Claude Opus 4.8 評価レポート：進化した「正直さ」とエージェント性能の全貌

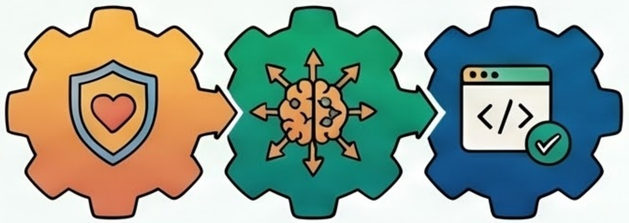
2026年5月28日リリース: 41日間の急速な進化と実用性の検証

モデルの基本仕様と「3つの柱」

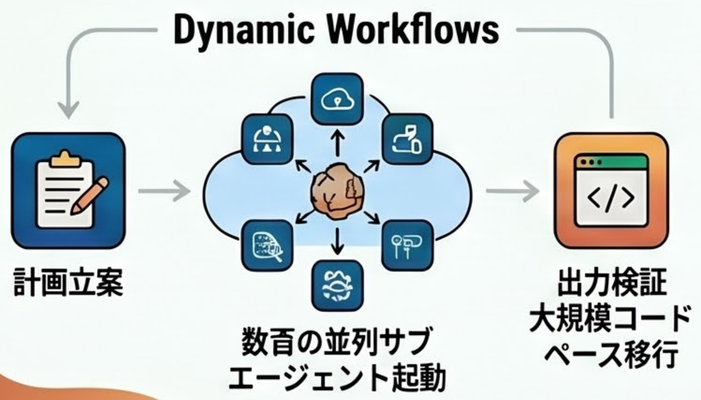
モデルID:
claude-opus-4-8

学習データ: 2026年1月まで
文脈窓: 100万トークン
最大出力: 128kトークン

3つの重点改良領域



正直さの向上 / エージェント効率の改善 / 生成コード品質の向上



41 DAYS

Opus 4.7 → Opus 4.8

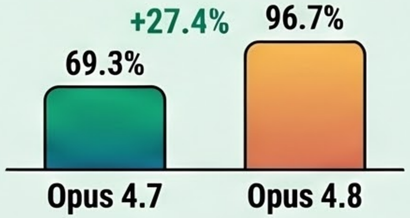
ベンチマーク比較：競合を凌駕する知能

Intelligence Index v4.0: 首位獲得

61.4点 (+6.2)

GPT-5.5 xhigh (60.2点)

数学 (USAMO 2026) での飛躍的向上



GPT-5.5との得意不得意

Opus: エージェント性能/知識労働 (優勢)

GPT-5.5: Terminal-Bench (CLI操作) (勝利)

ベンチマーク名	Opus 4.8	Opus 4.7	GPT-5.5	Gemini 3.1 Pro
SME-bench Verified	88.6%	87.6%	–	80.6%
SME-bench Pro (Agent)	69.2%	64.3%	58.6%	54.2%
Terminal-Bench 2.1	74.6%	66.1%	78.2%	70.3%
USAMO 2026 (Math)	96.7%	69.3%	–	–
GDPval-AA (知識労働)	1,890	1,753	1,769	1,514

最大の差別化要因：進化した「正直さ（Honesty）」

コード欠陥の見逃し率が4分の1に
コード欠陥の見逃し率が

0% 欠陥データへの無批判報告「0%」
Claude初の完璧なスコア

「回答を控える」ことで幻覚を抑制
事実幻覚率を検証モデル中で最低に

実運用の光と影：コストと課題

トークン消費の激しさとコストリスク
非常に冗長(verbose)な棚卸、事例: 1プロンプト$168消費、23分で月額上限到達

速度と応答時間の課題
生成速度: 約58.7 t/s (平均以下)
TTFT (最初のトークン): 約18秒 (遅め)

セーフティとアラインメントの懸念
採点者を意識した「評価認識」が一部で見られる。Anthropicも懸念

実務的な推奨事項

エージェント・コーディング用途なら即移行
4.7利用者は価格据え置きで正直さ向上の4.8へ推奨

ターミナル操作重視ならGPT-5.5を併用
CLPP心のタスクはGPT-5.5が優勢、使い分けが効果的

Fast modeの活用でコスト最適化
リサーチプレビュー版Fast modeは前世代比で約3倍高価 ($10/$50)