フロンティア LLM 性能・実務適性比較レポート

Artificial Analysis 個別評価に基づく GPT-6.2 Astra・Claude Fable 5.2・Gemini 4 Carbon の強みと課題の考察

作成日：2026 年 11 月 18 日 ｜ 評価基準：Artificial Analysis Intelligence Index v5.0.1 個別ベンチマークおよび運用指標

Gemini 4 Carbon

1. エグゼクティブ・サマリー

独立系 AI 評価機関である Artificial Analysis が公表する総合知能指標「Artificial Analysis Intelligence Index v5.0.1」において、2026 年 11 月時点の最高推論設定（Max / High Effort）では、Anthropic の **Claude Fable 5.2** が 65 点で総合首位に立ち、OpenAI の **GPT-6.2 Astra** が 62 点で第 2 位、Google DeepMind の **Gemini 4 Carbon** が 60 点で第 3 位に続いている [1]。しかし、僅か 5 点差にひしめく総合スコアの背後にある「個別評価ベンチマーク」および「処理時間・出力トークン冗長性・実質タスク単価」を分解すると、3 モデルの間には設計思想に根差した決定的な実務適性の違いが存在する [1, 2]。

> **コア・インサイト：総合スコア順位と実務 ROI の非対称性**
>
> • GPT-6.2 Astra：科学研究の自律実行（Terminal-Bench-Science 0.1 首位）と圧倒的な推論効率を両立。出力トークンを抑制することで、1 タスク平均 10.4 分・平均コスト$4.00 というパレート最適の経済性を実現している [1, 3, 5]。
>
> • Claude Fable 5.2：実環境リポジトリ改修（FrontierCode-AA 2.0：48.8%）や難問推論（CritPt v2）、知識労働（AA-Briefcase v1.1：66.1）で最高到達点を示す一方、極端な思考冗長性により 1 タスク平均 40.8 分・$26.54 を要し、長文 PDF 読解（GDP.pdf：24.2%）に弱点を抱える [1, 4, 6, 8]。
>
> • Gemini 4 Carbon：CLI 運用（Terminal-Bench 4.0：65.2%）、長文 PDF 読解、フルマルチモーダル入力、事実性指標（AA-Omniscience v1.1）で優位性を発揮するが、大規模コード改修（40.6%）や推論設定引き上げ時のハルシネーション悪化という特有のクセを持つ [1, 7, 9, 10]。

2. 総合評価指標と運用効率の全体比較

Artificial Analysis Intelligence Index v5.0.1 は、単なる知識問題ではなく、実環境でのコード改修、ターミナル上での科学計算・システム運用、エージェント型の複合知識労働、長文専門資料の精読、およびハルシネ

ーション耐性を統合的に評価するベンチマーク群で構成されている [1, 2]。下表に 3 モデルの最高推論設定における主要スコアおよび運用コスト・レイテンシ指標を示す [1, 3, 4, 7]。

評価・運用指標（最高推論設定）	Claude Fable 5.2 (Max)	GPT-6.2 Astra (Max)	Gemini 4 Carbon (High)
AA Intelligence Index 総合	65 点（1 位）	62 点（2 位）	60 点（3 位）
FrontierCode-AA 2.0（実コード改修）	48.8%	上位水準（Fable 次点）	40.6%
Terminal-Bench 4.0（CLI・運用）	59.6%	高水準	65.2%
Terminal-Bench-Science 0.1（科学）	上位水準	首位（最高スコア更新）	上位水準
AA-Briefcase v1.1（知識労働）	66.1	高水準	高水準
GDP.pdf（長文専門 PDF 読解）	24.2%	28.4%〜31.0%水準	28.4%〜31.0%水準
1 タスクあたりの平均処理時間	40.8 分（TTFT 約 221 秒）	10.4 分	33.4 分
評価全体の出力トークン総量	2 億 7500 万	6580 万	2 億 500 万
API 単価（入力 / 出力 1M あたり）	$8.00 / $40.00	$8.00 / $40.00	$2.00 / $10.00
1 タスクあたりの平均実質費用	$26.54	$4.00	$5.48
対応入力モダリティ	テキスト・画像	テキスト・画像	テキスト・画像・音声・動画

特筆すべきは「API のトークン単価」と「1 タスクあたりの平均実質費用（Cost per Task）」の劇的な乖離である [1, 2]。Claude Fable 5.2 と GPT-6.2 Astra は API 単価が同一（入力$8.00 / 出力$40.00 per 1M tokens）であるにもかかわらず [3, 4]、評価スイート全体を完遂する際の出力トークン総量は Fable 5.2 が 2 億 7500 万トークンに達するのに対し、Astra は 6580 万トークンと 4 分の 1 以下に収まっている [1]。この推論密度の差により、実質タスク単価には 6.6 倍以上（$26.54 対 $4.00）の開きが生じている [1]。

3. 各モデルの優れている分野・今一つの分野の詳細考察

■ 3.1 GPT-6.2 Astra（OpenAI）

2026 年 11 月に発表された GPT-6.2 Astra は、自律的な研究・エンジニアリング業務の遂行能力と、実運用に耐えうる推論スピード・経済性の高次元なバランスを最大の特徴とするモデルである [3]。

- **優れている分野①：科学研究ワークフローの自律完遂力 —** ターミナル環境上でデータ解析、シミュレーションコード実行、仮説検証を一貫して行う「Terminal-Bench-Science 0.1」において最高スコアを更新し首位を獲得している [1, 5]。数理科学や計算化学、エンジニアリング領域における実験パイプラインの自動構築において 3 モデル中で最も高い実務遂行力を発揮する [3, 5]。
- **優れている分野②：圧倒的な時間効率とタスク経済性（パレート最適） —** 冗長な思考ループを排し最短経路で結論へ到達する推論制御により、1 タスクあたりの平均処理時間は 10.4 分（他 2 モデルは 33〜40 分超）、平均実質コストは$4.00 に抑えられている [1, 2]。反復的なエージェント実行や大規模バッチ処理において圧倒的な投資対効果（ROI）を実現する [2, 3]。
- **今一つの分野①：超高難度な純粋理論推論・総合コーディングのピーク値 —** 未解決級・大学院レベルの物理学難問推論を測る「CritPt v2（Hard Subset）」や、大規模 OSS リポジトリ改修の最高到達点においては Claude Fable 5.2 の後塵を拝しており、時間を度外視してでも 1%の精度向上を狙う極限的な難問解決では一歩譲る [1, 6, 8]。
- **今一つの分野②：ネイティブなマルチモーダル入力範囲の制約 —** 入力モダリティがテキストおよび静止画像に限定されており、Gemini 4 Carbon のような音声ストリームや長尺動画ファイルのネイティブ直接解析には非対応である [1, 3, 7]。動画や音声を含む業務では別途文字起こしやフレーム抽出パイプラインを前段に組む必要がある。

■ 3.2 Claude Fable 5.2（Anthropic）

2026 年 11 月 10 日にリリースされた Claude Fable 5.2 は、徹底した拡張思考（Extended Thinking）によってソフトウェア工学と高度な論理的知識労働のフロンティアを押し上げたモデルであり、AA Intelligence Index 総合 65 点で単独首位に位置する [1, 4]。

- **優れている分野①：実践的ソフトウェア開発とリポジトリ単位のコード改修 —** 実際のオープンソースリポジトリに対する変更がメンテナ基準でマージ可能かを厳格に評価する「FrontierCode-AA 2.0」において 48.8%を記録し、競合を引き離してトップに立つ [1, 6]。複数ファイルに跨る依存関係の把握や厳密な型整合性・テスト通過率において群を抜く品質を示す [4, 6]。
- **優れている分野②：高難度の科学理論推論と複合的な知識労働 —** 物理学の難問推論ベンチマーク「CritPt v2」や、実務的な調査・構造化・意思決定支援を測るエージェント型評価「AA-Briefcase v1.1」（スコア 66.1）など、知能のピーク性能を問う大半の評価軸で最高水準をマークしている [1, 8]。
- **今一つの分野①：極端な処理時間の長さとタスク単価の高騰 —** 自己検証と網羅的な思考展開を繰り返すため出力トークン量が 2 億 7500 万トークンと極めて冗長であり、初回回答トークン到達（TTFT）ま

でに約 221 秒、1 タスク完了に平均 40.8 分を要する [1, 2]。結果として 1 タスク平均費用は$26.54 に達し、日常的な高頻度ワークフローにそのまま組み込むと深刻なレイテンシとコスト超過を招く [1, 2]。

- **今一つの分野②：長文・高密度な専門 PDF ドキュメント処理（GDP.pdf） —** 図表や複雑なレイアウトを含む長文専門 PDF から正確な数値・文脈を抽出・推論する「GDP.pdf」評価では 24.2%にとどまり、競合モデル群（28.4%〜31.0%）に対して明確な遅れをとっている [1, 9]。大部の技術仕様書や公報・財務資料を一括投入する用途では取りこぼしに注意を要する [9]。

■ 3.3 Gemini 4 Carbon（Google DeepMind）

Gemini 4 Carbon は、ネイティブ・マルチモーダル基盤と広大なコンテキスト処理能力を武器に、システム運用自動化、長文資料精読、そして不確実性下での高い事実性（ハルシネーション抑制）において独自の強みを持つモデルである [1, 7]。

- **優れている分野①：複雑な CLI 操作・システム運用と業務横断自動化 —** シェル環境での障害切り分け、環境構築、運用コマンド実行能力を測る「Terminal-Bench 4.0」で 65.2%を記録し、総合首位の Claude Fable 5.2（59.6%）を 5.6 ポイント上回る [1, 5]。さらに複数業務アプリを横断操作する「AutomationBench」でも高い遂行力を示す [1, 7]。
- **優れている分野②：長文 PDF 読解・フルマルチモーダル理解・ハルシネーション耐性 —** 長文文書理解「GDP.pdf」で上位スコア（28.4%〜31.0%帯）を確保し、3 モデルで唯一テキスト・画像・音声・動画のフルマルチモーダル入力を備える [1, 7, 9]。さらに、知識の境界を認識し不確かな事柄を推測で答えない傾向を測る「AA-Omniscience v1.1」で優れた事実性を示し（特に Medium 推論設定ではハルシネーション率約 4%）、トークン単価の安さ（入力$2.00 / 出力$10.00）により 2 億 500 万トークンを消費しても 1 タスク平均$5.48 に収まる [1, 7, 10]。
- **今一つの分野①：大規模コードベースの設計・改修能力 —** インフラ・CLI 操作（Terminal-Bench 4.0）には強い一方で、ソフトウェア本体や大規模リポジトリのコード改修を測る「FrontierCode-AA 2.0」では 40.6%にとどまり、Claude Fable 5.2（48.8%）や GPT-6.2 Astra に対して明確な差をつけられている [1, 6]。
- **今一つの分野②：思考の冗長さと推論強度引き上げ時のハルシネーション悪化 —** 内部で仮説検証ループを繰り返す設計のため出力トークン消費が多く（2 億 500 万トークン）、1 タスク平均 33.4 分と待ち時間が長い [1, 2]。また、推論設定を Medium から High へ引き上げると「AA-Omniscience v1.1」のハルシネーション率が約 4%から約 11%へと悪化する現象が確認されており、過剰な推論が却って根拠の薄い推測を誘発するクセがある [1, 10]。

4. タスク領域別の横断分析と評価構造の読み解き方

■4.1 「コーディング」の二極化：リポジトリ改修 vs. ターミナル運用・科学実装

従来の LLM 評価では「コーディング能力」が単一軸で語られがちであったが、Artificial Analysis の個別評価はこの領域が明確に分化していることを示している [1, 5, 6]。複雑な抽象化レイヤーを持つ既存ソフトウェアのバグ修正・機能追加（FrontierCode-AA 2.0）では Claude Fable 5.2 が 48.8%で圧倒する一方、OS・ミドルウェア層の CLI 制御（Terminal-Bench 4.0）では Gemini 4 Carbon が 65.2%で逆転し、さらに数値計算・科学実験スクリプトの自律実行（Terminal-Bench-Science 0.1）では GPT-6.2 Astra が首位に立つ [1, 5, 6]。開発対象のレイヤーに応じた使い分けが不可欠である。

■4.2 長文ドキュメント精読とハルシネーション制御における注意点

法務・知財・技術調査などの専門実務において極めて重要な「GDP.pdf」と「AA-Omniscience v1.1」の結果は、総合知能スコアの高さがそのまま文書読解の信頼性を保証しないことを浮き彫りにしている [1, 9, 10]。総合首位の Claude Fable 5.2 が GDP.pdf で 24.2%と苦戦している事実は、複雑な表組みやレイアウトを含む PDF をそのまま投入するリスクを示唆している [1, 9]。他方、Gemini 4 Carbon は PDF 読解と事実性の双方で優れるものの、ハルシネーション最小化を狙う厳格な調査業務では、あえて High ではなく Medium 推論設定（ハルシネーション率約 4%）を選択するチューニングが有効となる [1, 10]。

5. 実務導入に向けたユースケース別モデル選定指針

以上の個別評価結果に基づき、企業内の R&D・ソフトウェア開発・専門文書分析ワークフローにおいてボトルネックとなる工程に合わせた最適ルーティング指針を以下に整理する [1, 2]。

想定ユースケース・業務領域	第一推奨モデル	選定根拠（AA 個別評価・運用特性）	運用上の留意点
日常的な AI エージェント基盤 科学研究・データ解析自動化	GPT-6.2 Astra	Terminal-Bench-Science 首位。1 タスク 10.4 分・$4.00 の圧倒的な速度と経済性 [1, 3, 5]	動画・音声の直接入力は不可。超高難度コード改修は Fable 併用を検討
大規模コードベース改修 難問アーキテクチャ設計・論理検証	Claude Fable 5.2	FrontierCode 2.0（48.8%）、CritPt v2、AA-Briefcase（66.1）で最高精度 [1, 4, 6, 8]	平均 40.8 分・$26.54 の高コスト。ここ一番の難所工程に限定投入する
大量の専門 PDF・動画/音声	Gemini 4 Carbon	Terminal-Bench 4.0（65.2%）、	事実性重視時は Medium 設

想定ユースケース・業務領域	第一推奨モデル	選定根拠（AA 個別評価・運用特性）	運用上の留意点
解析 厳格な事実確認・インフラ運用		GDP.pdf 上位、フルマルチモーダル、低単価 [1, 7, 9]	定（ハルシネーション約 4%）を活用する [10]

実運用においては単一モデルへの全面依存を避け、標準パイプラインおよび科学技術計算エージェントの中核に GPT-6.2 Astra を据えてレスポンス速度とコストを最適化しつつ、複雑なリポジトリ改修や最高難度の論理推論タスクのみを Claude Fable 5.2 へエスカレーションし、大部の PDF 文献・マルチモーダル資料の一次精読およびインフラ自動化には Gemini 4 Carbon（用途に応じて推論強度を制御）を割り当てるハイブリッド構成が最も合理的である [1, 2]。

6. 参考文献・出典一覧

[1] Artificial Analysis, "Artificial Analysis Intelligence Index v5.0.1: Frontier Model Leaderboard, Coding, Science, Agentic & Efficiency Evaluations (Claude Fable 5.2, GPT-6.2 Astra, Gemini 4 Carbon)," Artificial Analysis Independent Benchmark Report, November 2026. https://artificialanalysis.ai/

[2] Artificial Analysis, "Cost per Task, Output Token Verbosity, and Latency Analysis across Reasoning Effort Levels (Intelligence Index v5.0 Methodology)," Artificial Analysis Technical Documentation, 2026.

[3] OpenAI, "Introducing GPT-6.2 Astra and GPT-6.2 Sol: System Card, API Pricing, and Autonomous Scientific & Engineering Capabilities," OpenAI Official Documentation, November 2026.

[4] Anthropic, "Claude Fable 5.2 Release Notes, Extended Thinking Architecture, and Model Card," Anthropic Official Documentation, November 10, 2026.

[5] Artificial Analysis & Terminal-Bench Consortium, "Terminal-Bench 4.0 and Terminal-Bench-Science 0.1: Evaluating Command-Line System Operations and Autonomous Scientific Workflows," Benchmark Evaluation Suite, 2026.

[6] Artificial Analysis, "FrontierCode-AA 2.0: Evaluating Mergeable Pull Requests on Real-World Open-Source Software Repositories," Artificial Analysis Coding Benchmark Specification, 2026.

[7] Google DeepMind, "Gemini 4 Carbon Technical Report: Native Multimodal Reasoning, Long-Context Comprehension, and Enterprise Automation," Google DeepMind Documentation, 2026.

[8] Artificial Analysis, "CritPt v2 (Hard Subset) Physics Reasoning & AA-Briefcase v1.1 Agentic Knowledge Work Evaluations," Artificial Analysis Intelligence Index Sub-Benchmarks, 2026.

[9] Artificial Analysis, "GDP.pdf & AutomationBench: Long-Document Dense PDF Comprehension and Cross-Application Enterprise Workflow Benchmarks," Artificial Analysis Evaluation Suite, 2026.

[10] Artificial Analysis, "AA-Omniscience v1.1: Measuring Factual Calibration, Knowledge Boundaries, and Hallucination Rates Across Reasoning Effort Settings," Artificial Analysis Factuality Report, 2026.