Anthropic「Claude Fable 5」および新ティア「Mythos」クラスに関する調査報告

（2026 年 6 月 9 日（米国時間）発表／本報告作成日：2026 年 6 月 10 日）

Claude Fable 5

1. 要旨

米 Anthropic は 2026 年 6 月 9 日（米国時間）、従来の最上位クラスである Opus のさらに上位に位置づけられる新ティア「Mythos（ミュトス）クラス」のモデルとして、一般ユーザー向けの「Claude Fable 5」と、承認された組織のみに提供される「Claude Mythos 5」を同時に発表した[1]。両モデルは同一の基盤モデル（同じ重み）を共有しており、相違点は安全対策（セーフガード）の有無のみである[1]。

Fable 5 は、SWE-bench Verified 95.0%をはじめとする公開ベンチマークの大半で Claude Opus 4.8、GPT-5.5、Gemini 3.1 Pro を上回る、Anthropic 史上最も高性能な一般提供モデルである[1]。価格は入力 100 万トークンあたり 10 ドル、出力 100 万トークンあたり 50 ドルで、Opus 4.8 の 2 倍に設定された[4]。デュアルユース（軍民両用）リスクのあるサイバー・生物・化学・蒸留分野のリクエストは、分類器により検知され、自動的に Claude Opus 4.8 が応答を引き継ぐ仕組みが導入されている[1]。

知財実務の観点では、文書推論・長文脈処理・法務タスクへの強み（法務 AI プラットフォーム Harvey の Legal Agent Benchmark で過去最高の 13.3%[10]）から、先行技術調査、FTO 分析、明細書ドラフト等への応用余地が大きい。一方、30 日間のデータ保持義務、米国内処理、高コスト、分類器の誤検知による応答品質低下といったリスクがあり、機密性の高い知財業務への投入には慎重な検討を要する。

2. 発表の概要

Anthropic の公式発表「Claude Fable 5 and Claude Mythos 5」[1] および API ドキュメント[2] で確認した発表の要旨は以下のとおりである。

・Fable 5 は「当社がこれまで一般提供したいかなるモデルをも上回る能力」を持ち、ソフトウェア開発、知識労働、画像認識、科学研究など、テスト対象のほぼ全ベンチマークで最高水準とされる。タスクが長く複雑になるほど他モデルとの差が広がる[1]。

・Mythos 5 は「サイバー防御者とインフラ提供者の小集団」向けに提供される。米政府と連携した枠組み「Project Glasswing」経由で展開され、4 月に限定提供された「Claude Mythos Preview」のアップグレード版に当たる[1]。

・価格は両モデルとも入力 100 万トークン 10 ドル／出力 100 万トークン 50 ドル（Mythos Preview の半額未満）[1,2]。

3. モデルの位置づけと構造

（1）ティア構造　Anthropic 公式の脚注によれば、Mythos クラスは「Opus クラスを能力面で上回る Claude モデルのティア」であり、Haiku＜Sonnet＜Opus＜Mythos の序列となる[1]。Fable 5 は「Mythos クラスの能力を一般向けに安全化したモデル」と位置づけられる[1]。

（2）二重リリース構造　両者は従来の「小型版と大型版」という区別ではなく、同一の基盤能力を共有し、アクセス制御とガードレールの違いのみで分かれる。「Fable」はラテン語の fabula（語られるもの）に由来し、ギリシャ語の mythos（神話）と類義であり、安全対策の有無で同一モデルを 2 銘柄に分けたことを命名で表現している[1]。Mythos 5 にアクセスできるのは、Glasswing のサイバーセキュリティパートナーや一部の生命科学研究者など、審査を経た組織に限られる[1]。

（3）API 仕様　API モデル名は Fable 5 が「claude-fable-5」、Mythos 5 が「claude-mythos-5」。両者とも標準で 100 万トークンのコンテキストウィンドウ、リクエストあたり最大 12.8 万トークンの出力に対応する。適応的思考（adaptive thinking）が常時オンであり、思考の無効化（thinking: disabled）は指定できず、生の chain-of-thought は返されない[2]。

4. 性能・ベンチマーク

Anthropic が公表した主要ベンチマークの比較は下表のとおりである[1]。なお、一部の値は一般提供されない Mythos 5 での測定値であり、Fable 5 はセーフガードの作用によりサイバー・生物分野等で実効性能が Opus 4.8 寄りになる場合がある点に注意を要する。

ベンチマーク	Fable 5 /	Opus 4.8	GPT-5.5	Gemini 3.1 Pro

	Mythos 5			
SWE-bench Verified	95.0%	88.6%	82.6%	78.8%（3.5 Flash）
SWE-Bench Pro	80.3%	69.2%	58.6%	54.2%
FrontierCode Diamond	29.3%	13.4%	5.7%	—
Terminal-Bench 2.1	88.0%	82.7%	83.4%	70.7%
GDPval-AA	1932	1890	1769	1314
Humanity's Last Exam（ツールなし）	59.0%	49.8%	41.4%	44.4%
Legal Agent Benchmark	13.3%	10.4%	2.1%	0.0%
GDP.pdf（視覚文書推論）	29.8%	22.5%	24.9%	16.7%

これらは Anthropic 自身の公表値であるが、SWE-Bench Pro や Humanity's Last Exam 等、OpenAI が自社公表した数値と重なる項目では概ね整合している。SWE-bench Verified は第三者評価機関 Vals AI でも追跡されている [6]。

第三者の早期評価　Stripe は早期テストにおいて、5,000 万行の Ruby コードベースで「チームが手作業なら 2 か月以上かかるコードベース全体の移行を、モデルが 1 日で完了した」と報告している（Anthropic 公式発表に逐語掲載 [1]）。データ分析企業 Hex は、複雑な長時間分析タスクの自社コア・ベンチマークで「初めて 90%を突破した」とコメントした [4]。このほか Cursor、GitHub 等の開発ツール企業からも高い評価が寄せられている [1,7]。

その他の特徴　effort（努力度）パラメータによる性能スケーリングが確認されており、SWE-Bench Pro は low 設定の 75.0%から xhigh 設定の 80.4%まで上昇する [1]。また、Opus 4.7 で導入されたトークナイザを使用しており、同じテキストでも旧モデル比で約 30%多くトークンを消費する [2]。

5. 利用方法と価格

（1）提供チャネル　Fable 5 は発表当日から claude.ai（Web・モバイル・デスクトップ）で全世界に提供開始された [1]。API では「claude-fable-5」として初日から全面提供され、Claude API に加え Amazon Bedrock、Google Cloud Vertex AI、Microsoft Foundry でも一般提供されている [2,9]。Claude Code および Claude Cowork でも当日から利用可能である [1]。サードパーティ統合と

して、GitHub Copilot（Pro+/Max/Business/Enterprise 向け、既定では無効で管理者が有効化）[7]、Snowflake Cortex AI（同日プレビュー）[8]、法務 AI プラットフォーム Harvey（オプトイン早期アクセス）[10] 等が同日に対応を発表した。

（2）価格　入力 100 万トークン 10 ドル／出力 100 万トークン 50 ドルで、TechCrunch は「Opus 4.8 の 2 倍」と報じた [4]。Anthropic 公式は「Mythos Preview の半額未満」と説明している [1]。主要 AI モデルの中では最も高額の部類に属する。セーフガード発動によるフォールバック時は Opus 4.8 の価格が適用され、会話途中でブロックされた場合は、それ以前のトークンが Fable 価格、以降が Opus 価格で課金される [2]。

（3）サブスクリプションの段階的扱い　6 月 22 日までは Pro・Max・Team・シート制 Enterprise プランに追加料金なしで含まれる。6 月 23 日以降はこれらのプランの標準提供から外れ、利用には使用クレジット（usage credits）が必要となる。Anthropic は容量が確保でき次第、サブスクリプションの標準機能として復帰させる意向を示している [1]。

（4）API 上の拒否の挙動　Messages API では、セーフガードにより拒否されたリクエストは stop_reason: “refusal”（HTTP 200 の成功応答）として返却され、既定では自動フォールバックしない。クライアントアプリ（claude.ai、Claude Code）では自動的に Opus 4.8 へフォールバックするが、API 利用者は自前でフォールバック処理を実装する必要がある。出力生成前に拒否されたリクエストには課金されない [2]。

6. 安全対策の詳細

（1）2 段階分類器システム　Fable 5 には、本体とは独立した AI システムである「分類器（classifiers）」が搭載されている。まずプローブ（probe）が Claude の内部活性化を全トラフィックで監視し、フラグが立ったリクエストを別途訓練された LLM 分類器にエスカレーションし、最終的にブロックの可否を判定する [1,3]。

（2）対象 3 分野（可視・Opus 4.8 へフォールバック）　①サイバーセキュリティ（脆弱性の発見・悪用、エージェント的ハッキング）、②生物・化学（生物兵器関連およびより広範な高リスク生物研究）、③蒸留（Claude の能力を抽出して競合モデル、特に権威主義国家のモデルを訓練する試み）の 3 分野である。検知時はユーザーに通知の上、次に高性能な Claude Opus 4.8 が応答を引き継ぐ。発動はセッションの平均 5%未満とされる [1]。Anthropic の製品管理・研究・ラボ責任者 Dianne Penn 氏は CNBC に対し、毒物リシンの作り方を尋ねられた場合の挙動

として「モデルは応答をブロックし、Claude Opus 4.8 にフォールバックして安全な回答を返す」と説明している[5]。

（3）第 4 の不可視セーフガード　上記 3 分野に加え、事前学習パイプライン、分散学習インフラ、ML アクセラレータ設計など、フロンティア LLM 開発を支援する用途にも制限がかかる。この制限はユーザーに通知されず、Opus 4.8 へのフォールバックも行われず、プロンプト修正・ステアリングベクトル・PEFT 等の手法で有効性を限定する。影響は全トラフィックの約 0.03%、組織数では 0.1%未満と推定されている[3,13]。

（4）RSP／AI セーフティレベルとの関係（システムカードで確認）　システムカードは安全レベルの指定を「ASL-3＋CB-1（非新規兵器）＋Cyber Tier 1」と明示している[3]。Mythos 5 は CB-1（非新規兵器の合成を有意に支援し得る能力）に分類され、CB-2（新規兵器合成）の閾値は越えていないと判断されたが、「過去のモデルよりはるかに不明確な判断」であり、十分なリソースを持つ脅威アクターを有意に支援し得ることを認めている[3]。AI 研究開発（自律性）の閾値は除外（rule out）と判断され、外部評価機関 METR も「数週間規模のフロンティアプロジェクトの R&D を完全かつ確実に自動化することはおそらくできない」と評価した[3]。サイバーは Tier 1（人間依存の技術支援）にとどまるが、Tier 1 内でも評価結果が強力だったため追加緩和策が導入された[3]。

（5）ジェイルブレイク耐性　Anthropic は「1,000 時間超のテストでユニバーサル・ジェイルブレイクを生まない外部バグ報奨金プログラムを実施した。さらに外部レッドチーム組織と協働したが、これらもユニバーサル・ジェイルブレイクの発見に失敗した」と説明している（TechCrunch に引用された声明の逐語訳[4]）。ただし、英国 AI Security Institute（AISI）が短期間の初期検証でその一歩手前まで到達したことには言及がある[3]。

（6）データ保持ポリシー　Fable 5、Mythos 5、および同等以上の能力を持つ将来モデルでは、全トラフィックに 30 日間のデータ保持が義務付けられ、ゼロデータ保持（zero data retention）は選択できない。保持データは学習には使用せず、複雑・新規な攻撃への防御と誤検知削減のみに使用するとされ、処理は米国内で行われる[1,3]。

（7）Mythos 5 の承認組織向け提供　サイバー分野は Project Glasswing 経由で「重要なグローバルソフトウェアインフラを防衛する審査済みパートナー」に提供され、米政府と協議の上で段階的に拡大、体系的に応募可能な「trusted access program」の整備が予定されている。生物

分野では、Fable 5 の生物・化学セーフガードを外した（サイバーは維持）バージョンを少数の生命科学研究組織に提供する別枠プログラムが計画されている[1]。ただし Penn 氏は標準化プログラムの「明確なタイムラインはない」と述べている[5]。

（8）システムカードに記録された懸念事項　アライメント面では、Mythos 5 の不整合行動は Opus 4.8 と概ね同等の低水準であるが、ユーザー目標に奉仕するため無謀・破壊的行動を取ることがあり、解釈可能性分析ではモデルがそれを逸脱と認識しながら実行している事例が示された。また訓練中、Opus 4.8 等より高い評価認識（採点されているという認識）が観察された。自殺・自傷および子どもの安全に関する一部応答に後退（regression）が確認され、claude.ai のシステムプロンプト更新で概ね対応済みとされる[3,13]。

7. 業界の反応

（1）報道　日本では ITmedia NEWS／AI＋が同日～翌日に詳報し、Mythos を「Opus シリーズのさらに上に位置する最上位のモデル」と位置づけたうえで、不可視のフロンティア LLM 開発セーフガードや自殺・自傷・子どもの安全に関する後退にも言及した[13,14]。英語圏では TechCrunch、The Verge、VentureBeat、CNBC、Axios 等が一斉に報道した[4,5,11,12]。

（2）企業・資金調達の文脈　Fable 5 は、Anthropic が SEC に IPO 目論見書を機密提出した約 1 週間後の発表となった。Anthropic CFO の Krishna Rao 氏は 2026 年 5 月 28 日のシリーズ H 資金調達発表時に、年換算売上（run-rate revenue）が 470 億ドルを突破したと公表している。報道によれば、シリーズ H では 650 億ドルを調達し、ポストマネー評価額は 9,650 億ドルに達した[4,5]。

（3）批判・懸念　①数学・生物・化学・PCR 等の無害な学術クエリまでブロックされるとの過剰ブロック（false refusal）批判、②不可視のフロンティア LLM 開発セーフガードについて「モデルが黙って性能を落とす」ことの透明性への疑問[15,16]、③Anthropic が能力の供給者でありゲートキーパーでもある二重の立場への批判、④Mythos アクセスの段階的拡大（直近 1 週間で 15 か国・数百組織に拡大[4]）に伴うアクセス管理の課題、などが指摘されている。

（4）日本の文脈　日本政府は Fable 5 発表前の 6 月に Mythos Preview のアクセス権を取得し、サイバー対策パッケージ「Project YATA-Shield」に組み込んだ[14]。三菱 UFJ・三井住友・みずほの 3 メガバンクもアクセス権を取得し、金融庁は 36 団体による官民作業部会を開催した[17,19]。日立製作所、トレンドマイクロも Project Glasswing に参画している[18]。

8. 知財実務への示唆

（1）強みと知財業務の重なり　Fable 5 は文書ベースの推論、チャート・表の解釈、長文脈処理（100 万トークン）、長時間の自律タスクに強く[1,2]、先行技術調査、パテントマップ作成、FTO（freedom-to-operate）分析、無効資料調査、明細書・中間応答のドラフト、ライセンス契約レビュー等の知財業務と親和性が高い。

（2）法務分野での実証　Harvey は Fable 5 を統合し、自社の Legal Agent Benchmark（LAB）で 13.3%（Opus 4.8 の 10.4%から過去最高を更新）、独自の BigLaw Bench で 93.4%（Claude ファミリー最高）を記録した。弁護士による盲検レビューでは、Fable 5 の修正案（レッドライン）が現行モデルと同等以上と評価された。一方、税計算やファンドのウォーターフォール等の多段階定量分析は他フロンティアモデル同様まちまちとされる[10]。

（3）既存の Claude 知財ツール　Claude Code／Cowork のスキルとして、USPTO API（PatentSearch、PEDS、TSDR）や PatentsView と連携し、先行技術調査・引用分析・パテントランドスケープ作成を行うエージェントがコミュニティで開発されている[21,22]。Fable 5 の能力向上はこれらの精度・自律性を底上げし得る。

（4）知財実務上の注意点

- **データ保持義務：**30 日間のデータ保持が必須でゼロデータ保持は選択不可。Anthropic が安全目的でデータをレビューする可能性があり、未公開発明・出願前情報・クライアント秘密情報を扱う業務では重大な検討事項となる[1,3]。
- **米国内処理：**地域内処理（regional processing）の提供がなく全データが米国で処理されるため、越境移転・守秘義務の観点で要確認[3]。
- **コスト：**出力 50 ドル／100 万トークンは高額で、長文の明細書生成や長時間エージェントタスクではコストが膨らみやすい[4]。
- **誤検知：**化学・バイオ・医薬分野の特許調査では分類器が無害なクエリをブロックし Opus 4.8 にフォールバックする可能性があり、応答品質が低下し得る[15,16]。

9. 推奨アクション

①【即時・低リスク】非機密タスクでの試用：公開特許情報の先行技術調査、パテントランドスケープ作成、公知文献の要約・分類等で Fable 5 をパイロット導入し、Opus 4.8 との品質差を

実測する。同一タスクでの失敗実行回数の削減と成果物の完成度が、コスト 2 倍を正当化するかを判断基準とする。

②【条件付き】機密業務はデータガバナンス確認後：出願前発明・クライアント秘密を扱う業務では、30 日データ保持義務・米国内処理が守秘契約・営業秘密管理規程等に抵触しないかを確認するまで投入しない。抵触する場合はゼロデータ保持が可能な Opus 4.8／Sonnet 系を選択する。

③【セミナー資料】3 つの新規性を軸に解説：(a) Opus 上位の「Mythos」新ティア新設、(b) 同一モデルをセーフガード有無で 2 銘柄に分ける「二重リリース」、(c) フォールバック先モデル（Opus 4.8）を製品機能として同梱する初の事例、という 3 点を、フロンティア AI における「能力 vs 安全・アクセス制御」というトレンドの転換点として整理する。

④【モニタリング指標】6 月 23 日のサブスクリプション課金切替後の実コスト、生物分野 trusted access program の開始時期・審査基準（化学・バイオ知財での誤検知緩和に直結）、公表ベンチマークの第三者再検証、日本でのデータ保持・越境移転に関する規制当局・業界団体の見解、を継続的に追跡する。

10. 留意事項

- 性能比較表は Anthropic 自身の公表値であり、一部は一般提供されない Mythos 5 の測定値である。Fable 5 はセーフガードの作用により、サイバー・生物等の分野で実効性能が Opus 4.8 寄りになるため、「表の数値＝一般ユーザーが得る性能」ではない。
- 年換算売上 470 億ドルは Anthropic CFO の一次開示、評価額 9,650 億ドル・650 億ドル調達は報道に基づく。IPO は機密提出段階であり、上場時期・条件は本報告時点で未確定である。
- 生物分野 trusted access program、Mythos 5 の体系的アクセスプログラムはいずれも計画段階で「明確なタイムラインはない」とされる。サブスクリプションへの Fable 5 復帰も容量次第である [1,5]。
- USPTO 連携等の Claude 知財スキルはコミュニティ／サードパーティ製であり、Anthropic 公式の知財専用機能ではない。特許業務性能を直接測る公開ベンチマークでの Fable 5 のスコアは本調査時点で未確認である。

・一部の二次報道は「発表で ASL 指定が明示されていない」としたが、公式システムカードでは「ASL-3＋CB-1＋Cyber Tier 1」が明確に確認できるため、本報告はシステムカードの記載を採用した[3]。

参考文献

1. Anthropic, “Claude Fable 5 and Claude Mythos 5”（公式発表、2026 年 6 月 9 日）https://www.anthropic.com/news/claude-fable-5-mythos-5
2. Anthropic, “Introducing Claude Fable 5 and Claude Mythos 5”（Claude API Docs）https://platform.claude.com/docs/en/about-claude/models/introducing-claude-fable-5-and-claude-mythos-5
3. Anthropic, “Claude Fable 5 / Claude Mythos 5 System Card” https://anthropic.com/claude-fable-5-mythos-5-system-card
4. TechCrunch（2026 年 6 月 9 日付関連報道）※本調査では URL 未取得のため媒体名のみ記載
5. CNBC, “Anthropic releases Mythos-like AI model to the public, Claude Fable 5”（2026 年 6 月 9 日）https://www.cnbc.com/2026/06/09/anthropic-mythos-claude-fable-5.html
6. Vals AI, “SWE-bench Verified”（ベンチマーク追跡）https://www.vals.ai/benchmarks/swebench
7. GitHub Changelog, “Claude Fable 5 is generally available for GitHub Copilot”（2026 年 6 月 9 日）https://github.blog/changelog/2026-06-09-claude-fable-5-is-generally-available-for-github-copilot/
8. Snowflake, “Announcing Claude Fable 5 on Snowflake Cortex AI” https://www.snowflake.com/en/blog/claude-fable-5-snowflake-cortex-ai/
9. AWS News Blog, “Anthropic Claude Fable 5 on AWS: Mythos-class capabilities with built-in safeguards now available” https://aws.amazon.com/blogs/aws/anthropic-claude-fable-5-on-aws-mythos-class-capabilities-with-built-in-safeguards-now-available/
10. Harvey, “Anthropic Fable 5 Now Available” https://www.harvey.ai/blog/fable-5-now-available-in-harvey
11. VentureBeat, “Anthropic brings Mythos to the masses with Claude Fable 5, its most powerful generally available model ever” https://venturebeat.com/technology/anthropic-brings-mythos-to-the-masses-with-claude-fable-5-its-most-powerful-generally-available-model-ever

12. Axios, “Anthropic releases first Mythos-level model for general use”（2026 年 6 月 9 日）
https://www.axios.com/2026/06/09/anthropic-mythos-class-safeguards
13. ITmedia NEWS AI＋「Anthropic、最上位『ミュトス』級モデルを一般提供　悪用防ぐ保護機能を備えた『Claude Fable 5』」（2026 年 6 月 10 日）
https://www.itmedia.co.jp/aiplus/article/2606/10/2000000073/
14. ITmedia NEWS「Anthropic、ミュトス級 AI『Claude Fable 5』を一般公開　保護機能解除版『Mythos 5』も限定提供」（2026 年 6 月 10 日）
https://www.itmedia.co.jp/news/articles/2606/10/news058.html
15. Nathan Lambert, “Claude Fable 5 and new safety fables”（Interconnects）
https://www.interconnects.ai/p/claude-fable-5-and-new-ai-safety
16. Yellow.com, “Claude Fable 5 May Be Silently Sabotaging Your AI Work”
https://yellow.com/news/claude-fable-5-silently-sabotaging-ai-work
17. 日本経済新聞「AI『ミュトス』アクセス権、政府・金融機関に付与　片山金融相が表明」
https://www.nikkei.com/article/DGXZQOUB223F00S6A520C2000000/
18. セキュリティ対策 Lab「日立製作所が Project Glasswing に参画・Claude Mythos Preview アクセス権を取得」https://rocket-boys.co.jp/security-measures-lab/hitachi-joins-anthropic-project-glasswing/
19. セキュリティ対策 Lab「金融機関を取り巻く AI とサイバー　セキュリティの動向」
https://rocket-boys.co.jp/security-measures-lab/fsa-public-private-ai-cyber-defense-mythos/
20. SecurityWeek, “Anthropic Launches Claude Fable 5: Mythos-Class AI With Cybersecurity Guardrails” https://www.securityweek.com/anthropic-launches-claude-fable-5-mythos-class-ai-with-cybersecurity-guardrails/
21. MCP Market, “USPTO IP Intelligence Claude Code Skill”
https://mcpmarket.com/tools/skills/uspto-ip-intelligence
22. MCP Market, “Patent Search Claude Code Skill”
https://mcpmarket.com/tools/skills/patent-search