フロンティア LLM の第三者評価の現状

― GPT-6.2 Astra 公開後の主要評価機関の原典確認 ―

調査基準日：2026 年 11 月 20 日（日本時間）

Claude Fable 5.2

---

1. 調査の目的と方法

本稿は、YouTube 動画「GPT-6.2 Astra は最強？Epoch AI は同点 1 位、Vals は科学 1 位、Agent Arena は 6 位！第三者評価を比較」[1]が取り上げた第三者評価について、各評価機関の原典を確認し、2026 年 11 月 20 日時点のフロンティア LLM の評価状況を整理したものである。

対象は、Artificial Analysis、Epoch AI、Vals AI、Arena の 4 機関に、動画が扱っていない ARC Prize の検証結果を加えた 5 つである。GPT-6.2 Astra は OpenAI が 2026 年 11 月 17 日に公開した高難度業務向けモデルで[9,13]、API 価格は 100 万トークンあたり入力 8 ドル、キャッシュ入力 0.40 ドル、出力 40 ドルである[2]。

記述は次の 3 区分で書き分けた。

- **確認済み**：評価機関自身のページまたは公式投稿で確認できた事実。
- **二次情報**：報道や集計サイトを経由して確認した数値。原典での再確認が望ましい。
- **分析**：筆者による整理や推測。

リーダーボードは随時更新されるため、数値はいずれも基準日時点のものである。各出典をどこまで確認できたかは、文末の参考文献に個別に付記した。なお、本稿の作成には評価対象の一つである Claude Fable 5.2 を用いており、順位と数値は原典の記載のみに基づいて記した。

2. 結論（要旨）

- 動画の数値は原典とほぼ一致した。相違は、公開後の更新による変動（Arena の指標刷新、WebDev のスコア、Artificial Analysis の速度値、Vals の順位）にとどまる。
- 総合指標では Claude Fable 5.2 が先頭に立つ評価が多い。Artificial Analysis、Vals、Arena の各指数で 1 位、Epoch の総合指標では GPT-6.2 Astra と同点 1 位である（分析）。

- GPT-6.2 Astra は科学・数学・抽象推論で最上位にあり、評価コストも低い。一方、実務型の開放タスク、金融・法務・税務、事実知識では相対的に弱い（分析）。
- Gemini 4 系（Carbon、Neon ほか）は Vals と Arena で上位に複数入り、単価が低い（分析）。
- 「最強」かどうかは評価と条件で変わる。母集団、ハーネス、推論努力の設定が評価ごとに異なるため、順位を横に並べて比べることはできない。

3. 評価機関別の現状

表 1　主要評価の一覧（2026 年 11 月 20 日時点）

評価（運営）	上位	GPT-6.2 Astra	他の主なモデル
Intelligence Index v5.0.1（Artificial Analysis）	Claude Fable 5.2（max）65[4]	62。モデル単位で 2 位[5]、構成単位で 96 中 3 位[3]	Gemini 4 Carbon 60、Claude Opus 5.5 59[6]
Epoch Capabilities Index（Epoch AI）	Claude Fable 5.2 と GPT-6.2 Astra が 174 で同点[9]	174（241 モデル中 1 位、同点）[9]	Gemini 4 Carbon 170、GPT-6 Astra 169[9]
Vals Index v2.2（Vals AI）	Claude Fable 5.2 71.31%[12]	67.66%（42 モデル中 8 位）[12]	Gemini 4 Neon 70.87%、Claude Sonnet 5.5 70.85%、Gemini 4 Carbon 70.45%[12]
Vals Science Index（Vals AI）	GPT-6.2 Astra 68.44%[13]	28 モデル中 1 位[13]	—
Coding Index（Arena）	Claude Fable 5.2 1,489[18]	1,338（3 位）[18]	Claude Opus 5.5 1,373[18]
Work Index（Arena）	Claude Fable 5.2 1,445[19]	1,280（6 位）[19]	Claude Opus 5.5 1,317、Gemini 4 Neon 1,306[19]
Code Arena WebDev（Arena）	Claude Fable 5.2 1,824[21]	1,797（3 位）[21]	Claude Opus 5.5 1,818[21]
ARC-AGI-3 標準ハーネス（ARC Prize）	Claude Fable 5.2 81.17%（二次情報）[26]	70.3%（max）[24]	Gemini 4 Carbon 65.12%（二次情報）[26]
ARC-AGI-3 Provider Adapter（ARC Prize）	GPT-6.2 Astra 100.0%[24,25]	100.0%（xhigh）[24]	GPT-6 Astra 99.95%（二次情報）[26]

注：ARC-AGI-3 の「上位」は、二次情報[26]が比較した 4 モデルの中での最高値であり、リーダーボード全体の首位かどうかは確認していない。

3.1　Artificial Analysis（Intelligence Index v5.0.1）

知識労働、コーディング、科学、一般知識にまたがる 9 つの評価を合成した指数である[7]。GPT-6.2 Astra（max）は 62 で、比較対象モデルの中央値 43 を大きく上回る[3]。首位は Claude Fable 5.2（max）の 65 で[4]、Gemini 4 Carbon（high）の 60、Claude Opus 5.5（max）の 59 が続く[6]。

順位の表示は 2 通りある。同社の公式投稿は Astra を「2 番手のモデル」とし[5]、モデルページは「96 中 3 位」と表示する[3]。比較ページでは Claude Fable 5.2（max）、同（xhigh）、GPT-6.2 Astra（max）の順に並ぶため[4]、推論努力ごとの構成を別々に数えた結果が 3 位だと考えられる（分析）。

評価 1 タスクあたりのコストは Astra（max）が 4.00 ドル[3]、Claude Fable 5.2（max）が 26.54 ドル（二次情報）[8]で、約 6.6 倍の開きがある（倍率は筆者の計算）。前モデルの GPT-6 Astra（max）は指数 57、6.24 ドルだった（二次情報）[8]。ただしトークン単価（入力 8 ドル、出力 40 ドル）は、比較対象の中央値（2 ドル、10 ドル）の 4 倍である[3]。

表 2　GPT-6.2 Astra の推論努力別の値（Artificial Analysis）[3]

推論努力	指数	1 タスク所要時間	初回応答まで	出力トークン
max	62	13.7 分	197.65 秒	6,580 万
xhigh	62	7.4 分	69.30 秒	2,710 万
high	60	5.9 分	27.93 秒	2,000 万
medium	57	4.5 分	3.54 秒	1,240 万
low	46	1.8 分	2.91 秒	631 万
中央値（比較対象）	43	12.3 分	—	6,510 万

注：max と xhigh は指数が同じで、xhigh は所要時間と出力トークンが半分前後に収まる。所要時間と初回応答時間は実測値で、日によって変動する。

コーディングと信頼性には課題が残る。同社の Coding Agent Index v2.1 では Astra（max）は 53 で前モデルと変わらず、Claude Fable 5.2（high）の 58 を下回る（二次情報）[6]。難問を集めた AA-Omniscience での誤答（ハルシネーション）率は 59%と報じられている（二次情報）[6,7]。

3.2　Epoch AI（Epoch Capabilities Index）

多数のベンチマークを一つの尺度に統合した総合指標（ECI）で、GPT-6.2 Astra は 174 となり、追跡対象 241 モデルの中で Claude Fable 5.2 と同点の 1 位である[9]。Gemini 4 Carbon は 170 で 3 位、前モデルの GPT-6 Astra は 169 で 4 位に位置する[9]。Epoch は、この値が公開前に自社ベンチマークだけで行った試験に基づくもので、今後変わりうると注記している[11]。

表 3　Epoch AI の個別ベンチマーク比較[9]

分野	ベンチマーク（記録上の最高値）	GPT-6.2 Astra	GPT-6 Astra	Claude Fable 5.2	Gemini 4 Carbon
総合	ECI（174）	174	169	174	170
数学	FrontierMath Tier 4 v2（100%）	100%	98%	100%	95%
数学	FrontierMath Erdős（10%）	10%	3%	n/a	3%
ゲーム	Mystery Game Puzzles（95%）	95%	84%	90%	73%
ゲーム	EBR-bench（92%）	92%	76%	90%	88%
ゲーム	SpireClimb（41%）	41%	34%	15%	n/a
ゲーム	Chess Puzzles（79%）	78%	76%	79%	78%
マルチモーダル	Furniture Assembly Benchmark（90%）	89%	80%	90%	78%
ソフトウェア開発	MirrorCode（98%）	68%	60%	n/a	n/a
世界知識	SimpleQA Verified（94%）	79%	76%	94%	78%

注：n/a は当該ページに値の掲載がないことを示す。MirrorCode の最高値 98%を記録したモデルは、この比較表には含まれていない。

同点 1 位は「全項目で 1 位」を意味しない。数学とパズル系は最高値に並ぶが、事実知識の SimpleQA Verified では Claude Fable 5.2 に 15 ポイント及ばない[9]。

動画が触れていない重要な結果として、Epoch 自身の業務を題材にした実務タスク評価（Epoch Automation Reports）がある。平均スコアは Claude Fable 5.2 が約 82%、Claude Opus 5.5 が約 78%、Gemini 4 Argon が約 67%、GPT-6.2 Astra が約 65%だった[11]。Astra は ECI で並んだ Fable 5.2 を 17 ポイント下回り、Epoch が社員に求める水準に達したモデルは一つもなかっ

た[10]。データセンター分析の課題では、Fable 5.2 が 98%、Opus 5.5 が 88%、Astra が 65%である[10]。

一方、未解決問題を集めた FrontierMath Erdős では、Astra が 68 問中 7 問を解き、それまでの記録（3 問）を更新した[11]。

3.3　Vals AI（Vals Index と Science Index）

Vals Index v2.2 は、金融、コーディング、法務、税務の 8 つのベンチマークで構成される[12]。GPT-6.2 Astra は 67.66%で 42 モデル中 8 位である[12,13]。公開当日の同社発表では 7 位だった[14]。上位 7 つは Anthropic の 3 モデルと Gemini 4 系の 4 モデルが占める[12]。

表 4　Vals Index v2.2 の上位 11 モデル[12]

順位	モデル	正答率	コスト／テスト	所要時間／テスト
1	Claude Fable 5.2	71.31%	$17.94	43 分 42 秒
2	Gemini 4 Neon	70.87%	$7.40	31 分 25 秒
3	Claude Sonnet 5.5	70.85%	$16.18	1 時間 26 分
4	Gemini 4 Carbon	70.45%	$10.51	47 分 45 秒
5	Claude Opus 5.5	70.14%	$24.77	1 時間 25 分
6	Gemini 4 Argon	69.39%	$13.37	44 分 58 秒
7	Gemini 4 Hydrogen	68.49%	$9.84	32 分 05 秒
8	GPT-6.2 Astra	67.66%	$10.68	52 分 16 秒
9	Wonder Full	66.79%	$6.81	59 分 19 秒
10	GPT-6.2 Sol	66.06%	$3.18	41 分 47 秒
11	GPT-6 Astra	65.02%	$21.28	33 分 39 秒

分野別に見ると評価が大きく分かれる。科学では Science Index と Terminal-Bench-Science で 1 位を取り、「科学者にとって世界最高のモデル」とする OpenAI の主張[2]と方向が一致する。反対に、金融、法務、税務は中位以下である。評価は最大の推論努力（max）で実施されている[13]。

表 5　GPT-6.2 Astra のベンチマーク別結果（Vals AI）[13]

区分	ベンチマーク	スコア	順位／対象モデル数
総合	Vals Index	67.66%	8／42
科学	Vals Science Index	68.44%	1／28
科学	Terminal-Bench-Science	71.43%	1／48
科学	BioMysteryBench	80.74%	3／34
科学	MysteryMechanism	53.15%	2／34
コーディング	Terminal-Bench 4.0	63.13%	5／69
金融	EMB v1.1	76.03%	5／47
金融	Vals Finance Index	57.31%	23／45
法務	Vals Legal Index	69.68%	23／43
法務	Legal Research Bench	42.79%	31／99
法務	HLAB-Vals-1.1	16.25%	12／43
税務	Tax Agent Bench	21.84%	56／90

注：「区分」は筆者による整理である。

3.4　Arena（Agent Arena と Code Arena WebDev）

Agent Arena は、エージェントが実際に使われたセッションから得られる信号（タスクの成否、利用者の満足度、操縦のしやすさなど）で順位をつける[17]。Arena は 11 月 18 日、総合指標の「Agent Arena Index」と 6 つの個別指数（Agentic Work、Coding、Chat、Task Outcomes、Tool Use、Steering）を導入した[15,16]。導入時点の総合首位は Claude Fable 5.2（Max）の 1463 点で、2 位に 134 点の差をつけていた[15]。

動画が紹介した「スコア 0.152 で 6 位」は、この刷新前の指標による順位である。11 月 18 日時点の旧スコアは、Claude Fable 5.2 が 27.4、Claude Opus 5.5 が 18.6、Gemini 4 Neon が 18.4、Gemini 4 Carbon が 16.1、Claude Sonnet 5.5 が 15.6、GPT-6.2 Astra が 15.2（6 位）、GPT-6 Astra が 14.6（7 位）だった（二次情報）[23]。

表 6　Arena の個別指数と WebDev の上位[18,19,20,21]

指数	1 位	2 位	3 位	GPT-6.2 Astra（Max）
Coding Index	Claude Fable 5.2（Max）1,489	Claude Opus 5.5（Max）1,373	GPT-6.2 Astra（Max）1,338	3 位。$5.42／タスク、19.3 分
Work Index	Claude Fable 5.2（Max）1,445	Claude Opus 5.5（Max）1,317	Gemini 4 Neon（High）1,306	6 位（1,280）。$4.06／タスク、13.7 分
Chat Index	未確認	Claude Opus 5.5（Max）1,341	Gemini 4 Neon（High）1,322	11 位（1,211）。$2.10／タスク、7.2 分
Code Arena WebDev	Claude Fable 5.2（xHigh）1,824	Claude Opus 5.5（Max）1,818	GPT-6.2 Astra（Max）1,797	3 位（1,797）

注：コストと時間はタスクあたりの中央値。Chat Index の 1 位は確認できた抜粋に含まれていなかった。総合の Agent Arena Index における GPT-6.2 Astra の順位は確認できていない。

WebDev は、同じ指示から匿名の 2 モデルが作った Web アプリを人が比べて投票する評価である。公開翌日の発表では Astra は 1,781 点で 3 位、GPT-6.2 Sol は 1,757 点で 7 位だった[22]。現在のページ表示では Astra が 1,797 点（3 位のまま）、Sol が 1,771 点（5 位）に上がっている[21]。4 位以下は Claude Sonnet 5.5（1,773）、GPT-6.2 Sol（1,771）、Gemini 4 Carbon（1,769）、Gemini 4 Neon（1,764）と僅差で並ぶ[21]。

3.5　ARC Prize（ARC-AGI-3）※動画では扱われていない

ARC Prize は、GPT-6.2 Astra が ARC-AGI-3 で初めて 100%に達したと検証結果を公表した[25]。ただし、この満点は Provider Adapter と呼ばれるハーネスを使った場合（xhigh）の値である。標準ハーネスでは max で 70.3%、xhigh で 56.5%にとどまる[24]。ARC-AGI-2 は 95.8%、ARC-AGI-1 は 98.5%である[24]。

標準ハーネスは、モデルが自分で残すと決めたメモだけを環境内で持ち越せる方式である。Provider Adapter は、リクエスト間で推論状態を保持し、長い対話を圧縮する仕組みを使う[24]。標準ハーネスで比べると順位が入れ替わり、Claude Fable 5.2 が Astra を上回る（二次情報）[26]。

表 7　ARC-AGI-3 のハーネス別スコア（モデル別の最高値）

モデル	標準ハーネス	Provider Adapter	区分
GPT-6.2 Astra	70.3%	100.0%	確認済み[24]
Claude Fable 5.2	81.17%	97.72%	二次情報[26]
Gemini 4 Carbon	65.12%	99.09%	二次情報[26]
GPT-6 Astra	62.71%	99.95%	二次情報[26]

4. 動画の記述と原典の照合

動画の主要な数値を原典と突き合わせた結果を表 8 に示す。食い違いはいずれも、動画の公開後に原典側が更新されたことによるものである。

表 8　動画の記述と原典の照合

項目	動画の記述	原典の値（基準日時点）	判定
API 価格	入力 8 ドル、キャッシュ 0.40 ドル、出力 40 ドル。前モデル比 20%・60%安	同じ[2]	一致
AA 指数（max）	62、96 モデル中 3 位、中央値 43	同じ[3]。公式投稿はモデル単位で 2 位[5]	一致
AA の 1 タスクコスト	Astra 4.00 ドル、Fable 5.2 26.54 ドル	4.00 ドル[3]、26.54 ドル（二次情報）[8]	一致
AA の所要時間	13.9 分（中央値 12.1 分）、初回応答 209.8 秒	13.7 分（中央値 12.3 分）、197.65 秒[3]	変動
AA の出力トークン	6,580 万（中央値 6,510 万）	同じ[3]	一致
AA の推論努力別	xhigh 62、high 60、medium 57、low 46	同じ[3]	一致
Epoch ECI	174、241 モデル中同点 1 位	同じ[9]	一致
Epoch の個別項目	FrontierMath Tier 4 100%、MirrorCode 68%、SimpleQA 79%など	同じ[9]	一致
Vals Index	67.66%、42 モデル中 8 位	同じ[12,13]。公開当日は 7 位[14]	一致
Vals の分野別	Science 1 位、金融・法務 23 位、税務 56 位など	同じ[13]	一致

項目	動画の記述	原典の値（基準日時点）	判定
Agent Arena	スコア 0.152、68 モデル中 6 位	旧スコア 15.2 で 6 位（二次情報）[23]。11 月 18 日に新指数を導入[15]	一致（旧指標）
WebDev	1,781 点、146 モデル中 3 位	発表値 1,781 点[22]、現在の表示 1,797 点。3 位は同じ[21]	変動
OpenAI 掲載の AA 値	Fable 5.2 64.6、Astra 62.1	OpenAI のページ本文からは抽出できず[2]。64.6 は二次情報と一致[28]	一部未確認

注：AA は Artificial Analysis の略。WebDev の母数（146 モデル）は原典で確認できていない。

5. 評価の読み方（分析）

本章は筆者の整理であり、事実の記述ではない。

- **Claude Fable 5.2**：総合指標の多くで首位に立つ。弱点は最上位設定のコストで、Artificial Analysis の 1 タスクコストは Astra の約 6.6 倍である。
- **GPT-6.2 Astra**：科学、数学、抽象推論と費用対効果が強みである。開放型の実務タスク、金融・法務・税務、事実知識、コーディングエージェントの伸びには課題が残る。
- **Gemini 4 系**：Vals Index の上位 7 つのうち 4 つを占め、Arena の Work Index と Chat Index でも 3 位から 5 位に入る。単価の低さが目立つ。
- **Claude Opus 5.5 と Sonnet 5.5**：Vals と Arena で上位にあり、フロンティアは 3 社の複数モデルが僅差で並ぶ状況になっている。

評価結果を使う際は、次の 3 点に注意が必要である。

- **測り方で結論が変わる。** ARC-AGI-3 は、ハーネスの違いで 100%と 70.3%に分かれた[24]。法務でも、OpenAI が掲載した Harvey LAB-AA v1.1 では Astra が Claude Fable 5.2 の 2 倍を超える合格率で最高値とされる一方[2]、Vals 側の設定（HLAB-Vals-1.1）では 16.25%で 12 位である[13]。
- **指数の同点は実務での同等を意味しない。** ECI で並んだ 2 モデルの間に、Epoch の実務タスク評価では 17 ポイントの差が出た[10]。
- **公開直後の数値は動く。** WebDev のスコアは発表から 2 日ほどで 16 点上がり、Vals Index の順位も 7 位から 8 位に変わった。利用者側の第一印象も数値とは別で、VentureBeat は公開直後のパワーユーザーの否定的な反応を報じている[27]。

自社の業務に近い評価を選び、推論努力の設定とコストを揃えて比べることが、順位表そのものより有用である。

6. 確認できなかった点

- 動画ページ本体と概要欄は取得できなかった（アクセス制限）。動画が挙げる出典リストとの突合は行っていない[1]。
- Arena の各ページは最新版を直接読めなかった。検索結果に収録されたページ内容（取得から約 1 日前）と同社の X 投稿で確認しており、信頼区間と投票数は未確認である[17,19,21]。
- 総合の Agent Arena Index における GPT-6.2 Astra の順位は確認できていない。
- OpenAI の発表ページに掲載されたグラフ内の数値（Artificial Analysis の 64.6 と 62.1 など）は、本文の抽出では読み取れなかった[2]。
- Gemini 4 Neon の Artificial Analysis と Epoch での値は確認できていない。
- X（旧 Twitter）の投稿は、検索結果の抜粋またはミラーサイト経由での確認である。

参考文献

[1] YouTube 動画（チャンネル名未確認）「GPT-6.2 Astra は最強？Epoch AI は同点 1 位、Vals は科学 1 位、Agent Arena は 6 位！第三者評価を比較」 https://www.youtube.com/watch?v=HVtq0iLKE-4（2026 年 11 月 20 日閲覧） ※文字起こしで内容を確認。動画ページ本体と概要欄は取得できず。

[2] OpenAI「Introducing GPT-6.2 Astra and Sol」2026 年 11 月 17 日 https://openai.com/index/introducing-gpt-6-2-astra-sol/（2026 年 11 月 20 日閲覧） ※本文を取得して確認（グラフ内の数値は抽出できず）。

[3] Artificial Analysis「GPT-6.2 Astra: Intelligence, Performance & Price Analysis」 https://artificialanalysis.ai/models/gpt-6-2-astra（2026 年 11 月 20 日閲覧） ※本文を取得して確認。

[4] Artificial Analysis「GPT-6.2 Astra vs Claude Fable 5.2: Release Comparison」 https://artificialanalysis.ai/models/comparisons/gpt-6-2-astra-vs-claude-fable-5-2（2026 年 11 月 20 日閲覧） ※検索結果の抜粋で確認（本文未取得）。

[5] Artificial Analysis（X 投稿）2026 年 11 月 17 日 https://x.com/ArtificialAnlys/status/2122750634210529569（2026 年 11 月 20 日閲覧） ※検索結果の抜粋で確認（本文未取得）。

[6] OfficeChai「GPT 6.2 Astra Places Second On Artificial Analysis Intelligence Index, Edges Out Gemini 4 Carbon By 2 Points」2026 年 11 月 https://officechai.com/ai/gpt-6-2-astra-artificial-analysis-intelligence-index/（2026 年 11 月 20 日閲覧） ※二次情報。本文を取得して確認。

[7] The Decoder「OpenAI launches GPT-6.2 Astra and Sol with lower API prices and some benchmark gains」2026 年 11 月 https://the-decoder.com/openai-launches-gpt-6-2-astra-and-sol-with-lower-api-prices-and-some-benchmark-gains/（2026 年 11 月 20 日閲覧） ※二次情報。本文を取得して確認。

[8] Stephan Miller「Claude Fable 5.2 Hands Its Riskiest Questions to Opus 4.8」2026 年 11 月 https://www.stephanmiller.com/model-buzz-roundup-week-of-1104/（2026 年 11 月 20 日閲覧） ※二次情報。検索結果の抜粋で確認（本文未取得）。

[9] Epoch AI「GPT-6.2 Astra」（モデルページ） https://epoch.ai/models/gpt-6-2-astra（2026 年 11 月 20 日閲覧） ※本文を取得して確認。

[10] Epoch AI「How good are Fable 5.2 and GPT-6.2 Astra at real work?」2026 年 11 月 17 日 https://epoch.ai/publications/how-good-are-fable-5-2-and-gpt-6-2-astra-at-real-work（2026 年 11 月 20 日閲覧） ※検索結果の抜粋で確認（本文未取得）。

[11] Epoch AI（X 投稿スレッド）2026 年 11 月 17 日 https://x.com/EpochAIResearch/status/2122762380463735010（2026 年 11 月 20 日閲覧） ※ミラーサイト（twitter-thread.com）経由で内容を確認。

[12] Vals AI「Vals Index Leaderboard and Methodology」2026 年 11 月 18 日更新 https://www.vals.ai/benchmarks/vals_index（2026 年 11 月 20 日閲覧） ※本文を取得して確認。

[13] Vals AI「GPT-6.2 Astra Benchmarks, Cost and Capabilities」 https://www.vals.ai/models/openai_gpt-6.2-astra（2026 年 11 月 20 日閲覧） ※本文を取得して確認。

[14] Vals AI（X 投稿）2026 年 11 月 17 日 https://x.com/ValsAI/status/2122750572533326298（2026 年 11 月 20 日閲覧） ※検索結果の抜粋で確認（本文未取得）。

[15] Arena（X 投稿、Agent Arena Index の導入）2026 年 11 月 18 日 https://x.com/arena/status/2123126646983332177（2026 年 11 月 20 日閲覧） ※検索結果の抜粋で確認（本文未取得）。

[16] Crypto Briefing「Arena launches Agent Arena Index to rank how AI agents perform in real work」2026 年 11 月 https://cryptobriefing.com/arena-launches-agent-arena-index/（2026 年 11 月 20 日閲覧） ※二次情報。検索結果の抜粋で確認（本文未取得）。

[17] Arena「Agent Arena Leaderboard」 https://arena.ai/leaderboard/agent（2026 年 11 月 20 日閲覧） ※検索結果の抜粋（11 月 13 日時点の表示）で確認。直接取得では列名を読めず。

[18] Arena「Coding Index」 https://arena.ai/leaderboard/agent/code（2026 年 11 月 20 日閲覧） ※検索結果の抜粋（取得から約 1 日前の表示）で確認。

[19] Arena「Work Index」 https://arena.ai/leaderboard/agent/work（2026 年 11 月 20 日閲覧） ※検索結果の抜粋（約 1 日前の表示）で確認。直接取得は失敗。

[20] Arena「Chat Index」 https://arena.ai/leaderboard/agent/chat（2026 年 11 月 20 日閲覧） ※検索結果の抜粋（約 1 日前の表示）で確認。

[21] Arena「Code Arena | WebDev」 https://arena.ai/leaderboard/code/webdev（2026 年 11 月 20 日閲覧） ※検索結果の抜粋（約 1 日前の表示）で確認。直接取得では古い版が返り、信頼区間・投票数は未確認。

[22] Arena（X 投稿、WebDev の結果）2026 年 11 月 18 日
https://x.com/arena/status/2123112696283582603（2026 年 11 月 20 日閲覧） ※検索結果の抜粋で確認（本文未取得）。

[23] BenchLeader「LMArena Agent leaderboard (Nov 2026)」2026 年 11 月 19 日時点
https://www.benchleader.com/benchmarks/lmarena_agent（2026 年 11 月 20 日閲覧） ※二次情報（集計サイト）。検索結果の抜粋で確認（本文未取得）。

[24] ARC Prize「GPT-6.2 - ARC-AGI Results」 https://arcprize.org/results/openai-gpt-6-2（2026 年 11 月 20 日閲覧） ※本文の前半約 3 分の 2 を取得して確認（後半は未読）。

[25] ARC Prize（X 投稿）2026 年 11 月 17 日 https://x.com/arcprize/status/2122751233442156678（2026 年 11 月 20 日閲覧） ※検索結果の抜粋で確認（本文未取得）。

[26] FourWeekMBA「OpenAI's GPT-6.2 Astra Is First to 100% on ARC-AGI-3: $4.1K」2026 年 11 月
https://fourweekmba.com/ai-openais-gpt-6-2-astra-is-first-to-100-on-arc-agi-3-4-1k/（2026 年 11 月 20 日閲覧） ※二次情報。検索結果の抜粋で確認（本文未取得）。

[27] VentureBeat「'Essentially unusable': OpenAI launches GPT-6.2 Astra and Sol to negative reception from AI power users」2026 年 11 月 https://venturebeat.com/technology/essentially-unusable-openai-launches-gpt-6-2-astra-and-sol-to-negative-reception-from-ai-power-users（2026 年 11 月 20 日閲覧） ※検索結果の抜粋（見出しと価格表）で確認（本文未取得）。

[28] LLM Learner「AA Intelligence Index v5.0.1 Leaderboard」2026 年 11 月 12 日更新
https://llmlearner.com/rankings/aa-intelligence-index-v5-0（2026 年 11 月 20 日閲覧） ※二次情報（集計サイト）。検索結果の抜粋で確認（本文未取得）。