Artificial Analysis Intelligence Index v4.2

改定の分析

― 改定内容・ランキング変動・反響と実務への示唆 ―

2026 年 9 月 6 日

Claude Fable 5.1

凡例：本稿の「確認済み事実」は、Artificial Analysis 公式記事・方法論ページ・公式 X 投稿などの一次ソース、および出典を明示した二次報道に基づく。「（分析）」と付した箇所は筆者の推論である。文中の上付き番号は文末「参考文献」の番号に対応し、検証できなかった URL・出典には※を付した。

1. 要旨

Artificial Analysis（以下「AA」）は 2026 年 9 月 4 日、LLM 総合評価指標「Artificial Analysis Intelligence Index」の v4.2 を公表した[1,2]。AA 自身が「v5 リリースの要素を前倒しした暫定的な中間更新（interim update）」と位置づける改定で、実務型エージェントおよび実文書処理タスクの比重を一段高めた[1]。具体的には、自社構築の非公開テスト「AA-Briefcase」と Surge AI の長文 PDF 読解ベンチマーク「GDP.pdf」を追加し、飽和した「GPQA Diamond」を除外、非公開（held-out）テストの重み比率を 20%から 40%へ倍増させた[1,3]。改定後の首位は Anthropic「Claude Fable 5.1」（57 点）、2 位は OpenAI「GPT-6 Astra」（55 点、前世代 GPT-5.6 Sol 比＋4 点）である[4,5]。

改定の直接の引き金は、9 月 3 日に発表された GPT-6 Astra が旧指数（v4.1.1）で前世代 Sol と同点（61 点）にとどまり、「期待外れ」と報じられた騒動である[6,7]。その翌日に v4.2 が公表されたため、「市場のコンセンサスに合わせて指数を後付けで調整したのではないか」という恣意性・タイミングへの批判が Hacker News 等で提起された[8,9]。他方、飽和ベンチマークの除去と非公開化によるコンタミネーション対策は標準的な保守作業として妥当だという肯定的評価も強く、両論が併存している[9,10]。

知財・調達実務への含意（分析）：単一の合成スコアはモデル選定の一次フィルタにとどめ、自社タスクでの精度・レイテンシ・コスト・ライセンス条件を個別に検証すべきである。v4.2 が重視するエージェント処理と長文実務文書（契約書・技術文書・財務書類）処理は、先行技術

調査、明細書ドラフト、長文の審査対応など知財業務と親和性が高い評価軸である。ただし非公開テスト依存の増加は、「ゲーミング耐性」と「外部からの検証可能性・独立再現性」のトレードオフを伴う点に留意が必要である。

2. 改定内容（確認済み事実）

2.1 公表日と位置づけ

公式記事のメタデータ上の公表日は 2026 年 9 月 4 日であり、同日に公式 X 上でも告知されている [1,2]。AA は記事中で、「v5 の要素を数か月かけて計画・構築してきた。1 月に Index v4 を公開してから 8 か月が経過し、直近数週間でフロンティアが急速に動いたため、即時の中間更新を出すことが重要だと判断した」旨を説明している [1]。

2.2 評価ベンチマークの変更

（a）追加①：AA-Briefcase。AA が自社開発した非公開（held-out）テストセットで、業界専門家が構築した「複数週にわたる実務プロジェクト」を模した長期エージェント型ナレッジワーク評価である。ルーブリック採点とペアワイズ採点を組み合わせ、タスク達成度・分析品質・プレゼンテーション品質を測る。指数内で最大の 15%のウェイトを持つ [1,3]。

（b）追加②：GDP.pdf（Surge AI 作成）。100 件の PDF、10 ドメイン、4,592 ページ、1,275 の専門家作成基準（atomic criteria）にわたる長文専門文書の読解ベンチマークで、全基準を満たした場合のみ加点する「All-pass Rate」を主指標とする [11,12]。AA の実装では GPT-5.6 Luna（Medium）が各基準を独立に採点する。Surge 本家は Gemini 3.5 Flash を採点に用いており、両者のスコアは直接比較できない [11,3]。Surge の元論文は、2026 年 7 月時点で評価した 17 のフロンティアモデルのうち、厳格指標で 3 分の 1 を超える項目に合格したモデルは皆無であったと報告しており、難度の極めて高いベンチマークである [12]。ウェイトは 10%[3]。

（c）除外：GPQA Diamond。フロンティアモデルがほぼ天井に達して識別力を失った「飽和」を理由に、レガシー評価へ移された [1]。

（d）採点基盤の改善。AA-LCR を v1.1 に更新（採点用システムプロンプトの追加、解答キー16 問の修正）、GDPval-AA v2 および AA-Briefcase の Elo を再アンカリング、SciCode の採点サンドボックスを堅牢化（実行が遅いが正しいコードを失敗扱いしない）した [1]。

2.3　重み構成

v4.2 は全 10 評価を 4 カテゴリに配分する。方法論ページに基づく構成は表 1 のとおりである[3]。

表 1　Intelligence Index v4.2 の評価構成とウェイト

カテゴリ	評価	ウェイト	概要
Agents（30%）	AA-Briefcase	15%	長期エージェント型ナレッジワーク（AA 自社・非公開）
	GDPval-AA v2	10%	実務タスク遂行（Elo、人間性能＝1000 に再アンカリング）
	τ^3-Bench Banking	5%	銀行業務ドメインのツール利用エージェント評価
Coding（20%）	Terminal-Bench v2.1	10%	ターミナル操作型のエージェント・コーディング
	SciCode	10%	科学計算コード生成（採点サンドボックスを堅牢化）
General（30%）	AA-Omniscience	15%	知識精度 10%＋非ハルシネーション率 5%（非公開）
	GDP.pdf	10%	長文専門 PDF 読解（All-pass Rate、Surge AI 作成）
	AA-LCR v1.1	5%	長文脈読解（v1.0 とは直接比較不可）
Scientific Reasoning（20%）	HLE	10%	Humanity's Last Exam
	CritPt	10%	物理学の研究レベル問題（解答は非公開）

2.4　非公開テストの倍増

AA は「指数ウェイトの 40%が非公開の held-out テストセットとなり、v4.1 の 2 倍になった。held-out データには AA-Briefcase、AA-Omniscience、および CritPt の解答が含まれる」と明記し[1]、OfficeChai も 20%から 40%への倍増を確認している[5]。AA はその目的を「ラボが公開テストに最適化して指数をゲーミングする余地の削減」とし、v5 で比率をさらに引き上げると予告している[1]。

3. バージョン系譜（v4.0 → v4.1 → v4.1.1 → v4.2）

- v4.0（2026 年 1 月）：4 カテゴリ均等 25%。AA-Omniscience・GDPval-AA・CritPt を追加し、MMLU-Pro・AIME 2025・LiveCodeBench を除外した[13,14]。

- v4.1（2026 年 6 月）：エージェント型ワークロードへのシフト。Terminal-Bench Hard を Terminal-Bench 2.1 へ、τ^2-Bench Telecom を τ^3-Bench Banking へ更新し、GDPval-AA を v2 へ（Elo を人間性能＝1000 に再アンカリング、フロンティアモデル判定パネルの導入、ターン上限を 100 から 250 へ）。IFBench を飽和により除外。当時のウェイトは Agents 34%／Coding 24%／Scientific 24%／General 18%であった [15]。
- v4.1.1：グレーダーモデルと τ^3-Banking を更新する小改定。大半のモデルのスコア変動は 1 点未満（最大の上昇は Muse Spark 1.2 の＋2.7 点）[16]。
- v4.2（2026 年 9 月 4 日）：本稿の対象 [1]。

なお、バージョン間でスコアは直接比較できない。AA 自身も AA-LCR v1.1 について「v1.0 と直接比較不可」と明記している [1,3]。

4. モデルランキング・スコアの変動

4.1 上位モデル（プロプライエタリ）

v4.2 公表時点の上位は表 2 のとおりである [4,5]。OfficeChai は、Anthropic が上位 4 枠中 3 枠を占めると報じた [5]。ラボ別の序列は、Anthropic＞OpenAI＞Meta＞SpaceXAI＞Moonshot／Kimi＞Z.AI＞Google である [4,9]。

表 2　Intelligence Index v4.2 上位モデル（プロプライエタリ）

順位	モデル（推論設定）	スコア	開発元
1	Claude Fable 5.1（Adaptive Reasoning, Max Effort, fallback 付）	57	Anthropic
2	GPT-6 Astra（max）	55	OpenAI
3	GPT-6 Astra（xhigh）	54	OpenAI
3	Claude Opus 5（max）	54	Anthropic
5	Claude Fable 5（fallback 付）	53	Anthropic

4.2 有利になった主体・不利になった主体

OpenAI（GPT-6 Astra）が最大の受益者である。旧 v4.1.1 では前世代 Sol と同点（61 点）で「頭打ち」と報じられたが [7,6]、v4.2 では Sol 比＋4 点となった。AA-Briefcase で約＋85 Elo、GDP.pdf では全モデル中首位（Astra 33.2%＞Sol 28.2%＞Fable 5.1 26.2%）であり [11,17]、Astra が得意とするエージェント処理・長文処理が加点される構成となった。トークン効率でも Astra

がフロンティアを支配している[9,18]。

一方、Google（Gemini 系）はラボ別で最下位グループにとどまり、中国系ラボおよびオープンウェイト勢は絶対順位で中位以下である[4,9]。

4.3 オープンウェイト対クローズド、米国対中国

オープンウェイトモデルの上位は表 3 のとおりで、首位は Kimi K3（max、50 点）である[19,4]。トップのクローズドモデル（Claude Fable 5.1、57 点）とトップのオープンモデル（Kimi K3、50 点）の差は 7 点（2 位 GPT-6 Astra〔55 点〕との差は 5 点）である。

表 3　Intelligence Index v4.2 上位モデル（オープンウェイト）

順位	モデル（推論設定）	スコア	開発元
1	Kimi K3（max）	50	Moonshot AI（中国）
2	GLM-5.3（max）	49	Z.AI／智譜（中国）
3	Qwen3.8 2.4T A95B	47	Alibaba（中国）
4	GLM-5.3-Flash	46	Z.AI／智譜（中国）
5	DeepSeek V4 Pro	42	DeepSeek（中国）
6	MiniMax-M3	36	MiniMax（中国）

中国系ラボは、Moonshot／Kimi が 5 位、Z.AI（GLM）が 6 位で、Google（7 位）より上位にあるものの、米国 4 ラボ（Anthropic／OpenAI／Meta／SpaceXAI）の後塵を拝している[4]。もっとも、Z.AI はコスト対性能（Cost per Task）のパレートフロンティアを共有する 4 ラボ（Anthropic・OpenAI・Meta・Z.AI）の一つであり、コスト効率面での存在感は大きい[9,4]。競争軸が「米中」から「オープン対クローズド」へ移りつつあるとの論調も見られる[20]。

5. 反響・評判

5.1 メディア報道

The Decoder は「GPT-6 Astra のスコアが懐疑を呼んだ後に AA が指数を刷新した」と報じ、改定は批判への応答である可能性が高い（likely in response to criticism）と分析した[6]。OfficeChai は改定を「Rejigs（組み替え）」と表現し[5]、explainx.ai、WinBuzzer 等が詳報している[9,10]。日本語圏の主要メディア（日経、ITmedia、GIGAZINE 等）による v4.2 固有の報道は、本稿執筆時点では確認できなかった。

5.2　批判的評価

（a）恣意性・タイミング。Hacker News では、あるユーザーが「Astra が Sol と同点なのは馬鹿げていると気づき、皆の期待に合うよう指数を急いで更新した。旧指数は明らかに悪かった（Astra は Sol より明確に上）が、こう弄るのも非科学的だ」と指摘した[8]。ほかにも「どのグラフが v4.1 と v4.2 のどちらの数値か分からない」「なぜ ARC-AGI-3 が指数に入らないのか」「旧版と横並びで比較できる形で出すべきだ」など、透明性を求める声が上がった[8,9]。

（b）体感との乖離。「AA のランキングは自分や多くの人の実体験と合わない」「Opus 5 がしばらく首位だったが、体感では Opus 4.8 と大差なく、Fable 5 より劣る」といった声もある[8]。

5.3　肯定的評価

飽和した GPQA Diamond の除去は、MMLU や HumanEval の退役と同様の業界標準の保守作業であり、非公開評価の追加と held-out 比率の引き上げはコンタミネーション対策として妥当と評価されている[9]。また、方法論・プロンプト・採点基準を詳細に開示している点は、「説明のない非公開リーダーボードより透明である」との評価がある[10]。

5.4　開発企業の反応

本稿執筆時点で、Anthropic、OpenAI、Google DeepMind、xAI（SpaceXAI）、中国系ラボによる v4.2 への公式な言及・反論は確認できていない（検証未達）。

5.5　改定理由の読み方 — 確認済み事実と分析の区別

確認済み事実：（a）改定日は GPT-6 Astra 発表（9 月 3 日）とその「期待外れ」報道の翌日である[6]。（b）AA は「v5 前倒しの中間更新」「フロンティアが急速に動いたため」と公式に説明している[1]。（c）新評価は Astra が得意なエージェント・長文処理を重視し、結果として Astra は 2 位に浮上、GDP.pdf では首位となった[11,4]。

分析（留保付き）：タイミングから「コンセンサス追認のための後付け調整」という疑義が生じるのは自然だが、これを断定する証拠はない。explainx.ai が整理するように、「改定は方法論上正しく、かつタイミング上も好都合だった」という二つの読みは両立し得る。AA の公開文書はどちらが支配的かを解決しておらず、未決着の論点として扱うべきである[9]。

6. 他のベンチマーク・リーダーボードとの関係

6.1 Epoch AI との評価不一致

複数の二次報道は、Epoch AI が GPT-6 Astra を「267 モデル中 1 位（ECI 169 点、50 超のベンチマークを統合）」と評価したと伝えている [21]。Epoch AI 公式の ECI 解説は「50 超の異なるベンチマークのスコアを用いる合成指標」と説明するが、169 点・267 モデル・Astra 1 位という具体値は Epoch の一次文書では直接確認できず、公式 X 投稿と二次報道に依拠している点に留保が必要である [22]。AA は（旧指数において）同モデルを前世代並みと評価しており、両者の結論は正反対であった。The Decoder は「独立系 2 ラボが多数のテストを一つの総合スコアにまとめながら、正反対の結論に至った」と対比している [21]。

6.2 ARC-AGI-3

GPT-6 Astra は ARC-AGI-3 で人間平均を上回る効率を示し、ARC Prize を主宰する François Chollet は自身の AGI 予測を前倒しした（進捗は想定の「2 倍速」）[21]。ただし AA の指数に ARC-AGI-3 は含まれておらず、この点はコミュニティでも指摘された [8]。

6.3 コンタミネーション対策の文脈

AA の非公開比率 40%化は、公開ベンチマーク汚染（学習データへの混入により「暗記」が「推論」に偽装される問題）への業界的対応の一環として位置づけられる。関連事例として、Scale AI の GSM1k 論文は査読版で「最大 8%の精度低下と、複数モデル系統での体系的な過学習の証拠」を報告している（Scale 公式ブログの要約では「最大 13%」と表現され、AA 関連の二次記事もこの 13%を引用している）[23,9]。このほか、リーダーボード上で同一モデルのスコアが大きく乖離した事例や、秘匿計算環境内での二重盲検評価の試みなども二次情報として言及されているが、本稿では詳細を検証できていない。

6.4 技術的な注意点

- 指数は主に英語・テキストベースの評価である。マルチモーダル性能や多言語（日本語を含む）性能は別指数（Multilingual Index 等、Global-MMLU-Lite ベース）で評価され、Intelligence Index には含まれない [3,24]。
- GDPval-AA v2 および AA-Briefcase は、3 ラボのフロンティア LLM（Claude Opus 4.8、GPT-5.5、Gemini 3.1 Pro Preview 等）を審判パネルとする LLM-as-judge 方式である。GDP.pdf・AA-LCR・AA-Omniscience の採点には GPT-5.6 Luna が用いられる [3]。審判に

LLM を用いる構成への懸念（同族モデルへのバイアス等）は残るが、AA は複数審判パネル化によりバイアス低減を図っている（分析）。

- 統計的信頼性について、AA は「一部モデルで 10 回超の反復実験に基づき、指数全体の95%信頼区間は±1%未満」と推定する一方、「個別評価はより広い信頼区間を持ち得る」とも明記している[3]。
- どのモデルが v4.2 で再採点済みかは公開資料から明確でなく、旧 v4.1.1 由来の数値が混在する横断比較チャートの読解には注意が必要である[8,9]。

7. 影響の分析

7.1 企業・投資家・政策における参照

AA の指数・価格データは、米中 AI 競争を論じる政策文書（USCC、CSIS、Brookings、Stanford AI Index 関連）においてコスト対性能の根拠として参照されている。特に「中国のオープンウェイトモデルが低コストで性能ギャップを縮めている」という文脈で頻用され、例えば USCC の報告書は、Kimi K2.5 と GPT-5.2 が同じ指数 47 点ながら Kimi が 4 倍安いと指摘している[25]。実務面では、「知能指数は僅差でもコストが数倍違う」ため意図的に安価なモデルを選ぶ現場もあると日本語の解説も指摘する[26]。指数の高いモデルほど推論トークンが増え、コストとレイテンシが悪化する傾向は、調達設計上の要注意点である（分析）。

7.2 日本・日本語 LLM への含意

Intelligence Index は英語・テキスト中心であるため日本語性能の評価には直接使えず、日本語圏では Nejumi LLM リーダーボード、ELYZA-tasks-100、Hugging Face Open LLM Leaderboard 等が併用されている[27]。v4.2 が重視するエージェント処理・長文実務文書処理は、規制文書・技術文書・契約書の多い日本企業のユースケースと親和性が高い一方、AA の指数上位に日本発モデルは現れておらず、グローバル指標上の可視性は限定的である（分析）。

8. 実務への示唆（推奨事項）

1. 一次情報で最新スコアを直接確認する。モデル選定の意思決定前に、AA サイト上のライブ順位（v4.2）と、対象モデルが「v4.2 で再採点済みか」を確認する。旧 v4.1.1 由来の数値が混在する比較図には注意する。総合スコアの差が約 5 点未満であれば「誤差圏内」とみなし、自社タスクでの実測で判断することを目安とする（分析）[4,3]。

2. 指数は一次フィルタに限定し、業務別サブスコアを見る。知財業務であれば、長文文書処理は GDP.pdf、長期エージェント作業は AA-Briefcase／GDPval-AA v2、長文脈読解は AA-LCR のサブスコアを個別に確認する。総合 1 位（Claude Fable 5.1）と GDP.pdf 1 位（GPT-6 Astra）が異なるように、用途別に最適モデルは変わる [11,4]。
3. コスト対性能（Cost per Task）で評価する。トークン単価ではなく「完了タスクあたりコスト」で比較する。v4.2 では Anthropic・OpenAI・Meta・Z.AI がパレートフロンティアを共有しており、コスト重視であればオープンウェイト（Kimi K3・GLM-5.3、および低コストの Z.AI／DeepSeek 系）が有力候補となる [9,19]。
4. 非公開テスト依存の増加を監査リスクとして認識する。評価の独立再現性が重視される用途（医療・法務・金融等）では、非公開ベンチマーク 40%の指数を鵜呑みにせず、社内の私的評価セット（private eval）で二重検証する（分析）[1]。
5. 複数リーダーボードを併用する。AA 単独ではなく、Epoch AI、LMArena、SWE-bench、日本語なら Nejumi／ELYZA 等を併用する。Epoch AI と AA が Astra 評価で正反対の結論に至った事例が示すとおり、単一指標への依存は誤判断リスクが高い [21,27]。

判断を見直すべきシグナル（分析）：（a）AA が v5 を公表し非公開比率をさらに引き上げた場合、（b）対象モデルが v4.2 で再採点され順位が 5 点以上変動した場合、（c）開発各社が v4.2 に公式に反論・言及した場合（現時点では未確認）。

9. 留保事項

- 本稿のスコア・順位はすべて Artificial Analysis の自己申告値であり、独立に再現したものではない。指数構成・個別スコアは v5 公表前に再度変わり得る。
- 改定の「恣意性」は外部から証明も反証も不可能である。「コンセンサス追認」説と「通常の保守作業」説は両立し、どちらが支配的かは未決着であるため、断定を避けた。
- 開発企業（Anthropic・OpenAI・Google DeepMind・xAI・中国系ラボ）の v4.2 への公式反応は確認できなかった（検証未達）。反響は主にメディアおよび Hacker News 等のコミュニティの声に基づく。
- 重みの表記に資料間で食い違いがある。方法論ページは Agents 30%／Coding 20%／General 30%／Scientific 20%であるが、一部二次情報や旧 FAQ は「4 カテゴリ均等 25%」や v4.1 時代の「34／24／24／18」を記載しており、WinBuzzer もこの不整合を指摘している [3,10]。本稿は方法論ページ（v4.2）の数値を採用した。

- Epoch AI の GPT-6 Astra 評価値（ECI 169 点／267 モデル中 1 位）は二次報道と Epoch 公式 X 投稿に基づき、Epoch の一次文書では直接確認できていない [22,21]。
- GSM1k の精度低下幅は、査読版論文が「最大 8%」、Scale 公式ブログ要約が「最大 13%」と表記が異なる（AA 関連の二次記事は 13%を引用）[23]。
- 米国対中国の「国別」専用チャートが v4.2 に紐づく形で存在するかは検証できていない。AA には「Intelligence by Lab Over Time」ビューはあるが、国別内訳は確認できなかった。
- 日本語圏の大手メディア（日経・ITmedia・GIGAZINE 等）による v4.2 固有の報道は確認できなかった。日本語の言及は LLM 比較解説サイト（AI PICKS、Qiita 等）における一般的な Index 解説が中心である [26]。
- 一部の二次情報（OfficeChai、explainx.ai、WinBuzzer、The Decoder、AlphaSignal 等）は一次発表を要約したものであり、細部（例：GPT-6 Astra の旧指数スコア「61」対「61.2」）に丸めや表記差がある。

参考文献

※：URL または出典の存在を本稿執筆時点で直接検証できなかったもの。

1. Artificial Analysis「Announcing Artificial Analysis Intelligence Index v4.2」（2026 年 9 月 4 日） https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-2
2. Artificial Analysis（@ArtificialAnlys）公式 X 投稿「Announcing Artificial Analysis Intelligence Index v4.2. We are accelerating elements of our upcoming v5 release…」（2026 年 9 月 4 日） https://x.com/ArtificialAnlys/status/2096001986110099767
3. Artificial Analysis「Intelligence Benchmarking Methodology」（v4.2 対応版、2026 年 9 月時点） https://artificialanalysis.ai/methodology/intelligence-benchmarking
4. Artificial Analysis「Artificial Analysis Intelligence Index v4.2」（評価・リーダーボードページ、2026 年 9 月時点） https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index
5. OfficeChai「Artificial Analysis Rejigs Intelligence Index With v4.2, Fable 5.1 Tops, GPT-6 Astra Placed Second」（2026 年 9 月） https://officechai.com/ai/artificial-analysis-rejigs-intelligence-index-with-v4-2-fable-5-1-tops-gpt-6-astra-placed-second/
6. The Decoder「Artificial Analysis overhauls its Intelligence Index after GPT-6 Astra scoring drew skepticism」（2026 年 9 月 5 日） https://the-decoder.com/artificial-analysis-overhauls-its-intelligence-index-after-gpt-6-astra-scoring-drew-skepticism/
7. OfficeChai「GPT-6 Astra Scores A Disappointing 61 On Artificial Analysis Intelligence Index…」（2026 年 9 月） （URL 未確認） ※
8. Hacker News「GPT-6 Astra makes major gains in the Artificial Analysis Coding Agent Index」（コメントスレッド、2026 年 9 月） https://news.ycombinator.com/item?id=49556147
9. Yash Thakker「AA Intelligence Index v4.2: What Actually Changed」explainx.ai Blog（2026 年 9 月 5 日） https://www.explainx.ai/blog/artificial-analysis-intelligence-index-v4-2-september-2026
10. 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/
11. Artificial Analysis「GDP.pdf Benchmark Leaderboard」（評価ページ、2026 年 9 月時点） https://artificialanalysis.ai/evaluations/gdp-pdf
12. Surge AI「GDP.pdf」ベンチマーク論文（arXiv:2607.11192v3、2026 年 7 月） https://arxiv.org/abs/2607.11192 ※
13. Artificial Analysis（@ArtificialAnlys）公式 X 投稿「Announcing Intelligence Index v4.0: incorporating 3 new evaluations…」（2026 年 1 月）

https://x.com/ArtificialAnlys/status/2008570646897573931

14. Sebastian Raschka「Artificial Analysis Intelligence Index」（LLM Architecture Gallery）https://sebastianraschka.com/llm-architecture-gallery/aa-intelligence-index/

15. Artificial Analysis「Artificial Analysis Intelligence Index v4.1: a shift toward agentic workloads」（2026 年 6 月） https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-1

16. Artificial Analysis「Launching v4.1.1 of the Artificial Analysis Intelligence Index」（2026 年）https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-1-1

17. AlphaSignal（ニュースレター）GPT-6 Astra／Intelligence Index v4.2 関連速報（2026 年 9 月） （URL 未確認） ※

18. Artificial Analysis「Benchmarking GPT-6 Astra」（2026 年 9 月）https://artificialanalysis.ai/articles/benchmarking-gpt-6-astra ※

19. Artificial Analysis「Open Source LLMs」（オープンウェイトモデル比較ページ、2026 年 9 月時点） https://artificialanalysis.ai/models/open-source ※

20. Fortune「Has the AI race shifted from U.S. vs China to open vs closed?」（2026 年 8 月 4 日）https://fortune.com/2026/08/04/has-the-ai-race-shifted-from-u-s-vs-china-to-open-vs-closed/

21. The Decoder「Benchmarks disagree on GPT-6 Astra, but its human-beating efficiency on ARC-AGI-3 pulls Chollet's AGI forecast forward」（2026 年 9 月） https://the-decoder.com/benchmarks-disagree-on-gpt-6-astra-but-its-human-beating-efficiency-on-arc-agi-3-pulls-chollets-agi-forecast-forward/

22. Epoch AI「Epoch Capabilities Index（ECI）」解説ページ、および公式 X（@EpochAIResearch）投稿（2026 年 9 月） https://epoch.ai/ ※

23. Zhang, H. et al.（Scale AI）「A Careful Examination of Large Language Model Performance on Grade School Arithmetic」（GSM1k、arXiv:2405.00332） https://arxiv.org/abs/2405.00332

24. ProjectCOZY「The Artificial Analysis Index Explained: What It Does Measure」https://projectcozy.me/blog/artificial-analysis-index-explained

25. U.S.-China Economic and Security Review Commission（USCC）「Two Loops」、CSIS「What to Know About Chinese AI Models」、Brookings、Stanford HAI「AI Index 2026」関連資料（米中 AI 競争における Artificial Analysis データの参照事例） （URL 未確認） ※

26. AI PICKS「知能指数（Artificial Analysis Intelligence Index）とは？意味・使い方を解説」https://aipicks.jp/glossary/intelligence-index

27. Since2020（Data Driven Knowledgebase）「【2026 年最新】LLM 比較表・性能ランキング！LLM 比較サイト一覧」 https://since2020.jp/media/llm-ranking/