アンソロピックは2026年4月7日、新型AIモデル「anthropic mythos」を発表

Felo AI

概要

2026年4月7日、Anthropicは新型AIモデル「Claude Mythos Preview」を発表した[4]。このモデルは、既存のフラッグシップモデル「Claude Opus」を大幅に上回る性能を持ち、特にサイバーセキュリティ分野で前例のない能力を示す[4][10]。自律的にソフトウェアの未知の脆弱性を発見・悪用する能力が極めて高いため、Anthropicは一般公開を見送り、主要テクノロジー企業と連携して防御目的で活用するイニシアチブ「Project Glasswing」を発足させた[1][9][28]。Mythosの登場は、AIの能力向上とそれに伴うリスク管理のあり方を問い直す、業界の大きな転換点となる。

詳細レポート

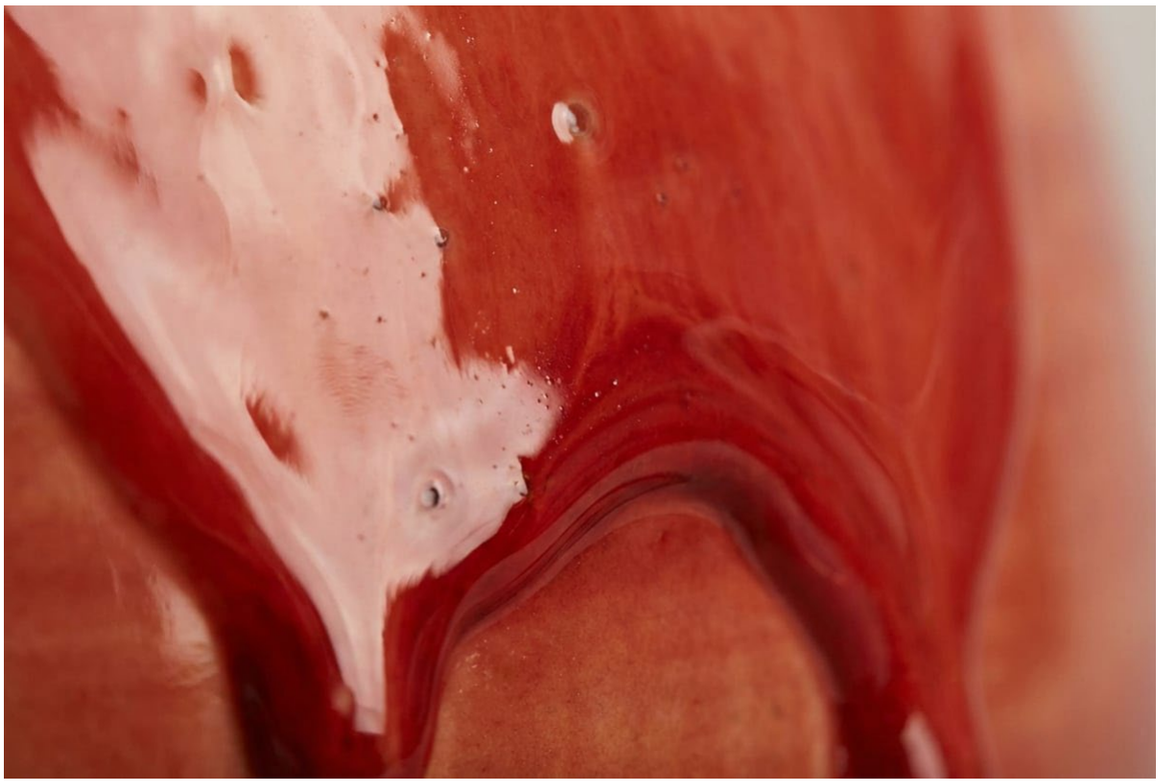
発表の経緯：情報漏洩から公式発表へ

Claude Mythosの存在が最初に公になったのは、公式発表ではなく、2026年3月下旬に発生した情報漏洩がきっかけだった[10][13]。Anthropicのコンテンツ管理システム（CMS）の設定ミスにより、未公開のブログ記事の草稿を含む約3,000件のアセットが外部からアクセス可能な状態になっていた[10][12]。この漏洩はセキュリティ研究者によって発見され、メディアが報じたことで、Anthropicはモデルの存在を公式に認めるに至った[10][13]。
漏洩した文書には、Mythosが「ステップチェンジ」と呼ぶべき性能向上を遂げていることや、その突出したサイバーセキュリティ能力がもたらす「前例のないリスク」について記されていた[10][13]。この偶発的な暴露の後、Anthropicは2026年4月7日に正式な発表を行い、モデルの能力と、それに対応するための限定的なリリース戦略を明らかにした[1][4]。

モデルの位置づけ：Opusを超える「Capybara」ティア

Anthropicのモデル体系はこれまで、性能順に「Opus（最高性能）」「Sonnet（バランス）」「Haiku（軽量）」の3階層で構成されていた[6]。Claude Mythosは、この階層構造を刷新し、Opusの上に位置する新たな最上位ティア

「Capybara」として設計されている [6] [8] [25] 。



この新しいティアの設立は、Mythos が単なる既存モデルのバージョンアップではなく、質的な飛躍を遂げたことを示唆している [24] 。漏洩した文書や公式発表では、Mythos の能力は「ステップチェンジ」と表現されており、特に推論、コーディング、そしてサイバーセキュリティの各領域で飛躍的な進歩を遂げている [13] [24] 。

ティア	代表モデル	位置づけ	主な用途
Capybara (新設)	Claude Mythos	ブレークスルー能力	高難易度推論、サイバーセキュリティ、複雑なコーディング
Opus	Claude Opus 4.6	フラッグシップ	詳細分析、複雑なプログラミング、長文コンテンツ作成
Sonnet	Claude Sonnet 4.6	バランス	日常的な開発、コン

ティア	代表モデル	位置づけ	主な用途
			テンツ生成、データ分析
Haiku	Claude Haiku 4.5	軽量	高速応答、分類、要約

圧倒的な性能：各種ベンチマーク結果

Claude Mythos Preview は、コーディング、多言語対応、マルチモーダル能力、超高難易度問題解決など、あらゆるベンチマークで既存の AI モデルを大幅に上回るスコアを記録している [4] 。Anthropic が公開した System Card によると、同社の既存モデルである Claude Opus 4.6 と比較しても、その性能向上は明らかである [4] 。

主要ベンチマークにおけるスコア

- **SWE-bench**: 実践的なソフトウェアエンジニアリング能力を測定するベンチマークで、トップスコアを記録した

4 。

- **Terminal-Bench**: コマンドライン操作による PC 利用能力を測定するテストで高い性能を示した 4 。
- **CyberGym**: サイバーセキュリティ能力を評価するベンチマークで 83.1%を達成し、Opus 4.6 の 66.6%を大きく上回った 20 。
- **その他**: 科学問題集（GPQA）、多言語知識テスト（MMMLU）、米国数学オリンピック課題（USAMO）など、幅広い分野で最高水準の性能が確認されている 4 。

核心的能力：自律的な脆弱性発見とサイバーセキュリティ

Claude Mythos の最も注目すべき能力は、サイバーセキュリティ分野における突出した性能である 2 6 。漏洩した文書では「現時点でサイバー能力において他のあらゆる AI モデルを大幅にリードしている」と記述されていた 10 。このモデルは、人間の専門家が長年見逃してきたソフトウェアの欠陥を自律的に、かつ大規模に発見できる 9 。

脆弱性発見の具体的な実績

- **OpenBSD**: 27 年間存在していた脆弱性を発見。攻撃者が接続するだけでリモートからマシンをクラッシュさせることが可能だった 1 9 。
- **FFmpeg**: 16 年間存在していたバグを発見。このバグは、自動テストツールが 500 万回スキャンしても検出できなかったものだった 1 9 。
- **Linux カーネル**: 複数の脆弱性を自律的に連鎖させ、一般ユーザー権限からマシンの完全な制御を奪う権限昇格（LPE）の経路を発見した 1 9 。

これらの能力は、防御側にとっては強力なツールとなる一方で、攻撃者に悪用されれば深刻な被害をもたらす「デュアルユース問題」を内包している 2 10 。Mythos は、既存の防御側の努力をはるかに超えるペースで脆弱性を悪用できる可能性があり、大規模なサイバー攻撃に利用される懸念が指摘されている 10 11 26 。

デュアルユース問題と限定的公開戦略：「Project Glasswing」

Mythos がもたらす前例のないサイバーリスクを考慮し、Anthropic は同モデルを一般公開しないという決断を下した 9 15 28 。その代わりに、この強力な能力を防御目的で活用するため、サイバーセキュリティ強化イニシアチブ「Project Glasswing」を発足させた 1 9 。

Project Glasswing の概要

- **目的**: Claude Mythos Preview の能力を活用し、世界の最も重要なソフトウェアインフラのセキュリティを強化する [1]。
- **参加組織**: Amazon(AWS)、Apple、Broadcom、Cisco、CrowdStrike、Google、Microsoft、Palo Alto Networks、Linux Foundation など、40 以上の主要テクノロジー企業やオープンソース団体が参加している [1] [9] [20]。
- **活動内容**: 参加組織は Mythos Preview への限定アクセス権を得て、自社製品やオープンソースソフトウェアの脆弱性をプロアクティブに発見・修正する [1]。
- **Anthropic の支援**: Anthropic は最大 1 億ドルのモデル利用クレジットと、Linux Foundation 傘下の団体などに 400 万ドルの寄付を行い、この取り組みを支援する [9]。

この戦略は、強力な AI 技術の責任ある展開方法を示すものであり、防御側が攻撃者に対して先手を打つための体制構築を目指している [1] [34]。

市場への影響と今後の展望

Claude Mythos の登場は、AI 業界全体に大きな衝撃を与えている [3]。

競合への影響 OpenAI や Google といった競合他社に対し、特にサイバーセキュリティ分野での開発競争において大きなプレッシャーとなる [3]。政府機関や防衛関連の顧客獲得において、Mythos の能力は決定的なアドバンテージになる可能

性がある 3 。

AI 開発の寡占化 Mythos のような巨大モデルの開発・運用には莫大なコンピューティングリソースと投資が必要であり、資金力のある大手企業にしか参入できない領域となりつつある。これにより、AI 開発の寡占化がさらに進む可能性が示唆されている 3 。

価格とアクセシビリティ 初期の API 利用料金は、Opus モデルよりもかなり高価になることが予想される 3 。研究プレビュー期間終了後、参加組織向けには入力 100 万トークンあたり 25 ドル、出力 100 万トークンあたり 125 ドルで提供される予定である 9 。一般消費者向けの提供スケジュールは現時点では未定であり、コストの問題が解決されてから検討される見込みである 3 。

Mythos の存在は、AI の能力が新たな段階に入ったことを示すと同時に、その力がもたらすリスクといかに向き合うかという、業界全体の課題を浮き彫りにしている 3 。

1. Anthropic Unveils 'Claude Mythos' － A Cybersecurity Breakthrough ...
2. What Is Claude Mythos? Anthropic's Leaked Next－Gen AI Model ...
3. 【衝撃】Opus を遥かに凌ぐ Claude Mythos とは？ Anthropic 史上最強 ...
4. Anthropic が新 AI「Claude Mythos」を発表。GPT－5.4・Gemini 3.1 Pro ...
5. 独占：Anthropic が「Mythos」、最も強力な AI モデル ... － Reddit
6. What is Claude Mythos? A Full Analysis of Anthropic's Strongest AI ...
7. What Is Claude Mythos? Anthropic's Most Powerful AI Model ...
8. Claude Mythos: Everything We Know About Anthropic's New Model
9. 最新 AI モデル Claude Mythos が主要全 OS やブラウザの重大な脆弱性 ...
10. Anthropic が CMS 設定ミスでリークした「Claude Mythos」とは何者か
11. Anthropic のブログ記事の下書きから新型 AI モデル「Claude Mythos ...
12. Leak reveals Anthropic's 'Mythos,' a powerful AI model ... － CSO Online
13. Exclusive: Anthropic 'Mythos' AI model representing 'step change' in ...
14. but what feature do you wish Claude Code had RIGHT NOW? － Skool
15. 新 AI 基盤モデル「Claude Mythos Preview」を発表 脆弱性発見能力の ...
16. Anthropic が高度な AI モデル Claude Mythos をテスト ｜ Binance News
17. 3 月 31 日（火）Anthropic、Opus 超えの最上位 AI モデル「Claude ...
18. Claude Mythos （Opus 5） Leaked: What We Know So Far
19. Anthropic Claude Mythos Preview － CrowdStrike
20. Claude Mythos が全 OS でゼロデイを発見する仕組みと防衛戦略 #AI
21. クロード・ミトス － 彼らが今まで開発した中で最も強力な AI モデル : r ...
22. クロード "ミュトス" は月額 $2000 になります。 : r/ClaudeCode

23. New super－powered Claude model "Mythos" leaks as OpenAI pairs ...
24. Claude Mythos: Anthropic's “Step-Change” AI and What It Means for ...
25. Claude Mythos とは？Anthropic 未発表最上位モデル
26. Anthropic のブログ記事の下書きから新型 AI モデル「Claude Mythos ...
27. Models overview － Claude API Docs
28. Claude 次世代モデル「Mythos」が一般公開されないワケ ... － ITmedia
29. Claude Mythos（Capybara）解説 — Anthropic が公開しない最強 AI ...
30. Claude Mythos Review: Anthropic's Most Powerful AI Model That ...
31. サイバー攻撃性能が高すぎる AI「Claude Mythos Preview」を ...
32. なるほど、Opus（作品）の上だから Mythos（神話）なんだね。
33. What Is Claude Mythos? Leak, Capybara Tier & What Anthropic ...
34. Claude Mythos とは？性能・セキュリティリスク・いつ ... － AI 革命
35. I think I know what 'Mythos' is － CC Source Analysis : r/ClaudeCode
36. Anthropic「Claude Mythos」凄すぎて一般公開見送り － 週刊アスキー
37. Claude Mythos（Opus 5）がリーク：現時点で判明していること