Artificial Analysis Intelligence Index v4.2 改定の全貌：評価体系の構造転換とAI産業への波紋

Gemini 3.8 Flash

指標改定の背景と基本思想の転換

AIモデルのベンチマーク評価プラットフォームであるArtificial Analysisは、旗艦総合指標「Artificial Analysis Intelligence Index」のバージョン4.2を公表した[1]。本改定は、2026年1月に展開されたv4およびその派生であるv4.1からわずかな期間で実施された異例の暫定改定（Interim Update）であり、水面下で8か月間開発が進められている次期メジャー改定「v5」の主要コンポーネントを前倒しで統合する構造をとっている[2]。

改定の直接的な契機となったのは、フロンティアモデルの急速な性能伸長に対し、既存の評価基準が弁別能を失った点にある[1]。とりわけOpenAIの「GPT-6 Astra」の初期評価において、旧版の指標では前世代のGPT-5.6 Solと同等水準のスコアが付与されたことが技術コミュニティで激しい議論を呼んだ[3]。当時、Epoch AIの総合評価（267モデル中首位、50超のベンチマークで169ポイント）やARC-AGI-3による外部検証ではAstraが飛躍的な推論能力の向上を示していたのに対し、Artificial Analysisの指標ではその進化が不可視化されていたため、ベンチマーク自体の陳腐化と妥当性に対する懐疑論が急速に浮上した[3]。

これを受け、Artificial Analysisはモデル提供企業（AIラボ）による公開テストセットへの過学習や特化チューニング（ベンチマークハッキング）を排除し、現実のエンタープライズ業務におけるタスク遂行能力を厳密に測定すべく、評価フレームワークの大幅な再構築を余儀なくされた[1]。

評価アーキテクチャの変更点

v4.2における改定の要点は、非公開（Held-out）テストセットの比重拡大、飽和した旧世代ベンチマークの除外、および複雑な実務ワークフローを模した新規ベンチマークの導入にある[1]。

評価構成と重み付けの再配分

本改定において、指標を構成する10種類のベンチマークは「Agents」「Coding」「General」「Scientific Reasoning」の4領域に再編された[2]。最大の特徴は、全体の重み付けの40%を非公開データセット（AA-Briefcase、AA-Omniscience、CritPtの解答群など）が占めている点であり、これはv4.1における比率から倍増している[1]。

評価カテゴリ（全体比重）	個別ベンチマーク	個別比重	テスト規模・形式	採点手法および特性
Agents (30%)	**AA-Briefcase**	15%	4シナリオ／91タスク	非公開実務環境、E2B Linuxサンドボックス、

				ルーブリック採点＋ペアワイズ Elo[2]
	GDPval-AA v2	10%	220タスク	44職種・9業界の実務タスク、ブラウジング・シェル操作を伴うペアワイズ Elo[6]
	τ^3-Banking	5%	97タスク	銀行業務対話シミュレーション、DB状態検証、pass@1[6]
Coding (20%)	**Terminal-Bench v2.1**	10%	89タスク	ターミナル操作・システム管理・データ処理、pass@1[6]
	SciCode	10%	288サブ問題	16科学分野の数値計算コード生成、単体テスト実行検証、pass@1[6]
General (30%)	**AA-Omniscience**	15%	6,000問	正解率（10%）と非幻覚率（5%）の複合、推測による誤答にペナルティ[6]
	GDP.pdf	10%	100タスク／4,592頁	10ドメインの長大PDF文書推論、1,275の原子基準による All-pass Rate[1]
	AA-LCR v1.1	5%	100タスク	1万〜10万トークンの長文脈推論、一致判定

				LLMによるpass@1[6]
Scientific Reasoning (20%)	**Humanity's Last Exam (HLE)**	10%	2,158問	数学・自然科学・人文科学の最難関専門問題、pass@1[6]
	CritPt	10%	70タスク	物理・数理科学の計算問題、公式判定サーバー検証、pass@1[6]

新設および除外されたベンチマークの特性

新規導入されたベンチマークの中で中核をなす「AA-Briefcase」は、単一の評価項目として最大の重み（15%）を持つ独自評価である[2]。業界の専門家によって設計された複数週にわたるナレッジワークプロジェクトを模しており、数千件の一次資料を参照しながら、スプレッドシート、プレゼンテーション資料、分析メモを作成させる[1]。インターネットアクセスを遮断したE2B Linuxサンドボックス環境下で最大500ターンの自律実行を行わせ、タスク達成度、分析の緻密さ、成果物の表現品質の3要素からEloレーティングを算出する[6]。

もう一つの主要な追加項目であるSurge AI開発の「GDP.pdf」（比重10%）は、10領域にわたる計4,592ページの長大PDF文書から、図表、脚注、例外条項を統合して単一ターンで解答を導く能力を検証する[1]。モデルの回答は専門家が作成した1,275件の原子判定基準（Atomic Criteria）と照合され、全条件を完全に満たした場合にのみ正解とみなす「All-pass Rate」を指標とすることで、曖昧な部分点を排した極めて厳格な評価体系を確立している[1]。

これら実務タスクの拡充と引き換えに、長年にわたり大学院レベルの科学推論のベンチマークとして君臨してきた「GPQA Diamond」は、最新フロンティアモデルによる正答率が天井に達し、モデル間の差異を検出する能力を失った（飽和した）と判断され、指標から除外された[1]。

採点インフラストラクチャの改修

指標の安定性を損なっていた技術的要因に対しても是正処置が講じられた[1]。長文脈推論を測定する「AA-LCR v1.1」では、判定用システムプロンプトの再設計と解答キーの曖昧さの排除が行われ、採点精度が改善された[1]。「GDPval-AA v2」および「AA-Briefcase」ではサンプリング手法の刷新とElo尺度の再アンカー（人間の専門家水準を1,000として固定）が実施され、新モデルの追加に伴うレーティングのブレを抑制している[1]。さらに「SciCode」においては実行サンドボックスの堅牢化が図られ、正解を出力しているものの処理時間を要するコードがタイムアウトによって誤って不正解と処理される欠陥が修正された[1]。

フロンティアモデルの序列変動と実力分析

評価軸が静的な知識問題から動的なエージェント作業および長文ドキュメント処理へとシフトした結果、主要モデルの序列と実務適性に明確な差異が表れている[1]。

モデル名	開発元	総合スコア（v4.2）	AA-Briefcase 評価	GDP.pdf (All-pass Rate)	運用効率・特性
Claude Fable 5.1	Anthropic	57	第1位（Opus 5と同列首位）[1]	26.2%（第3位）[1]	自律エージェントの長期文脈維持に秀でる[1]
GPT-6 Astra	OpenAI	55	第3位（Sol比 +85 Elo）[1]	33.2%（第1位独走）[1]	複雑PDF解析で突出、最高水準の出力トークン効率[1]
Gemini 3.8 Flash	Google	59（特定高設定時）[10]	中位グループ	19.0%[12]	コスト対知能のパレートフロンティアを更新[1]
GPT-5.6 Sol	OpenAI	51（推定）[1]	Astraの下位	28.2%（第2位）[1]	旧世代フロンティア、PDF解析では依然上位[1]

総合首位にはAnthropicの「Claude Fable 5.1」（スコア57）が立ち、わずか2ポイント差の第2位にOpenAIの「GPT-6 Astra」（スコア55）が続いている[1]。旧指標で並走状態にあった前身モデルGPT-5.6 Solに対してAstraが4ポイントの有意なリードを獲得したことで、Astraの進化が定量的に是認される形となった[1]。第3位にはMetaが位置し、SpaceXAI、Moonshot/Kimi、Z.AI（智譜AI）、Googleがそれに続いている[1]。

しかし、総合スコア以上に重要な洞察は、ドメイン別の適性分離にある[2]。Claude Fable 5.1およびClaude Opus 5は「AA-Briefcase」において他モデルを圧倒しており、数週間に及ぶ複雑な文脈管理、散在する指示の発見、整合性の取れた成果物生成といった長期的エージェント作業において強固な優位性を示している[1]。

これに対し、GPT-6 Astraは「GDP.pdf」において33.2%のAll-pass Rateを達成し、2位のGPT-5.6 Sol（28.2%）や3位のClaude Fable 5.1（26.2%）を引き離している[1]。数千ページ規模の文書から例外条項や微細な数値を漏れなく抽出・統合する単一ターンの精査作業では、Astraの注意機構と推論精度が突出していることが示された[2]。

運用経済性の観点では、タスクあたりコスト（Cost per Task）のパレートフロンティアをAnthropic、OpenAI、Meta、Z.AIの4社が分け合っており、特定企業の独占は生じていない[1]。特筆すべきはGPT-6 Astraの出力トークン効率であり、同一タスクを解く際に他社モデルに比べて大幅に少ない

トークン数で正解に達している[1]。冗長な推論ステップを抑えつつ解に直行する設計が、推論コストとレイテンシの双方を圧縮している[5]。また、Googleの「Gemini 3.8 Flash」は極めて安価なトークン単価を維持しながらスコア59を記録し、実務における高費用対効果の選択肢としてパレートフロンティア上に食い込んだ[1]。

開発者コミュニティおよび業界の反響

v4.2の公開に伴い、AIエンジニアや研究コミュニティでは、新指標の実効性を歓迎する声と、ベンチマーク運営体制に対する批判的な議論が交錯している[15]。

実務適用性を重視する開発者層からは、今回の改定方針に対して強い支持が寄せられている[15]。静的な学術テストから実務ワークフローへと重心を移したAA-BriefcaseやGDP.pdfは、実際の企業開発における体感性能と整合しているとの見解が支配的である[2]。加えて、比重15%を占める「AA-Omniscience」の採点設計が評価されている[6]。同ベンチマークは正答を加点する一方で、事実無根の幻覚（Hallucination）に対して重い減点を科し、未回答（知識の欠落を認めて回答を棄権すること）にはペナルティを与えない[6]。実運用において最も危険な「確信を持った誤答」を排除し、不確実性を率直に認める信頼性の高いモデルを適正に評価する仕組みは、エンタープライズのシステム統合において極めて実用的であるとみなされている[15]。

対照的に、改定プロセスと方法論の科学的妥当性をめぐっては厳しい批判が相次いでいる[15]。最大の論争点は、改定の契機がGPT-6 Astraの初期スコアに対するソーシャルメディア上での不満噴出であった点にある[3]。コミュニティの期待に反してAstraのスコアが伸びなかった直後に非公開データセットを用いて指標が修正され、結果としてAstraの順位が引き上げられた経緯に対し、「世間の評価や世論（Vibes）に合わせるよう、事後的にベンチマークの組み合わせや重み付けを改ざんしたのではないか」というP値ハッキング的な恣意性を疑う声が上がった[15]。

さらに、非公開テストセットの比重を40%に高めた措置は、評価の追試や第三者による監査を事実上不可能にするため、科学的厳密性と再現性を犠牲にしたブラックボックス化であるとの懸念も根強い[15]。また、依然として採用されている「Terminal-Bench v2.1」が既に最上位層で飽和しており、より高難度な新版への移行が見送られた点や、欠陥が指摘される「SciCode」の比重を10%に設定した点など、一部のサブベンチマークの選定基準にも疑問が呈されている[15]。

モデル選定とAIエコシステムに与える実務的影響

Intelligence Index v4.2の導入は、エンタープライズにおけるモデル選定の意思決定プロセスと、AIラボの開発方針に構造的な変化をもたらしている[2]。

従来のモデル調達では、リーダーボードの最上位に位置する単一の総合スコアを基準にモデルを採用するアプローチが一般的であった[2]。しかしv4.2が浮き彫りにしたのは、総合首位のClaude Fable 5.1であっても長大PDF解析では後塵を拝し、逆にGDP.pdfで圧倒的なGPT-6 Astraが長期エージェント作業ではFableに及ばないという、用途ごとの明確なトレードオフである[1]。この結果は、企業がAIシステムを構築する際、単一のフロンティアモデルに業務のすべてを委ねるのではなく、長大な一次資料の精査にはAstraを、複数ツールを連動させる長期的な業務自動化にはFableを配分するといった、タスク特性に応じた動的ルーティングの重要性を示している[2]。

同時に、モデル開発企業側に対する影響も小さくない[2]。公開された学術データセットを過剰に学習させることで見かけ上のスコアを釣り上げる戦略（いわゆるBenchmaxxing）は、非公開データセットが4割を占め、厳格なAll-pass基準や幻覚への減点を課す新体系の前では通用しなくなっている[2]。実

質的な汎用推論能力と長期の文脈保持力を伴わないモデルはスコアを維持できず、AIラボ各社はより実践的な問題解決能力の向上に研究開発リソースを集中せざるを得ない状況が生まれている[2]。

総括と次世代評価体系への展望

Artificial Analysis Intelligence Index v4.2は、フロンティアモデルの急速な進化によって旧来の評価基準が機能不全に陥るなか、実務適性を軸として測定基準の再定義を試みた戦略的転換点である[1]。非公開テストセットの大幅な拡充と、長大PDF解析や多段階エージェント作業への傾斜は、産業界の実需とベンチマークの乖離を埋める上で論理的な帰結といえる[1]。

しかしその代償として、評価プロセスの再現性低下や、結果に対する作為的なチューニング疑惑といった、民間の独立系評価プラットフォームが本質的に抱えるガバナンス上の課題が顕在化した[15]。今後数か月以内に順次ロールアウトが予告されている「バージョン5」においては、非公開比率のさらなる引き上げが計画されている[1]。次世代評価体系が業界における権威あるデファクトスタンダードとして定着できるか否かは、ゲーミングの防止と実務的有用性の向上を追求しつつ、評価手法の透明性と客観的再現性をいかに担保できるかにかかっている[4]。

引用文献

1. Announcing Artificial Analysis Intelligence Index v4.2, https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-2
2. AA Intelligence Index v4.2: Fable 5.1 Leads, Astra Efficient (2026), https://www.explainx.ai/blog/artificial-analysis-intelligence-index-v4-2-september-2026
3. Artificial Analysis overhauls its Intelligence Index after GPT-6 Astra scoring drew skepticism, https://the-decoder.com/artificial-analysis-overhauls-its-intelligence-index-after-gpt-6-astra-scoring-drew-skepticism/
4. Artificial Analysis v4.2 Doubles Private Test Weight to 40% | AI Weekly, https://aiweekly.co/alerts/artificial-analysis-v42-doubles-private-test-weight-to-40
5. Artificial Analysis Intelligence Index v4.2, https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index
6. https://artificialanalysis.ai/methodology/intelligence-benchmarking
7. AA-Briefcase: Agentic Knowledge Work Benchmark - Artificial Analysis, https://artificialanalysis.ai/evaluations/aa-briefcase
8. AI Model Evaluations - Artificial Analysis, https://artificialanalysis.ai/evaluations?skill=reasoning
9. Muse Spark 1.3 (max)とGPT-5.6 Luna (medium)：モデル比較, https://artificialanalysis.ai/ja/models/comparisons/muse-spark-1-3-vs-gpt-5-6-luna-medium
10. Google has released Gemini 3.8 Flash, its fourth Flash model in, https://artificialanalysis.ai/articles/gemini-3-8-flash
11. Gemini 3.8 Flash: Release Intelligence, Performance & Price, https://artificialanalysis.ai/ja/models/releases/gemini-3-8-flash
12. GDP.pdf Benchmark Leaderboard - Artificial Analysis,

https://artificialanalysis.ai/evaluations/gdp-pdf
13. GPT-6 weiter hinter Fable 5.1 nach Überarbeitung des "Intelligence Index", https://www.trendingtopics.eu/gpt-6-benchmarks-2/
14. GPT 6 Astra is not worth it : r/LocalLLaMA - Reddit, https://www.reddit.com/r/LocalLLaMA/comments/1w6ynpx/gpt_6_astra_is_not_worth_it/
15. Artificial Analysis Intelligence Index v4.2 | Hacker News, https://news.ycombinator.com/item?id=49571632
16. What the Artificial Analysis / GPT-6 Astra mess actually teaches us, https://www.reddit.com/r/LocalLLaMA/comments/1w89mdu/what_the_artificial_analysis_gpt6_astra_mess/