Artificial Analysis の個別評価にみる GPT-6.2 Astra・Claude Fable 5.2・Gemini 4 Carbon の 得意分野と今一つの分野

2026 年 11 月 18 日

Claude Fable 5.2

1. 要旨

独立系評価機関 Artificial Analysis（以下 AA）の Intelligence Index v5.0.1 において、Claude Fable 5.2（max 設定）は 65[1)]、GPT-6.2 Astra（max 設定）は 62[2)]、Gemini 4 Carbon（high 設定）は 60[3)]を記録している。総合点の差は 5 点以内だが、指数を構成する個別評価をみると、3 モデルの得意・不得意は明確に分かれる。

- **Claude Fable 5.2**：知識・推論、物理推論、高難度コーディング、長期の知識労働で 3 モデル中最高である。一方、1 タスク当たりのコストと所要時間が突出し、長文 PDF 読解と端末コーディングは今一つである。
- **GPT-6.2 Astra**：科学研究ワークフローの実行、3D 構築、ロボティクス、コスト効率に強みがある。幻覚率の高さと、長期の知識労働（成果物作成）が弱点である。
- **Gemini 4 Carbon**：幻覚の少なさ、業務自動化、端末コーディングに強みがある。知識の幅、高難度コーディング、処理時間が弱点である。

なお、GPT-6.2 Astra の個別評価値の一部は二次情報経由であり、未確認の項目もある。確認状況は第 6 章に整理した。本稿では、確認できた事実と本稿の考察を区別して記す。

2. 評価の枠組みと比較条件

Intelligence Index v5.0.1 は、AA-Briefcase v1.1、AutomationBench-AA、Terminal-Bench 4.0、FrontierCode-AA 2.0、AA-Omniscience v1.1、GDP.pdf、HLE-Diamond、CritPt v2（Hard Subset）、Terminal-Bench-Science 0.1 の 9 評価で構成される [2)]。本稿は、この 9 評価に指数外の AA-3D、AA-Robotics v0.1 等を加えて比較する。

比較に用いた設定は、GPT-6.2 Astra と Claude Fable 5.2 が最上位の max、Gemini 4 Carbon が同モデルで最上位の high である。Carbon の設定は high・medium・low の 3 段階で [3]、推論量の条件は揃っていない。また、AA は Claude Fable 5.2 の構成を「Default Fallback」と表記している [1]。

3. 個別評価の比較

表 1　AA 個別評価の比較（太字は 3 モデル中の最高値）

評価（測る内容）	GPT-6.2 Astra（max 設定）	Claude Fable 5.2（max 設定）	Gemini 4 Carbon（high 設定）
総合 Intelligence Index	62[2]	**65**[1]	60[3]
HLE-Diamond（知識＋推論）	2 位（数値未確認）[4]	**80.2%**[5]	69.1%[5]
AA-Omniscience v1.1（知識の信頼性）	Net 未確認／幻覚率 59%[4]	**67**（正答率 79%）[5],[6]	54（正答率 58%、幻覚率 11%）[5],[7]
CritPt v2 Hard（物理推論）	71.5%[8]	**76.5%**[5]	69.0%[5]
Terminal-Bench-Science 0.1（科学研究ワークフロー）	**78.1%**[9],[8]	73.3%[5]	64.8%[5]
FrontierCode-AA 2.0（高難度コーディング）	45.4%（xhigh は 49.0%）[8],[9]	**48.8%**（high は 49.1%）[5]	40.6%[5]
Terminal-Bench 4.0（端末コーディング）	64.1%[8]	59.6%[5]	**65.2%**[5]
AutomationBench-AA（業務 SaaS ワークフロー）	77.2%[8]	76.4%[5]	**77.8%**[5]
AA-Briefcase v1.1（長期の知識労働、Elo）	1644[10],[8]	**1785**[5]	1611[5]
GDP.pdf（長文 PDF 読解、全項目合格率）	未確認	24.2%[5]	28.4%[5]
1 タスク当たりコスト／所要時間	**$4.00／10.4 分**[2]	$26.54／40.8 分[1]	$5.48／31.1 分[3]
（指数外）AA-3D（Elo）／AA-Robotics v0.1	**1318／79**[4]	1271／75[4]	確認できず

注：「未確認」は本稿作成時点で数値を確認できなかった項目である。コスト・所要時間は Intelligence Index の 1 タスク当たりである。GDP.pdf は GPT-6.2 Astra の値が未確認のため、最高値の太字表示をしていない。

4. モデル別の考察

4.1 GPT-6.2 Astra

（1）得意分野

- **科学研究ワークフローの実行：**Terminal-Bench-Science 0.1 は 78.1%で 3 モデル中最高である [9]。AA も、OpenAI が同評価で首位に立つと報告している [4]。
- **3D 構築・ロボティクス：**AA-3D は 1318（Fable 5.2 は 1271）、AA-Robotics v0.1 は 79（同 75）で、いずれも首位である [4]。
- **コスト効率：**Intelligence Index の 1 タスク当たり$4.00、所要 10.4 分である [2]。Fable 5.2（$26.54、40.8 分）[1]の約 15%の費用、約 4 分の 1 の時間で、総合 3 点差に収まる（比率は本稿の計算）。AA は、OpenAI のモデルがコストと出力トークンの両面でパレート最前線の大半を占めるとしている [4]。
- **知識・推論：**CritPt v2、AA-Omniscience の正答率、HLE-Diamond は、いずれも Fable 5.2 に次ぐ 2 位である [4]。HLE-Diamond の推論パートは 84.8%で、Fable 5.2（85.6%）とほぼ並ぶ [11]。
- **業務自動化・端末コーディング：**AutomationBench-AA は 77.2%、Terminal-Bench 4.0 は 64.1%で [8]、首位の Carbon との差は 0.6〜1.1 ポイントにとどまる。
- **法務（指数外）：**AA の法務評価では、通常版の全項目合格率が 15.0%、法務特化構成「Astra for Law」が 20.8%と報じられている [12]。OpenAI は Fable 5.2 の 2 倍超と主張するが、これは同社の公表値である [13]。相対的には強みだが、絶対水準は低い。

（2）今一つの分野

- **幻覚率：**AA-Omniscience の幻覚率は max 設定で 59%である [4]。前世代 GPT-6 Astra の 50%から 9 ポイント悪化したとされ [8]、知らない問いに推測で答える傾向が強いと読める。
- **長期の知識労働：**AA-Briefcase v1.1 は 1644 で [10]、Fable 5.2（1785）[5]と 141 点の差がある。
- **コーディングエージェント：**AA の Coding Agent Index（Codex ハーネス、max 設定）は前世代から横ばいである [4]。数値は Astra の 53 に対し、Claude Code 上の Fable 5.2 が 58 と報じられている [8],[14]。

- **高難度コーディング：**FrontierCode-AA 2.0 は max 設定で 45.4%と [8]、Fable 5.2 を下回る。ただし xhigh 設定では 49.0%との報告もあり [9]、設定による振れが大きい。

4.2 Claude Fable 5.2

（1）得意分野

- **知識・推論：**HLE-Diamond は 80.2%[11]、AA-Omniscience は 67（正答率 79%）[6]で、いずれも AA 掲載モデル中の最高である。HLE-Diamond の知識問題では、Carbon の 59%に対し 75%と大差がついた [15]。
- **物理推論：**CritPt v2（Hard Subset）は 76.5%で 3 モデル中最高である [5]。
- **高難度コーディング：**FrontierCode-AA 2.0 は 48.8%（high 設定では 49.1%）で、Carbon を 8 ポイント以上上回る [5]。
- **長期の知識労働：**AA-Briefcase v1.1 は 1785 で 3 モデル中最高である [5]。ただし全体では、Claude Sonnet 5.5（1816）と Claude Opus 5.5（1801）が上位にある [10]。
- **公開時の到達点：**公開時点では、9 評価中 5 評価（HLE-Diamond、AA-Omniscience、Terminal-Bench-Science、CritPt v2、FrontierCode）で最高値を更新した [16]。このうち Terminal-Bench-Science は、その後 GPT-6.2 Astra に首位を譲っている [4]。

（2）今一つの分野

- **長文 PDF 読解：**GDP.pdf は 24.2%で、Carbon（28.4%）を下回る。low 設定（26.8%）の方が max 設定より高く、推論量を増やしても改善していない [5]。
- **端末コーディング：**Terminal-Bench 4.0 は 59.6%で 3 モデル中最下位である [5]。
- **コスト・時間・冗長性：**1 タスク当たり$26.54、40.8 分を要し、出力トークンは 2 億 7500 万に達する。AA のコスト順位では 86 モデル中最下位である [1]。
- **幻覚率：**max 設定の幻覚率は GPT-6.2 Astra と同じ 59%とする二次情報がある（AA 公式での確認は未了）[8]。正答率が高いため、総合スコアが高くなる構図と考えられる（本稿の考察）。
- **3D・ロボティクス：**AA-3D、AA-Robotics v0.1 とも、GPT-6.2 Astra に及ばない [4]。

4.3 Gemini 4 Carbon

（1）得意分野

- **幻覚の少なさ：**AA-Omniscience の幻覚率は high 設定で 11%（正答率 58%）である [7)]。medium 設定では 4%で、AA 掲載モデル中で最も低い水準にある [6)]。知らないことは答えない傾向が強いと読める。
- **業務自動化：**AutomationBench-AA は 77.8%で 3 モデル中最高である [5)]。ただし 3 者の差は 1.4 ポイント以内である。
- **端末コーディング：**Terminal-Bench 4.0 は 65.2%で 3 モデル中最高である。全体でも、Kimi K3.1 Pro（67.2%）に次ぐ 2 位である [17)]。
- **長文 PDF 読解：**GDP.pdf は 28.4%で、Fable 5.2 を約 4 ポイント上回る [5)]。
- **単価：**API 価格は入力 100 万トークン当たり$2、出力同$10 で、1 タスク当たりコストは $5.48 である [3)]。

（2）今一つの分野

- **知識の幅：**AA-Omniscience の正答率 58%[7)]、HLE-Diamond の 69.1%[5)]は、Fable 5.2（79%、80.2%）と大きな差がある。
- **高難度コーディング：**FrontierCode-AA 2.0 は 40.6%で 3 モデル中最下位である [5)]。
- **科学研究ワークフロー・物理推論：**Terminal-Bench-Science 0.1 は 64.8%で GPT-6.2 Astra と約 13 ポイントの差があり、CritPt v2 も 69.0%で最下位である [5)]。
- **長期の知識労働：**AA-Briefcase v1.1 は 1611 で 3 モデル中最下位である [5)]。
- **処理時間・冗長性：**1 タスク当たり 31.1 分と GPT-6.2 Astra の約 3 倍を要し、出力トークンも 2 億 500 万と多い [3)]。

5. 総合考察

本章は、前章までの事実に基づく本稿の考察である。3 モデルの性格は次のように整理できる。

- **Claude Fable 5.2 は「知識と深い推論」型：**知識の正確さと検討の深さが成果を左右する調査・分析に向く。ただし、コストと時間の制約が大きい。

- **GPT-6.2 Astra は「科学実行とコスト効率」型：**科学技術系のエージェント作業や大量処理に向く。ただし、事実確認を要する用途では幻覚率の高さへの備えが要る。
- **Gemini 4 Carbon は「堅実・自動化」型：**誤答を避けたい定型業務の自動化に向く。ただし、知識の幅と高難度の開発作業では他の 2 モデルに劣る。

もっとも、いずれのモデルも絶対水準がなお低い評価がある。GDP.pdf の全項目合格率は、掲載モデルの最高でも 32.8%にとどまる [18]。法務評価の全項目合格率も 2 割前後である [12]。専門業務での利用は、人による検証を前提とすべきである。

6. 留意点（確認状況）

- **一次確認済み（AA 公式ページ）：**Fable 5.2 と Carbon の 9 評価値、GPT-6.2 Astra の総合指数・幻覚率・AA-3D・AA-Robotics、各モデルのコスト・所要時間・出力トークン。
- **二次情報経由：**GPT-6.2 Astra の Terminal-Bench-Science、Terminal-Bench 4.0、FrontierCode-AA、AutomationBench-AA、CritPt v2 の数値は、AA 値を引用する個人ブログ等 [8),9)]から取得した。AA 公式ページの本文では照合できていない。AA-Briefcase の 1644 は、AA 公式ページの検索結果抜粋と二次情報により確認した。Carbon の正答率 58%・幻覚率 11%、HLE-Diamond 知識問題の内訳、Coding Agent Index の数値、法務評価の合格率も二次情報による。
- **未確認：**GPT-6.2 Astra の GDP.pdf、HLE-Diamond の総合値、AA-Omniscience の Net Score。Net Score については、Carbon が Astra を 4 点上回るとの報道がある [14]。
- **条件差：**Carbon は high が最上位設定であり、他の 2 モデルの max とは同条件でない。AutomationBench-AA の 3 者差（1.4 ポイント以内）は、実質同等とみるのが妥当である（本稿の評価）。
- **幻覚率の意味：**AA-Omniscience の幻覚率は、答えを知らない問いに対し、回答を控えずに誤答した割合を測る指標である [6]。極めて難しい事実問題が対象であり、日常利用の誤答率を示すものではない [13]。
- **鮮度：**GPT-6.2 Astra は 2026 年 11 月 17 日公開 [19]直後の値であり、AA の掲載値は今後更新される可能性がある。本稿の数値は 2026 年 11 月 18 日閲覧時点のものである。

参考文献

1) Artificial Analysis「Claude Fable 5.2: Intelligence, Performance & Price Analysis」（AA 公式）
https://artificialanalysis.ai/models/claude-fable-5-2（2026 年 11 月 18 日閲覧）

2) Artificial Analysis「GPT-6.2 Astra: Intelligence, Performance & Price Analysis」（AA 公式）
https://artificialanalysis.ai/models/gpt-6-2-astra（2026 年 11 月 18 日閲覧）

3) Artificial Analysis「Gemini 4 Carbon: Intelligence, Performance & Price Analysis」（AA 公式）
https://artificialanalysis.ai/models/gemini-4-carbon（2026 年 11 月 18 日閲覧）

4) Artificial Analysis「OpenAI's release of GPT-6.2 Astra and GPT-6.2 Sol delivers the #2 model in the Artificial Analysis Intelligence Index and most of the Cost per Task Pareto frontier」（AA 公式記事）
https://artificialanalysis.ai/articles/openai-has-the-2-model-in-the-intelligence-index-with-the-release-of-gpt-6-2-astra-and-gpt-6-2-sol（2026 年 11 月 18 日閲覧）

5) Artificial Analysis「Claude Fable 5.2 vs Gemini 4 Carbon: Release Comparison」（AA 公式）
https://artificialanalysis.ai/models/releases/comparisons/claude-fable-5-2-vs-gemini-4-carbon（2026 年 11 月 18 日閲覧）

6) Artificial Analysis「AA-Omniscience v1.1: Knowledge and Hallucination Benchmark」（AA 公式）
https://artificialanalysis.ai/evaluations/omniscience（2026 年 11 月 18 日閲覧）

7) XenoSpectrum「Google's Gemini 4 Carbon Tops One AI Index by Just 1 Point, With Big Leads Only in Select Tests」（二次情報） https://xenospectrum.com/en/gemini-4-carbon-benchmark-comparison/（2026 年 11 月 18 日閲覧）

8) Jake Handy「Model Drop: GPT-6.2 Astra & GPT-6.2 Sol」Substack（二次情報）
https://handyai.substack.com/p/model-drop-gpt-62-astra-sol（2026 年 11 月 18 日閲覧）

9) Kingy AI「GPT-6.2 Astra vs Claude Fable 5.2: Benchmarks and real costs」（二次情報）
https://kingy.ai/blog/gpt-6-2-astra-vs-claude-fable-5-2/（2026 年 11 月 18 日閲覧）

10) Artificial Analysis「AA-Briefcase v1.1: Agentic Knowledge Work Benchmark」（AA 公式）
https://artificialanalysis.ai/evaluations/aa-briefcase（2026 年 11 月 18 日閲覧）

11) Artificial Analysis「HLE-Diamond Benchmark」（AA 公式）
https://artificialanalysis.ai/evaluations/hle-diamond（2026 年 11 月 18 日閲覧）

12) Implicator.ai「GPT-6.2 Astra and Sol: What OpenAI's New Models Can Do」（二次情報）
https://www.implicator.ai/openai-gpt-6-2-astra-sol-cost-per-task/（2026 年 11 月 18 日閲覧）

13) The Decoder「OpenAI launches GPT-6.2 Astra and Sol with lower API prices and some benchmark gains」（二次情報） https://the-decoder.com/openai-launches-gpt-6-2-astra-and-sol-with-lower-api-prices-and-some-benchmark-gains/（2026 年 11 月 18 日閲覧）

14) OfficeChai「GPT 6.2 Astra Places Second On Artificial Analysis Intelligence Index, Edges Out Gemini 4 Carbon By 2 Points」（二次情報） https://officechai.com/ai/gpt-6-2-astra-artificial-analysis-intelligence-index/（2026 年 11 月 18 日閲覧）

15) OfficeChai「Anthropic Jumps To Top Of Artificial Analysis Intelligence Index With Fable 5.2, Beats Gemini 4 Carbon By 5 Points」（二次情報） https://officechai.com/ai/anthropic-jumps-to-top-of-artificial-analysis-intelligence-index-with-fable-5-2-beats-gemini-4-carbon-by-5-points/（2026 年 11 月 18 日閲覧）※検索結果の抜粋のみ確認（本文未取得）

16) Artificial Analysis（X 投稿）Claude Fable 5.2 公開時の評価結果（AA 公式） https://x.com/ArtificialAnlys/status/2119878995386982508（2026 年 11 月 18 日閲覧）※検索結果の抜粋のみ確認（本文未取得）

17) Artificial Analysis「Terminal-Bench 4.0 Benchmark」（AA 公式） https://artificialanalysis.ai/evaluations/terminalbench-4-0（2026 年 11 月 18 日閲覧）

18) Artificial Analysis「GDP.pdf Benchmark」（AA 公式） https://artificialanalysis.ai/evaluations/gdp-pdf（2026 年 11 月 18 日閲覧）

19) Crypto Briefing「GPT-6.2 Astra takes first place on the AA-3D benchmark for 3D experiences」（二次情報） https://cryptobriefing.com/gpt-6-2-astra-tops-aa-3d-benchmark/（2026 年 11 月 18 日閲覧）※検索結果の抜粋のみ確認（本文未取得）