フロンティア LLM の第三者評価

2026 年 11 月 20 日（日本時間） ｜ Artificial Analysis・Epoch AI・Vals AI・Arena

GPT-6.2 Astra

GPT-6.2 Astra は、科学の研究課題で最も明確な優位を示した。Claude Fable 5.2 は、知識・推論・実務を含む総合指数と、利用者が実際にエージェントに任せた仕事で首位に立つ。Gemini 4 Neon は米国の法務・金融課題で首位を取り、エージェント総合でも 3 位に入った。どの仕事を任せるかによって第一候補が変わる。［1–9］

主要 6 モデルの比較

モデル	AA 指数	Epoch ECI	Vals 総合 42 モデル	Vals 科学 28 モデル	Agent Arena 68 設定	WebDev 公式画面
GPT-6.2 Astra	62	174	8 位	1 位	6 位	3 位
Claude Fable 5.2	65	174	1 位	3 位	1 位	1 位
Gemini 4 Neon	61	—	2 位	4 位	3 位	7 位
Gemini 4 Carbon	60	170	4 位	—	4 位	6 位
Claude Opus 5.5	59	169	5 位	7 位	2 位	2 位
GPT-6.2 Sol	57	—	10 位	8 位	11 位	5 位

AA・ECI は得点、他は順位。「—」は参照表で数値を確認できない欄。AA は各モデルの代表設定（Neon は xhigh、Carbon は high、他 4 モデルは max）。Arena では Neon が high、WebDev の Fable が xHigh。AA はモデル単位で Astra が 2 位、設定別では Fable の max と xhigh に続く 3 位。［1–9, 20］

動画の見出しをどう読むか

Epoch AI の「同点 1 位」は、公表される整数表示で Astra と Fable が 174 に並ぶという意味である。Vals の「科学 1 位」は 3 種類の研究課題の平均で、Vals の産業別総合では Astra は 8 位。Agent Arena の「6 位」は、仕事の完了や指示修正への反応などを実利用セッションから推定した順位である。［3–6, 8, 10］

比較対象を広げると見えること

Vals 総合では Claude Sonnet 5.5 が 70.85%で 3 位。Neon の 70.87%との差は 0.02 ポイントにとどまる。Meta の Wonder Full は米国法務で 77.86%の 2 位。Qwen 4 Max は Vals 総合で 14 位、WebDev 公式画面で 9 位と評価が分かれる。［5, 7, 9］

総合指数と科学の成績が示すこと

同じ 174 点でも推定の幅は異なる

Epoch の ECI は、各試験の難易度とモデル能力を統計的に推定し、異なる試験を受けたモデルを一つの尺度に置く。241 モデルのうち、Astra は 11 試験から 174（90%信用区間 170–179）、Fable は 21 試験から 174（170–177）、Carbon は 15 試験から 170（167–173）と推定された。Astra と Fable のどちらが上かまで、この公表値から決めるのは難しい。［3, 4］

Epoch の個別課題	GPT-6.2 Astra	Claude Fable 5.2	Gemini 4 Carbon
FrontierMath Tier 4 v2	100%	100%	95%
規則を推理するゲームパズル	95%	90%	73%
SimpleQA Verified（事実知識）	79%	94%	78%
MirrorCode（ソフトウェア開発）	68%	未掲載	未掲載

出典［4］。MirrorCode の全掲載モデル中の最高値は 98%。Fable・Carbon の比較欄は未掲載のため、両者との優劣は判定できない。

科学では研究手順と分野を分けて読む

Vals Science Index で Astra は 68.44%。旧 GPT-6 Astra（65.09%）を 3.35 ポイント、Fable（63.74%）を 4.70 ポイント上回る。Neon は 62.63%。研究データの解釈、模擬実験を通じた機序解明、ターミナルを使う研究作業を等しく重み付けしている。［6］

Astra の科学 3 課題	成績・順位	得意分野の違い
Terminal-Bench-Science	71.43%・1 位／48	数学・工学は Astra、生命科学は Fable、地球科学は Neon が首位＊
BioMysteryBench	80.74%・3 位／34	実データから問いに答える。首位は Fable
MysteryMechanism	53.15%・2 位／34	模擬環境で機序を推定する。旧 GPT-6 Astra が首位

＊分野別順位は Vals 科学指数に含まれる 28 モデル内での順位。個別ベンチマークの全参加モデル数とは異なる。［6, 11］

研究を設計し報告まで仕上げると差が開く

Epoch が自組織の仕事を人手採点した別の検証では、平均は Astra が 65%、Fable が 17 ポイント上回った。データセンター調査は Fable 98%、Opus 88%、Astra 65%。Astra が作った別の研究報告には、冒頭の結論が後段の条件付き結果と食い違う例もあった。各モデル各課題 1 回で、採点には職員の判断を含む。［12］

エージェントと専門業務の評価

実利用では Fable が首位　Astra の推定順位には幅がある

Agent Arena の 11 月 18 日版で Fable（max）は 1,448 点。Opus（max）1,336 点、Neon（high）1,333 点、Carbon（high）1,303 点、Sonnet（max）1,296 点、Astra（max）1,291 点と続く。利用者が仕事の完了を示したか、称賛・苦情を述べたか、途中の指示に応じたか、シェル操作を正常に扱えたかを統合する。［8, 10］

Astra の公式モデル欄は 1,291±35、順位の幅は 2〜11 位。データセットでは 4,785 セッション・430,272 観測、スコア 0.1522（区間 0.1248–0.1796）。公式画面は基準モデルに合わせた尺度を使う。旧尺度の 0.152 を成功率 15.2%に換算することはできない。なお、コーディングに限った Agentic Coding Index では Astra は 3 位である。［8, 10, 13, 19］

Web 制作の好みと法律業務の合格条件

同じ指示で作った Web アプリを比較する WebDev では、公開データの Astra は 1,780.61 点で 3 位（1,102 票）。公式画面で確認できた値も 3 位で 1,797 点。モデルや設定の収録範囲・スコアが表示面で異なるため、それぞれの出典を併記した。公式画面で首位は Fable（xHigh）1,824 点、2 位は Opus（max）1,818 点。［9, 14］

米国の専門業務	GPT-6.2 Astra	比較すると分かること
Vals Legal Index	69.68%・23 位／43	Neon 78.36%で 1 位、Carbon 75.93%で 4 位、Fable 69.17%で 24 位
Vals Finance Index	57.31%・23 位／45	Neon 68.31%で 1 位、Carbon 67.05%で 2 位
Tax Agent Bench	21.84%・56 位／90	Neon は 54.89%。米国法人税の調査課題
Harvey の法務課題	AA 版 15.0% Vals 版 16.25%	AA 版の Fable は 7.2%。Vals 版の Astra は 12 位／43。実装と採点指標を併記する必要がある

出典［5, 7, 11, 15, 16］。AA の 15.0%は、誤った事実の有無を含む全要件を満たした課題の割合。Vals 法務は米国法律事務所の分野別収入で重み付けし、知財は企業・商取引分野の中に含まれる。

開発者発表の読み方

OpenAI は Astra を科学者向けの最高モデルと位置づける。第三者評価は数学・工学の研究作業でこの見方を支える一方、AA の難問物理 CritPt では Fable が 76.5%、Astra が 71.5%だった。科学の中でも、ターミナルで研究を進める力と物理問題を解く力では首位が入れ替わる。［2, 6, 17］

推論設定と費用から見たモデル選択

Astra は設定を一段下げても表示上は 62 点

モデルと推論設定	AA 指数	1 課題の費用	1 課題の時間	最初の回答まで
Fable 5.2　max	65	26.54 ドル	41.6 分	227.7 秒
Astra　max	62	4.00 ドル	14.3 分	206.7 秒
Astra　xhigh	62	2.52 ドル	7.7 分	66.3 秒
Astra　high	60	2.07 ドル	6.2 分	25.7 秒
Astra　medium	57	1.62 ドル	4.8 分	3.9 秒
Astra　low	46	0.66 ドル	1.9 分	2.8 秒

［1, 2］の 11 月 20 日閲覧表示。費用は AA 指数の課題をその配点で重み付けした平均、時間は AA の算定値。最初の回答までの時間と課題全体の時間は別指標。Fable は AA の既定フォールバックを含む設定。

Astra の max は Fable の max に対して、AA 課題では費用が約 15%、所要時間が約 34%。Astra の xhigh は max より費用が 37%少なく、時間も約半分である。整数表示が同点でも、HLE-Diamond のように max 72.2%、xhigh 67.3%と差の出る課題がある。［2］

費用の順序も業務で変わる。AA の標準価格ベースでは Astra が 4.00 ドル、Neon が 6.44 ドル、Carbon が 7.76 ドル。一方、Arena の利用記録に基づく中央値は、Astra 約 3.3 ドル、Neon 約 2.6 ドル、Carbon 約 1.4 ドルだった。作業の構成と集計方法が異なる。［1, 8, 18］

知財部門での検証に結びつける

業務と候補モデル	実案件で確かめたいこと
発明発掘・実施例の検討 GPT-6.2 Astra	研究報告から、観測事実、推定した作用機序、追加で確かめるべき点を分ける。担当研究者が、裏付けのない効果と条件の取り違えを数える。
複数資料を使う競合調査 Claude Fable 5.2	特許公報、企業発表、論文を同じ案件で照合する。重要文献の取りこぼし、根拠まで追える主張の割合、調査をやり直した回数を比較する。
契約・法務資料の調査 Gemini 4 Neon	日本語の契約レビューや拒絶理由対応で、引用箇所の正確さ、論点の漏れ、修正時間を採点する。Vals で測った米国法と自社業務の違いを検証する。
新着公報の定期監視・分類 GPT-6.2 Sol／Gemini 4 Carbon	既知の重要公報を含む評価用セットを使い、同じ再現率を達成するための費用と人の確認時間を比較する。

上表は公開結果を踏まえた用途仮説。公報、引用段落、請求項との対応を同じ入力・同じ検索環境で検証して初めて、日本語の特許実務における適否を判断できる。

原典と参照箇所

参照日　2026 年 11 月 20 日（日本時間）。下の資料名から各原典を開ける。ライブ表は随時更新される。

［1］**Artificial Analysis**　LLM Leaderboard　指数・各社代表設定・費用

［2］**Artificial Analysis**　Release Comparison Tool　Astra・Fable の設定別内訳

［3］**Epoch AI**　Epoch Capabilities Index　推定値・90%信用区間・試験数

［4］**Epoch AI**　GPT-6.2 Astra　公式同順位・個別課題のモデル比較

［5］**Vals AI**　Vals Index v2.2　4 産業の重みと総合順位

［6］**Vals AI**　Vals Science Index　3 課題の構成と分野別順位

［7］**Vals AI**　Vals Legal Index　米国法律事務所の収入構成と法務課題

［8］**Arena**　Leaderboard　Agent Arena と Agentic Coding の公式順位

［9］**Arena**　Code Arena WebDev　公式画面の投票順位とスコア

［10］**Arena**　Introducing the Agent Arena Index　シグナルと尺度変更

［11］**Vals AI**　GPT-6.2 Astra　テスト別成績・母数・発表時の数値

［12］**Epoch AI**　How good are Fable 5.2 and GPT-6.2 Astra at real work?

［13］**Arena**　Agent 公式配布データ　11 月 18 日版の観測数・区間

［14］**Arena**　WebDev 公式配布データ　11 月 18 日版の投票数・区間

［15］**Vals AI**　Vals Finance Index　米国金融・税務評価

［16］**Artificial Analysis**　Harvey LAB-AA v1.1　誤情報を考慮した全要件合格率

［17］**OpenAI**　Introducing GPT-6.2 Astra and Sol　開発者による評価

［18］**Artificial Analysis**　Gemini 4 Neon　Carbon の比較に適用した価格条件

［19］**Arena**　GPT-6.2 Astra on the Agent Arena Index　推定順位の幅

［20］**Artificial Analysis**　GPT-6.2 発表時の比較　モデル単位で 2 位とする説明

比較条件の補足　AA が発表記事でいうモデル単位の 2 位と、全推論設定を別行にした 3 位は両立する。［1, 20］　Agent での表示順位と推定の幅は公式モデル欄による。［19］

調査の起点　指定の YouTube 動画と付属の文字起こし