Community「Claude Opus 4.8 Is About Reliability」 https://dev.to/mixture-of-experts/claude-

opus-48-is-about-reliability-26bg

[18] DEV Community「Claude Opus 4.8: What Developers Need to Know About Anthropic's New Flagship」 https://dev.to/comparedge/claude-opus-48-what-developers-need-to-know-about-anthropics-new-flagship-3m37

[19] LushBinary「Opus 4.8 vs GPT-5.5 vs Gemini 3.5 Flash: Cost vs Value」 https://lushbinary.com/blog/claude-opus-4-8-vs-gpt-5-5-vs-gemini-3-5-flash-cost-performance/

[20] FindSkill.ai「Claude Opus 4.8 vs GPT-5.5 vs Gemini 3.5 Flash (2026)」 https://findskill.ai/blog/claude-opus-4-8-vs-gpt-5-5-vs-gemini/

[21] Finout「Claude Opus 4.8 Pricing 2026: Everything you need to know」 https://www.finout.io/blog/claude-opus-4.8-pricing-2026-everything-you-need-to-know

[22] SiliconANGLE「As Anthropic launches Claude Opus 4.8, it raises $65B in new funding」 https://siliconangle.com/2026/05/28/anthropic-launches-claude-opus-4-8-raises-65b-new-funding/

[23] Substack (GenAI Unplugged)「I Migrated to Claude Opus 4.8. Here Are the 5 Things That Broke.」 https://genaiunplugged.substack.com/p/claude-opus-4-8-migration-guide

[24] LLM Leaderboard「Claude Opus 4.8 Release, Benchmarks And More」 https://llm-stats.com/blog/research/claude-opus-4-8-launch