

Claude Fable 5.1／GPT-6 Astra／Gemini 3.8 Flash／Muse Spark 1.3

— 一次ソース検証・独立評価・日本企業実務への含意 —

作成日：2026 年 9 月 6 日

Claude Fable 5.1

1. 要旨

- 2026 年 9 月 1 日から 3 日にかけて Anthropic、Google、Meta、OpenAI の 4 社が相次いで新モデルを発表したことは、各社の公式発表により確認できる^{1,2,3,4}。ただし本調査の前提には 2 点の補正が必要である。第一に、Anthropic のモデルの正式名称は「Claude Fable 5.1」であり、「5.1」は最上位系列「Fable」の版番号である^{1,5}。第二に、「GPT-6 Astra」は 2026 年 8 月 1 日の未解決数学問題解決の発表時に名前が挙がった内部モデルと同一系列であり、正式リリースは 9 月 3 日である^{4,6}。
- 独立系評価機関の集計では、総合知能で Claude Fable 5.1 が首位である。Artificial Analysis Intelligence Index において、Fable 5.1 (max effort) は同機関計測史上最高の 66 点を記録し、Index v4.2 への改訂後も首位を維持した^{7,8}。GPT-6 Astra は自社ベンチマークでは競合をほぼ全項目で上回ると主張するが、第三者集計では 2 位相当であり、数学・コンピュータ操作・サイバーなど特定領域で突出する^{4,7,9}。Gemini 3.8 Flash は軽量系列であり、他 3 モデルとはティアが異なる²。Muse Spark 1.3 はコスト効率の点で「フロンティア到達」と評価された¹⁰。
- **【分析】** 日本企業の知財・R&D 部門の実務では、深い推論・長時間エージェント・科学研究支援は Fable 5.1、コスト効率重視の大量処理は Gemini 3.8 Flash または Muse Spark 1.3、コンピュータ操作・数学的検証・サイバー防御は GPT-6 Astra が有力候補となる。ただし、デュアルユース規制 (Mythos 5.1 および Gemini 3.8 Flash Cyber は日本企業から原則アクセス不可) と、各ベンチマークの測定条件差に留意が必要である。

凡例：本書では、各社公式発表・公式ドキュメント等の一次ソースで確認した事項を通常の記述とし、報道・二次情報に基づく事項には「報道」「とされる」を付し、筆者の解釈・推論には**【分析】**を付して区別する。参考文献の URL は実際にアクセス確認できたものを掲載し、確

認できなかったものは「URL 未確認」と注記した。

2. 前提の検証（一次ソース確認）

2.1 Claude Fable 5.1

Anthropic は公式ブログ「Introducing Claude Fable 5.1 and Claude Mythos 5.1」で両モデルを発表した¹。発表日は 9to5Mac、VentureBeat、MacRumors 等が 2026 年 9 月 1 日と報じている^{11,12,13}。Fable は同社の最上位（フラッグシップ）系列であり、下位に Opus・Sonnet・Haiku が位置する⁵。Claude Mythos 5.1 は Fable 5.1 と同一の基盤モデルであり、両者の差はセーフガードの水準にある（第 7 章参照）¹。したがって、ユーザー表記の「Claude 5.1」は「Claude Fable 5.1」を指すと解するのが適切である。

2.2 GPT-6 Astra

OpenAI は公式ページ「GPT-6 Astra: A new generation of intelligence」で発表した⁴。リリース日は Axios、CNBC 等が 2026 年 9 月 3 日と報じている（一部媒体は 9 月 4 日と表記）^{6,14,15}。2026 年 8 月 1 日に「10 の未解決数学問題を解いた」と発表されたモデルは、当時未公開の内部版「Astra」であり、本モデルと同一系列である^{4,6}。よって 8 月時点では未リリースであり、正式公開は 9 月 3 日と整理できる。名称「GPT-6 Astra」は公式表記と一致する。

2.3 Gemini 3.8 Flash

Google は公式ブログ（2026 年 9 月 2 日付、Tulsee Doshi 氏・Raluca Ada Popa 氏執筆）で「Gemini 3.8 Flash」および「Gemini 3.8 Flash Cyber」を発表した²。最も重要な留意点は、本モデルが軽量「Flash」系列であり、他 3 モデル（いずれもフラッグシップ級）とはティアが異なることである^{2,16}。信頼できる組織向けの限定提供枠「Fairwind Program」も同ブログで公式に確認できる²。

2.4 Muse Spark 1.3

Meta（Meta Superintelligence Labs）は、研究ブログ「Introducing Muse Spark 1.3」および AI at Meta 公式 X アカウントで 2026 年 9 月 2 日に発表した^{3,17,18}。

2.5 「Fable 5 に匹敵」報道の検証

報道の根拠は 2 系統ある。第一に、Meta 最高 AI 責任者 Alexandr Wang 氏は、Muse Spark 1.3 がコーディングで GPT-5.6 Sol を超え、Claude Fable 5.1 と「同等 (on par)」であり「これまでで最大の性能飛躍」であると発言したと報じられている (報道)¹⁹。第二に、Artificial Analysis は独立記事「Muse Spark 1.3: Meta reaches the frontier」(9 月 2 日)において、限定プレビュー版の Muse Spark 1.3 (max) が Intelligence Index 62 点で Claude Fable 5.1 および Opus 5 に次ぐ位置にあり、一般提供版 (xhigh) は 61 点で GPT-5.6 Sol (max) ・ Grok 4.6 (high) と同点であると評価した¹⁰。旧 Claude Fable 5 (max) が 62 点であるため⁷、「(旧) Fable 5 に匹敵」という表現は独立評価上おおむね妥当である。

【分析】他方、最新の Fable 5.1 (max、66 点) には及ばない。Wang 氏の「Fable 5.1 と同等」発言はコーディング特化ベンチマーク (自社ハーネスによる測定で第三者再現なし) に基づく主張であり、総合知能では独立評価上 Fable 5.1 が上位に位置する。

3. 基本情報の比較

項目	Claude Fable 5.1	GPT-6 Astra	Gemini 3.8 Flash	Muse Spark 1.3
発表主体	Anthropic	OpenAI	Google (DeepMind)	Meta (Superintelligence Labs)
発表日	2026 年 9 月 1 日	2026 年 9 月 3 日	2026 年 9 月 2 日	2026 年 9 月 2 日
系列内の位置	フラッグシップ (Fable = 最上位)	フラッグシップ (GPT-6 世代最上位)	軽量系列 (Flash)	フラッグシップ級 (Muse Spark 系最上位)
提供形態	API／消費者向け (Pro・Max・ Team・Enterprise) ／主要クラウド 3 社	API／ChatGPT 全プ ラン／Azure・ Bedrock (段階展 開)	API／AI Studio／ Antigravity／消費者 向け (Pro・Ultra)	Muse Code／Meta Model API
コンテキスト長	1M トークン級と報 道 (公式本文では明 示なし)	1M トークン	1M トークン	1M トークン
モデル ID	claude-fable-5-1	gpt-6-astra	gemini-3.8-flash	(Muse Code／API 経由)

項目	Claude Fable 5.1	GPT-6 Astra	Gemini 3.8 Flash	Muse Spark 1.3
重み公開	クローズド	クローズド	クローズド	クローズド（オープンウェイト版は「近日」とされるが未履行）

出典：各社公式発表^{1,4,2,3}、報道・二次情報^{12,15,20,21}。

4. 料金・提供条件

モデル	入力（/1M トークン）	出力（/1M トークン）	備考
Claude Fable 5.1	\$10	\$50	キャッシュ読取\$0.25（75%減）。典型的ワークロードで実効約 25%減、高度エージェント作業で最大約 45%減。米限定推論は 1.1 倍 ^{1,12}
GPT-6 Astra	\$10	\$50	GPT-5.6 Sol の 2.5 倍。キャッシュ読取 90%割引・書込 25%割増。Fast モードは 2 倍速・2 倍価格 ^{4,22}
Gemini 3.8 Flash	\$0.75	\$3.75	導入価格（2026 年 12 月 31 日まで）。2027 年 1 月 1 日以降は\$1.50/\$7.50 に倍増予定 ²
Muse Spark 1.3	\$1.25	\$4.25	Muse Spark 1.2 と同額。contributor tier は\$0.10/\$0.20。88%キャッシュ割引 ^{18,20}

出典：各社公式料金表および報道（表中に個別記載）。

【分析】 Gemini 3.8 Flash は Fable 5.1・Astra と比べ、入力・出力ともに約 13 分の 1 の価格である。軽量系列である点を割り引いても、コスト効率は際立つ。Muse Spark 1.3 の contributor tier（学習データ提供と引き換えの低価格枠）は最安級に位置する。

5. 性能評価

5.1 第三者評価：Artificial Analysis Intelligence Index

- Artificial Analysis（AA）によれば、Claude Fable 5.1（max effort）は 66 点で同機関計測史上最高であり、Claude Opus 5（max、63 点）、Claude Fable 5（max、62 点）、GPT-5.6 Sol（max、61 点）、Grok 4.6（high、61 点）を上回る（フォールバック出力は出力トークンの約 4%）。リーダーボード表示値（フォールバック込み）は 65.7 点⁷。Muse Spark 1.3

は 62 点 (max) / 61 点 (xhigh) ¹⁰、GPT-6 Astra は 61.2 点で GPT-5.6 Sol とほぼ同点、Gemini 3.8 Flash は 59 点 (high) とされる ^{7,23}。

- 9 月 4 日公表の Index v4.2 (より難度の高い問題と非公開テスト 40%を導入) でも、首位は Claude Fable 5.1、2 位 GPT-6 Astra、3 位 Meta という順位である ⁸。The Decoder は、GPT-6 Astra のスコアが実際の進歩を捉えきれていないとの批判を受けて AA が改訂に至ったと報じている (報道)。Astra は前世代 Sol から 4 点上昇した ⁸。
- **【分析】** OpenAI 自社ベンチマークでは Astra が Fable 5.1 にほぼ全項目で勝つが、独立集計では Fable 5.1 が総合首位となる。両者は「測定対象が異なる」ため両立し得る ^{24,25,26}。

5.2 第三者評価：Epoch AI

- Epoch AI 公式 X によれば、GPT-6 Astra は Epoch Capabilities Index (ECI) で 169 点の新記録を樹立した (従来最高は 163 点=Claude Fable 5.1)。ただし Epoch 自身が、この上昇は推論時代の ECI トレンドの不確実性範囲内であると注記し、長時間コーディングベンチマーク MirrorCode では Astra は Opus 4.7 と Fable 5 の間に位置するとした ⁹。The Decoder の「267 モデル中 1 位」は二次的表現である ⁸。Epoch の Greg Burnham 氏は Astra を一つの時代の終わりと次の時代の始まりを画するものと評し、OpenAI は公式ページでこれを引用している ⁴。
- Epoch は OpenAI から事前アクセスを得てテストしたことを開示している ⁹。

5.3 LMArena (人間選好 Elo)

マルチモーダル (vision) 部門では claude-fable-5 が 1330 Elo で首位である (BenchLM 集計) ²⁷。9 月上旬に発表された新モデル群の Elo は確定途上にある。

5.4 主要ベンチマーク一覧 (自己申告値を含む)

下表のうち OpenAI 公式表の数値は自社研究環境・最大 effort での測定であり、同表中の競合スコアも OpenAI が再測定・引用したものである。Anthropic 公式表の数値は Anthropic の測定である。測定条件が異なるため、厳密な横並び比較はできない点に注意を要する。

ベンチマーク	出典	Fable 5.1	GPT-6 Astra	Opus 5	Gemini 3.8 Flash
FrontierMath Tier 4 (v2)	OpenAI 表	87.8%	97.6%	73.2%	—
Humanity's Last Exam (ツール使用)	両社表	65.0%	57.2%	63.6%	—
GPQA Diamond	OpenAI 表	93.7%	96.0%	93.7%	95.3%
Terminal-Bench 4.0	両社表	55.8%	57.9%	52.6%	19.1%
DeepSWE v1.1	OpenAI 表	67.4%	74.1%	73.7%	73.8%
Terminal-Bench-Science 0.1	両社表	52.6%	64.6%	30.0%	—
ARC-AGI-3	OpenAI 表	—	99.9%※	30.2%	—
ExploitBench	OpenAI 表	—	100%	70%	—
AutomationBench	両社表	31.4%	41.4%	26.9%	—

出典：OpenAI 公式⁴、Anthropic 公式¹、二次的整理^{28,22}。※ARC-AGI-3 の 99.9%は OpenAI の Provider Adapter ハーネスによる値であり、ARC Prize 財団の標準ハーネスでは 62.7%との報道がある²⁶。【分析】「飽和」は特定のモデル＋ハーネス構成に対する表現である点に注意。

5.5 FrontierMath に関する検証

Epoch AI は FrontierMath の運営者である。OpenAI 公式は Astra が FrontierMath Tier 4 を 98% で飽和させたと主張し⁴、Epoch も最後の未解決 1 問が解かれたとして Tier 4 の「完全飽和」を宣言した²⁹。ただし、FrontierMath は OpenAI の資金提供で開発され、OpenAI が一部に排他アクセスを有するという利益相反が、Epoch 自身および複数の媒体により明記されている^{29,30}。さらに Epoch は同時により難度の高い「FrontierMath Erdős」（68 の Erdős 未解決問題）を公開し、Astra は標準予算では 68 問中 2 問（約 3%）のみを解決した（唯一 0%でないモデル）³¹。

【分析】Epoch の一連の公表は、「Tier 4 飽和」が「数学研究が解けるようになった」ことを意味しないことを示唆している。

6. 能力の特徴（一次ソースベース）

6.1 Claude Fable 5.1

科学研究支援が突出する。同一基盤の Mythos 5.1 は、タンパク質バインダー設計において

Adaptyv Bio コンペティション最良比で 10 倍の結合親和性を達成し、12 ターゲットで約 50% のヒット率（通常 10～15%）を示した。また、金星の 3 分の 1 に相当する高解像度標高マップをニューラルネットワークで作成し Creative Commons で公開し、計算生物学では 7 つのオープンソースモデルを最大 2.5 倍高速化した¹。長時間の自律作業では、Ramp 社で 38 時間の無人実行、Millennium 社で数年間未解明であったクラッシュ原因の特定が顧客証言として示されている。CursorBench 3.2.0 では 73.4%（Cursor 証言）¹。

6.2 GPT-6 Astra

コンピュータ操作が世界最高水準にある。OSWorld 2.0 で 72.6% を記録し、GPT-5.6 Sol より約 47% 短時間で完了、ScreenSpot-Pro で 92.7%。Codex ハーネス更新により 1.9 倍高速化した⁴。サイバー能力は Preparedness Framework 上で初めて「Critical」に到達し、ExploitBench 100%、テスト中に 2 件の未知のゼロデイを発見・使用した（開示済み）⁴。数学では素数ギャップに関する新結果 2 件（bounded gap を 246 から 186 へ改善等）を公表した。長文処理では MRCR v2 の 8-needle・512K～1M トークンで 96.3%⁴。Greg Brockman 社長は AGI の到来かもしれないと発言しているが、これは経営陣の主張であり、判断はユーザーに委ねる旨も併記されている⁶。

6.3 Gemini 3.8 Flash／3.8 Flash Cyber

軽量系列ながら、DeepSWE v1.1 で多くの大型フロンティアモデルを低コストで上回ると Google は主張する。HLE-Verified 54.9%、Vals Finance Agent V2 および Harvey 社の Legal Agent Benchmark で 3.7 Flash および他のフロンティアモデルを上回るとする²。Gemini 3.8 Flash Cyber は、CyberGym における自律的脆弱性発見でフロンティア級の性能を示し、CWE-Bench（Collinear 運営）では pass@1 で 47.2%（フロンティアモデル 47.8% に匹敵）を低コストで達成した。Chrome Security チームでは最良の大型商用モデルの 2.6 倍の正しいパッチを生成し、Wiz 社では検出率が 7.5～9.7 ポイント向上しコストは 2.3～5.2 分の 1 であった。Cyber 版は Fairwind Program（政府機関・重要インフラ運用者・ソフトウェア保守者向け）経由でのみ提供される²。

6.4 Muse Spark 1.3

エージェント／コーディングに特化する。Muse Spark 1.2 比でツール呼出が約 20% 減、トークン消費が約 25% 減（Meta 社内比較）。長時間・複数ワークフローの並行処理と、明確化のため

の質問・停滞の報告・重要操作前の確認といった協調行動を特徴とする。max reasoning モードは追加の安全テスト完了まで限定提供。入力テキスト・画像・動画のマルチモーダル、出力はテキストである^{3,18}。

7. 安全性・ガバナンス

項目	Anthropic	OpenAI	Google	Meta
安全枠組み	Responsible Scaling Policy／Frontier Compliance Framework	Preparedness Framework	Frontier Safety Framework	Advanced AI Scaling Framework v2
デュアルユース区分	Fable 5.1（一般）／Mythos 5.1（信頼アクセス限定）。同一基盤・セーフガード差	一般版は高度サイバーを拒否。Daybreak 経由で緩和	3.8 Flash（一般）／3.8 Flash Cyber（Fairwind 限定）	max reasoning は限定提供。CBRN 拒否
サイバー到達水準	全モデル中最強だが Frontier Compliance Framework 上は下位カテゴリ	初の「Critical」到達	フロンティア級（防御特化、攻撃能力は抑制）	詳細非公開
生物	Mythos 5.1 は Mythos 5 と同水準のセーフガードを維持（次のリスク層には未到達）	—	CBRN セーフガード	CBRN 拒否
電子透かし	EU AI Act 対応のテキスト透かし＋検出 API（プライベートプレビュー）	—	—	—

出典：各社公式^{1,4,2}、Meta については研究ブログおよび報道^{17,20}。

Fable／Mythos 区分の要点は次のとおりである。両者は完全に同一のモデルであり、Fable 5.1 は本番セーフガード付きで一般提供、Mythos 5.1 はサイバー・生命科学の検証済み組織向け（Cyber Verification Program／Life Sciences Verification Program。米国政府と連携し、現状は米国組織限定）である¹。Fable 5.1 自体も Covered Model 指定を受け、データ保持・安全審査・アクセス方針に関する追加措置の対象となる¹。Fable 5.1 は脆弱性の「発見」は許容するが「エ

クスプロイト開発」は許容しない。Claude Code におけるサイバー関連の誤検知は約 60%減少し、生物分野のセーフガードは初等生物・医学質問での誤発火が 85%減少したとされる¹。

【分析】「Mythos はデュアルユース能力の追加安全措施がない」という理解はやや不正確である。Mythos は追加措置が「ない」のではなく、Fable に課される一部のサイバー・生物制限が緩和・解除される代わりに、組織審査という別の統制がかかる構造である。

8. 市場・専門家の評価と各社の戦略的位置づけ

- **Anthropic**：エンタープライズと安全性を訴求する。Enterprise Frontier Safeguards (EFS、顧客管理クラウドでのゼロデータ保持) を 100 社超と共同開発し、AI for Science・Model Hardware Standard など科学研究への投資を打ち出す¹。独立評価で総合首位を維持し、企業向けで優位に立つ^{7,8}。
- **OpenAI**：AGI 言説を前面に掲げる。史上最大規模の学習 (Stargate Texas で 10 万 GPU 超) を経て、米国政府の正式レビュー後に公開した^{4,6}。一方、ベンチマークの「飽和」表現とハーネス条件差、段階展開 (Pro 優先を巡る混乱で Altman 氏が謝罪) に対する批判がある (報道)²⁶。
- **Google**：低コスト・高効率戦略を採る。6 週間で 3 回目の Flash 系リリースである。Hassabis 氏の DeepMind CEO 退任・会長就任や Gemini 3.5 Pro の遅延により「フロントティア競争への関与に疑問」との報道が出る中、軽量系列で存在感を示した (報道)¹⁶。
- **Meta**：オープン／クローズ戦略が揺れている。2026 年 4 月に Muse Spark を初のクローズドモデルとして投入し Llama から転換した後³²、8 月に Muse Glimmer (30B、Apache 2.0) でオープンに回帰し³³、Muse Spark 1.2/1.3 のオープンウェイト化を「近日」と約束するも未履行である。エンタープライズからは「6 カ月で 2 回ライセンス方針を変えた予測可能性の問題」との批判がある³⁴。

9. 中国モデルとの相対的位置

- **DeepSeek V4-Pro** (8 月 13 日 GA、総パラメータ 1.6T、MIT ライセンス)：SWE-bench Verified 96.40%で 2 位 (Opus 5 に次ぐ)、コストは約 36 分の 1。ただしエージェンティ

ックコーディングでは 7 モデル中最下位 (LiveBench) であり、スコアは scaffolding 依存との指摘がある^{35,36}。

- **GLM-5.3** (Z.ai、8 月) : DeepSWE v1.1 で 66.9、SWE-Marathon v1.1 で 42.5 (ベンダー申告)。AA の Cost per Task パレートフロンティアを米 3 社と共有する^{36,23}。
- **Qwen3.8-Max** (Alibaba、8 月 3 日、2.4T sparse-MoE) : オールラウンダー志向。AA Intelligence Index 58 点³⁷。
- **Kimi K3** (Moonshot AI) : AA Index 上位 (一部集計で 3 位)、Frontend Code Arena で首位²³。

【分析】中国のオープンウェイトモデルは、コスト効率と特定ベンチマークで米フロンティアに肉薄するが、総合知能・エージェント能力では Claude Fable 5.1/GPT-6 Astra が依然としてリードする。ただし、Gemini 3.8 Flash・Muse Spark のコスト効率帯では中国勢と直接競合の関係にある。

10. 日本企業（知財・R&D 部門）の実務への含意【分析】

本章は前章までの確認事実に基づく筆者の分析であり、選定は各社の利用規約・データ取扱条件・社内規程を踏まえた個別検討を要する。

10.1 用途別の選定指針

- 深い推論・長時間自律研究・科学 R&D 支援 (サイエンスエージェント化) → Claude Fable 5.1。タンパク質設計・計算生物学・データ解析での実証、独立評価での総合首位、長時間無人実行の安定性が根拠となる。知財では先行技術調査、複雑な明細書ドラフティング、R&D の仮説生成に適する。
- コンピュータ操作の自動化・数学的検証・サイバー防御 → GPT-6 Astra。OSWorld / ScreenSpot-Pro での最高水準、FrontierMath Tier 4 飽和、ExploitBench 100% が根拠。CAD 再構成 (BenchCAD 95.9%) や社内システム操作の自動化に有力である。

- 大量文書処理・コスト効率重視の定型業務→Gemini 3.8 Flash（他の約 13 分の 1 の価格）または Muse Spark 1.3 contributor tier。大量の公開特許・論文のスクリーニング、要約、一次分類に向く。
- 社内エージェント・コーディング支援のコスト最適化→Muse Spark 1.3。AA のタスク当たりコストは最良水準（\$0.55／task）である。

10.2 デュアルユース・規制の確認（選定前に必須）

- 生命科学・創薬・サイバーの高度用途では、Fable 5.1 にセーフガード上の制限がかかり得る。Claude Mythos 5.1（LSVP／CVP）は現状米国組織限定であり、日本企業は原則としてアクセスできない。Gemini 3.8 Flash Cyber も Fairwind Program（政府・重要インフラ中心）に限られ、一般企業の利用は困難である。高度サイバー・生物用途は、当面フル能力版に日本からアクセスしにくい前提で計画する必要がある。
- EU AI Act の対象となる業務では、Anthropic の電子透かし・検出 API が有用となる。

10.3 調達リスク管理

- Meta はライセンス方針を 6 カ月で 2 回変更している。複数年の展開では Meta の予測可能性リスクを織り込む必要があり、オープンウェイト待ちについては「近日」の約束が未履行である点に注意する。
- Gemini Flash は 2027 年 1 月 1 日に価格が倍増する予定であり、コスト前提の見直しが必要となる。

10.4 判断を見直す閾値

- 独立系評価（AA Index v5、Epoch ECI）で GPT-6 Astra が Fable 5.1 を明確に上回れば、推論・科学用途の第一候補を見直す。
- Muse Spark のオープンウェイト版が実際にリリースされ、かつライセンスが商用に適するものであれば、オンプレミス／データ主権要件のある知財業務で有力化する。
- 中国モデル（DeepSeek／GLM／Qwen）が独立エージェントベンチで米フロンティアと同等になれば、コスト最適化の選択肢が拡大する。

11. 総括：各モデルの強み・弱み・適する用途

モデル	強み	弱み	適する用途
Claude Fable 5.1	総合知能で独立評価首位。科学研究・長時間エージェント・企業向け安全性	高価格。生物・サイバー領域で強いセーフガード制約	R&D 支援、知財の高度分析、規制産業
GPT-6 Astra	コンピュータ操作・数学・サイバー・トークン効率	高価格。独立総合では 2 位。ハーネス依存の誇大表現への批判。段階展開	業務自動化、数学的検証、（許可される場合の）サイバー防御
Gemini 3.8 Flash	圧倒的なコスト効率。軽量・高速	軽量ティアであり最高難度の推論は対象外。価格倍増予定	大量処理、コスト重視業務
Muse Spark 1.3	タスク当たりコスト効率が最良。エージェント効率	クローズド（オープンウェイト未履行）。自己申告中心。予測可能性リスク	コスト最適化されたエージェント／コーディング

本表は第 3～9 章の確認事実に基づく筆者の整理【分析】である。

12. 今後数カ月で注視すべき点

- Artificial Analysis Index v5（8 カ月をかけて開発中とされる）および Epoch ECI の更新で、フロンティア順位がどう動くか。
- Muse Spark オープンウェイト版の実際のリリースとライセンス条件。
- GPT-6 Astra の Daybreak（サイバー制限緩和）の拡大、および Mythos 5.1／Fairwind の国際展開（日本企業のアクセス可否）。
- Gemini Flash の 2027 年価格改定の実施と、Google フラッグシップ（次期 Gemini Pro／Deep Think）の動向。
- 中国オープンウェイト勢のエージェント能力の追い上げと、それに対する各社のコスト戦略。
- AGI 言説の検証：OpenAI が掲げる「研究インターン級（2026 年 9 月目標）→自律 AI 研究者（2028 年初頭）」ロードマップの進捗。

13. 留意事項

- **測定条件の非対称性**：各社公式ベンチマークは自社に有利な条件（ハーネス、effort、フールバック）で測定されている。特に OpenAI の ARC-AGI-3 「99.9%」や Astra 公式表中の競合スコアは条件差が大きく、独立系（Artificial Analysis、Epoch AI）の数値を優先すべきである。
- **利益相反**：FrontierMath は OpenAI の資金提供を受け、OpenAI が一部に排他アクセスを有する。Epoch は独立運営だが、この点を自ら開示している。
- **自己申告と第三者評価の区別**：Muse Spark 1.3 のコーディングベンチマーク（DeepSWE 75.4 等）および Wang 氏の「Fable 5.1 と同等」発言は自社ハーネスによる測定で第三者再現がない。Meta 社内の「20%/25%削減」も社内比較である。
- **報道ベースの未確認情報**：発表日の一部（Astra：9月3日と9月4日の表記混在）、Fable 5.1 のコンテキスト長の公式明示値、Muse Spark オープンウェイトの時期は流動的または未確認である。
- **AGI 言説**：OpenAI の AGI 主張は経営陣の見解であり、独立評価は「大きいが不均一な進歩」と評価している。断定は避けるべきである。
- **法的不確定性**：デュアルユースモデルに関する各国の輸出管理・アクセス可否は変動しており、日本からの Mythos/Cyber 系モデルへのアクセス可否は個別確認を要する。

参考文献

番号は本文中の上付き番号に対応する（初出順）。URL は 2026 年 9 月 6 日時点でアクセスを確認したもの。確認できなかったものは「URL 未確認」と記した。

[1] Anthropic 「Introducing Claude Fable 5.1 and Claude Mythos 5.1」 2026 年 9 月 1 日。

<https://www.anthropic.com/claude-fable-and-mythos-5-1>

[2] Google (The Keyword) 「Introducing Gemini 3.8 Flash and 3.8 Flash Cyber」 2026 年 9 月 2 日。

<https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/>

[3] AI at Meta (X 公式アカウント) 「Muse Spark 1.3 リリース告知投稿」 2026 年 9 月 2 日。

<https://x.com/AIatMeta/status/2095234385129963666>

- [4] OpenAI 「GPT-6 Astra: A new generation of intelligence」 2026 年 9 月 3 日。
<https://openai.com/index/gpt-6-astra/>
- [5] Anthropic 「Claude Fable（製品ページ）」 2026 年 9 月閲覧。 <https://www.anthropic.com/claude/fable>
- [6] Axios 「OpenAI releases new model GPT-6 Astra, says it may represent AGI」 2026 年 9 月 3 日。
<https://www.axios.com/2026/09/03/openai-astra-gpt-6-agi-brockman>
- [7] Artificial Analysis 「Benchmarking GPT-6 Astra」 2026 年 9 月。
<https://artificialanalysis.ai/articles/benchmarking-gpt-6-astra>
- [8] The Decoder 「Artificial Analysis overhauls its Intelligence Index after GPT-6 Astra scoring drew skepticism」 2026 年 9 月。 <https://the-decoder.com/artificial-analysis-overhauls-its-intelligence-index-after-gpt-6-astra-scoring-drew-skepticism/>
- [9] Epoch AI (X: @EpochAIResearch) 「GPT-6 Astra の Epoch Capabilities Index（169 点）および MirrorCode 評価に関する投稿」 2026 年 9 月。（URL 未確認）
- [10] Artificial Analysis 「Muse Spark 1.3: Meta reaches the frontier」 2026 年 9 月 2 日。（URL 未確認）
- [11] 9to5Mac 「Anthropic upgrades Claude with new Fable 5.1 model, details here」 2026 年 9 月 1 日。
<https://9to5mac.com/2026/09/01/anthropic-upgrades-claude-with-new-fable-5-1-model-details-here/>
- [12] VentureBeat 「Anthropic's Claude Fable 5.1 and Mythos 5.1 arrive with a 75% cost reduction for Fable cache reads」 2026 年 9 月。 <https://venturebeat.com/technology/anthropics-claude-fable-5-1-and-mythos-5-1-arrive-with-a-75-cost-reduction-for-fable-cache-reads>
- [13] MacRumors 「Claude Fable 5.1 発表に関する報道」 2026 年 9 月 1 日。（URL 未確認）
- [14] CNBC 「GPT-6 Astra 発表に関する報道」 2026 年 9 月 3 日。（URL 未確認）
- [15] Yotta Labs 「GPT-6 Astra: Release Date, Pricing, Benchmarks, and Rollout (2026)」 2026 年 9 月。
<https://www.yottalabs.ai/post/gpt-6-release-date-rumors-what-is-known-2026>
- [16] The Register 「With Gemini 3.8 Flash, Google reminds everyone it's still in the race」 2026 年 9 月 2 日。
<https://www.theregister.com/ai-and-ml/2026/09/02/with-gemini-38-flash-google-reminds-everyone-its-still-in-the-race/5294049>
- [17] Meta Superintelligence Labs（研究ブログ） 「Introducing Muse Spark 1.3」 2026 年 9 月 2 日。
（URL 未確認）
- [18] MarkTechPost 「Meta AI Released Muse Spark 1.3: An Agentic Coding Model That Uses ~20% Fewer Tool Calls and ~25% Fewer Tokens Than Muse Spark 1.2」 2026 年 9 月 3 日。
<https://www.marktechpost.com/2026/09/03/meta-ai-released-muse-spark-1-3-an-agentic-coding-model-that-uses-20-fewer-tool-calls-and-25-fewer-tokens-than-muse-spark-1-2/>
- [19] Axios／BigGo Finance 「Alexandr Wang 氏（Meta 最高 AI 責任者）の Muse Spark 1.3 性能に関する

発言報道」2026年9月。（URL未確認）

- [20] WinBuzzer 「Meta Releases Muse Spark 1.3 Model for Longer Tool-Based Work」 2026年9月4日.
<https://winbuzzer.com/2026/09/04/meta-releases-muse-spark-1-3-model-longer-tool-based-work-xcxwbn/>
- [21] Amazon Web Services (What's New) 「Claude Fable 5.1, Anthropic's new frontier model is now available on AWS」 2026年9月. <https://aws.amazon.com/about-aws/whats-new/2026/09/claude-fable-5-1-aws/>
- [22] DataCamp 「GPT-6 Astra: Features, Benchmarks, and Pricing」 2026年9月.
<https://www.datacamp.com/blog/gpt-6-astra>
- [23] Swfte 「AI Model Leaderboard September 2026 — LMSys Arena, LLM, Image & Coding Rankings」 2026年9月. <https://www.swfte.com/ai/leaderboard>
- [24] DEV Community (G. Anhaia) 「GPT-6 Astra vs Fable 5.1 vs Gemini 3.8 Flash: The Ultimate Comparison」 2026年9月. <https://dev.to/gabrielanhaia/gpt-6-astra-vs-fable-51-vs-gemini-38-flash-the-ultimate-comparison-24g0>
- [25] MindStudio 「GPT-6 Astra Benchmarks: Is It Really Better Than Fable 5.1?」 2026年9月.
<https://www.mindstudio.ai/blog/gpt-6-astra-benchmarks-analysis>
- [26] WinBuzzer 「GPT-6 Astra Arrives With Major Gains, Staged Access, and New Questions About Its Benchmarks」 2026年9月4日. <https://winbuzzer.com/2026/09/04/gpt-6-astra-arrives-with-major-gains-staged-access-and-new-questions-about-its-benchmarks-xcxwbn/>
- [27] BenchLM.ai 「LLM Leaderboard History: Arena Rankings 2023–2026 (September 2026)」 2026年9月. <https://benchlm.ai/llm-leaderboard-history>
- [28] Emergent 「GPT-6 Astra Benchmarks: What the Numbers Actually Show」 2026年9月.
<https://emergent.sh/learn/gpt-6-astra-benchmarks>
- [29] Epoch AI 「FrontierMath Tier 4 (v2) (ベンチマークページ)」 2026年9月閲覧.
<https://epoch.ai/benchmarks/frontiermath-tier-4-v2>
- [30] The New Stack/Vellum 「FrontierMath に対する OpenAI の資金提供・排他アクセスに関する記事」 2026年. （URL未確認）
- [31] Epoch AI 「FrontierMath Erdős (68 の Erdős 未解決問題ベンチマーク) 公表」 2026年9月. （URL未確認）
- [32] StartupHub.ai 「Mark Zuckerberg's Dual AI Strategy: Llama 5 and Muse Spark 2026」 2026年6月5日. <https://www.startuphub.ai/ai-news/ai-figures/2026/figure-mark-zuckerberg-dual-track-open-closed-2026-06-05>
- [33] TechCrunch 「Meta's new Glimmer AI model offers a hint at Zuckerberg's personal intelligence

- vision」2026 年 8 月 10 日. <https://techcrunch.com/2026/08/10/metas-new-glimmer-ai-model-offers-a-hint-at-zuckerbergs-personal-intelligence-vision/>
- [34] Futurum Group 「Meta Reopens Its Models. Is This a PC Play or a Policy Play?」2026 年. <https://futurumgroup.com/insights/meta-reopens-its-models-is-this-a-pc-play-or-a-policy-play/>
- [35] Codersera 「DeepSeek V4-Pro 0813 GA: Benchmarks & Pricing (2026)」2026 年 8 月. <https://codersera.com/blog/deepseek-v4-pro-0813-guide-2026/>
- [36] Developers Digest 「GLM-5.2 vs DeepSeek V4 vs Qwen3: The Open-Weights Coding Model Showdown (2026)」2026 年. <https://www.developersdigest.tech/blog/glm-5-2-vs-deepseek-v4-vs-qwen3-open-weights-coding-showdown>
- [37] OrcaRouter 「DeepSeek V4 Pro vs Qwen3.8 Max: Specialist vs All-Rounder」2026 年. <https://www.orcarouter.ai/blog/deepseek-v4-pro-vs-qwen-3-8-max>