ChatGPT

Artificial Analysis Intelligence Index v4.2 深掘り調査

エグゼクティブサマリー

Artificial Analysis（AA）は2026年9月4日、LLMの総合能力指標 **Artificial Analysis Intelligence Index を v4.2 に改定**した。今回の変更は単なるベンチマーク追加ではなく、①実務型エージェント評価 **AA-Briefcase** の新設、②4,592ページの専門PDFを扱う **GDP.pdf** の組み込み、③飽和した **GPQA Diamond** の除外、④非公開・held-out テストの総ウェイトを **20%から40%へ倍増**、⑤AA-LCR、SciCode、GDPval-AAなどの採点基盤修正、⑥全評価項目の大規模なウェイト再配分、という複合的な測定方法の更新である。AA自身は目的を「より複雑・現実的なタスク」と「benchmark gaming（ベンチマークへの過適合）の抑止」と説明している。 [1]

結論から言えば、v4.2は「学術問題を解く知能」から「現実の資料を読み、成果物を作り、業務を遂行できる能力」への移行をさらに進めた改定であり、方向性は妥当である。 特に公開ベンチマークの飽和・汚染・過適合が深刻化するなか、private setを40%まで増やした点は測定上の大きな改善である。反面、第三者が完全再現できない比率も同時に40%になり、**anti-gaming と reproducibility のトレードオフ**が鮮明になった。 [2]

また、v4.2によってランキング自体の意味が変わった。直前のv4.1.1ではGPT-6 Astraと前世代GPT-5.6 Solがほぼ同等だったが、新指標ではAstraがSolを4ポイント上回り、Claude Fable 5.1に次ぐ位置になった。専門メディアThe Decoderはこのタイミングを、Astraの旧指標スコアに対する懐疑が出た直後の「overhaul」と報じた。ただし、**AA自身は「Astraへの批判が理由」とは述べておらず**、「大型モデルのリリース中は指標を安定させるため改定を控えていたが、フロンティアの進展が速すぎたため暫定更新した」と説明している。因果関係は区別すべきである。 [3]

日本の利用者にとって最重要の注意点は、**Intelligence Index本体は主として英語・テキストの評価**だということである。日本語能力は別のArtificial Analysis Multilingual Index（Global-MMLU-Lite、日本語を含む）で測られる。したがって、日本市場のモデル調達でv4.2総合点だけを使用するのは不十分である。 [4]

改定内容とv4.1.1比較

直前のv4.1.1は2026年8月6日の小規模パッチで、τ^3-Bankingのv1.0.1化と、HLE・AA-LCR・AA-Omniscienceのgraderを GPT-5.6 Luna (medium)に統一することが中心だった。AAによれば、多くのモデルの総合点変動は1ポイント未満だった。これに対してv4.2は評価セットそのものとウェイト構造を変更する、実質的なメジャー改定に近い。 [5]

評価領域・項目	v4.1 / v4.1.1	v4.2	具体的な変更と理由
Agents 合計	34%	**30%**	カテゴリ総量は4pt減。ただし構成を業務型へ再設計。 [6]
GDPval-AA v2	20%	**10%**	半減。単独の実務エージェント評価への依存を低減。 [7]

評価領域・項目	v4.1 / v4.1.1	v4.2	具体的な変更と理由
τ^3-Banking	14%	**5%**	9pt減。 [8]
AA-Briefcase	—	**15%**	新設。業界専門家が作る複数週・多数ファイルの知識労働タスク。private held-out。 [9]
Coding 合計	24%	**20%**	4pt減。 [8]
Terminal-Bench v2.1	16%	**10%**	比重低下。 [10]
SciCode	8%	**10%**	v1.0.1で再採点。遅いが正しいコードをsandbox timeout等で誤失格にしにくくした。 [1]
General 合計	18%	**30%**	最大のカテゴリ増加（+12pt）。文書理解・知識・幻覚耐性を大幅強化。 [6]
AA-Omniscience	12%（Accuracy 8 + Non-hallucination 4）	**15%（10 + 5）**	知識と「知らないことを捏造しない」能力の比重増。 [6]
GDP.pdf	—	**10%**	新設。100 PDF、10領域、計4,592ページ、1,275の専門家基準。 [11]
AA-LCR	6%	**5%、v1.1**	grading system prompt追加、answer keyの誤り・曖昧さ修正。 [1]
Scientific Reasoning 合計	24%	**20%**	4pt減。 [8]
HLE	12%	**10%**	若干減。 [12]
GPQA Diamond	6%	**0%**	**除外**。AAはフロンティアモデルで「saturated」と判断。 standaloneでは継続測定。 [13]
CritPt	6%	**10%**	物理・科学推論のprivate solutionを含む評価の比重増。 [14]
private / held-out比率	約20%	**40%**	倍増。モデル開発側による公開問題への直接・間接最適化を難しくする。 [15]

新評価が実際に測るもの

AA-Briefcase は91タスク・4シナリオで、Slackログ、スプレッドシート、PDF、インタビュー、調査資料、規格、取締役会資料、メールなど、現実のホワイトカラー業務に近い大量ファイルから成果物を作らせる。rubricによる成否に加え、分析品質・プレゼン品質をpairwise/Elo方式で評価する。最大500ターンのsandbox作業を許すため、従来の単発Q&Aとは性質が大きく異なる。 [16]

ただし、「multi-week project」という名称は若干割り引いて読む必要がある。公式方法論によれば、シナリオ内の各タスクは**現状では独立runで実行され、モデル自身の前週の成果物を記憶・継承しない**。後続タスクには全モデル共通のstandardized base-caseを与える場合がある。したがって、これは「数週間自律稼働するAI社員」の忠実なシミュレーションというより、**長期プロジェクトを構成する複雑な個別業務を切り出した評価**である。 [17]

GDP.pdf はSurge AIが2026年4月に公開した専門文書ベンチマークで、金融、医療、法務、研究、工学、建設、製造・サプライチェーン、保険、不動産、HRの10分野を含む。AA実装では100タスク・4,592ページ・1,275個のatomic criteriaを使い、headlineのAll-passは必要条件を**すべて**満たしたときだけ成功とする。Surgeはデータと再現用harnessも公開しているため、GDP.pdf自体はAA-Briefcaseとは違って公開評価である。[18]

さらにv4.2ではGDPval-AAとAA-BriefcaseのsamplingおよびEloのanchorを修正し、新モデル追加時のレーティング安定性を高めた。これは評価精度を上げる一方、**同じモデルの数字が「モデルが変わらなくても物差しの変更で変わる」可能性**を意味する。旧版スコアとの直接比較には注意が必要である。[15]

技術・規制・商業・研究への影響

技術・研究

短期的な最大の利点は、**benchmark saturation と contamination への耐性向上**である。AAはv4.1でも飽和したIFBenchを除外しており、v4.2では同じ理由でGPQA Diamondを外した。公開問題でフロンティアモデルが天井に近づけば、モデル間の差を測る「識別力」がなくなるため、難しい評価へ順次入れ替えるのは健全なベンチマーク運用である。[19]

private test 40%も、訓練データへの混入やbenchmark-specific tuningを抑えるという点では強い。ただし研究者から見ると、公開性との交換条件になる。AAは方法論・prompt・一部例題・Stirrup harnessを詳しく公開している一方、headline scoreの40%に関わる問題または解答を第三者が完全には検査できない。したがって今後は、**「再現できる公開ベンチマーク」と「汚染されにくい非公開ベンチマーク」のどこに信頼を置くか**がv5でも中心問題になる。[20]

また、AAはv4.2総合スコアについて一部モデルで10回超の反復実験から**95%信頼区間を±1%未満と推定**しているが、個々のevaluationはそれより広い場合があり、統計分析の詳細は今後公開するとしている。1〜2ポイント程度の順位差を「能力差」と断定するには慎重さが必要である。[21]

長期的には、LLM研究の最適化対象が単なる問題回答から、**ファイル生成、tool use、長文資料統合、作業品質、幻覚抑制**へ広がる可能性がある。一方、複数の異質なスコアを人為的なウェイトで一つの「Intelligence」に集約する根本問題は残る。v4.1を批判したRedditの詳細投稿は、総合値は統計的に自然発生した因子ではなく、**“The weights are an editorial choice”** と指摘していた。この批判はv4.2でも完全には解消していない。[22]

商業

企業のモデル選定には短期的にかなり大きな影響がある。GDPvalのウェイトが20→10%、τ^3-Bankingが14→5%になり、代わってBriefcase 15%、GDP.pdf 10%、Omniscience 15%が重くなったため、**同じモデルでも「何を重要視するか」が変わっただけで順位が変わり得る**。日本語の医療AI系ブログ「医知創造ラボ」も、9月4日の改定後は「旧指標の点数と現在の点数を混ぜて比較してはいけない」と明示している。[23]

一方で調達担当者にとっては、文書・表・チャート・メール・スプレッドシートから実際の成果物を作る評価が増えたことで、一般的な「賢さ」よりも**業務ROIに近いシグナル**になった。AAはコスト、時間、tokens per taskも併記しているため、「最高点」だけでなく価格性能を選べるという実務上の強みがある。[24]

規制・ガバナンス

規制への影響は**直接ではなく間接的**である。EU AI Actの現行統合テキストでは、systemic-risk GPAIの判定に「appropriate technical tools and methodologies, including indicators and benchmarks」を用いうるほ

か、systemic-riskモデルの提供者にはstate-of-the-artな標準プロトコル・ツールによるmodel evaluationとadversarial testingの実施・文書化が求められる。 [25]

しかしv4.2は主として**能力・業務性能**を測る指数で、サイバーセキュリティ、CBRN、欺瞞、モデル制御などのsystemic safety benchmarkではない。したがって、EU AI Act上のモデル評価を補助する能力情報にはなり得ても、**v4.2高得点＝規制適合、または安全**とは解釈できない。これはAAの評価構成とEU法上の要求を比較した分析上の推論である。 [26]

日本についても同様で、今回調べた公式・報道資料の範囲では、AAII v4.2を日本の行政機関が正式な規制・認証指標として採用した事実は確認できなかった。特に日本企業では、本指数が「primarily text-based, English-language」である以上、別途日本語評価・業界固有評価・安全性評価を組み合わせる必要がある。 [4]

外部反応と評判

リリース後約48時間という短い観測期間のため、反応は**専門AIメディア・開発者コミュニティが中心**で、査読論文や主要シンクタンクによるv4.2固有の分析はまだ確認できない。

層	確認できた反応	評価
専門ニュース	The Decoderは、Astraが旧AA指標で前世代並みに見えたことへの懐疑の直後にv4.2が登場したと報道。新指標ではAstraがSolを+4pt。 [27]	やや懐疑的。「指標が現実に追いついた」と見る一方、改定タイミングにも注目。
業界ブログ	ExplainXはprivate held-out、GDP.pdf、GPQA除外を中心に「What Actually Changed」と分析。 [28]	おおむね肯定的。anti-gamingと実務性を評価。
ベンチマーク業界	BenchLMはAA Indexを「display-only external reference」とし、自社総合ランキングには算入していない。単一公開task set/harnessとして扱えないことを理由に挙げる。 [29]	有力参考値ではあるが、独立検証可能な単一benchmarkとは区別。
LinkedIn	実務との整合を歓迎する一方、「成功/失敗だけでなく、決定・権限・実行過程まで評価すべき」との意見。 [30]	概ね好意的だが、agent governance評価を求める声。
Reddit	r/LocalLLaMAでは「v4.2は悪評への反応に見える」との疑念や、実体験とランキングが一致しないという意見。 [31]	指標の安定性・重み付けに対する懐疑が強め。
Hacker News	リリース翌日のスレッドは約140 points・約60 comments規模まで伸長。 [32]	技術者層からの関心は高い。
日本語ブログ	医知創造ラボは新版/旧版のスコア混同を警告し、総合点が正答率ではないことも強調。 [33]	比較的慎重・実務的。
日本語X	茶谷和弘氏は、非公開テスト比重倍増を「評価のゲーム化対策」として妥当な流れとコメント。別の日本語AI速報アカウントも「実践的で不正対策」と紹介。 [34]	private set導入への評価は肯定的。

層	確認できた反応	評価
note / Qiita	v4.0やv4.1.1、AAII一般の日本語解説は複数存在するが、9月6日時点でv4.2を深く論評するnote/Qiita記事は検索上ほぼ確認できず。 35	反応形成途上。
学術・think-tank	Epoch AIはGDP.pdf自体を専門業務由来のmultimodal reasoning benchmarkとして整理しているが、v4.2改定そのものの論評ではない。 36	**v4.2固有の学術的評価は現時点で未確認。**
Mastodon	v4.2固有の実質的な公開コメントを検索で確認できず。	**情報不足。**

主要一般紙・Reuters/Bloomberg級、TechCrunch/Wired/Ars Technica等による**v4.2そのものを主題とした独立記事は、2026年9月6日時点の検索では確認できなかった**。したがって現状の「評判」をメディアコンセンサスと呼ぶのは早く、AI評価コミュニティの初期反応と見るべきである。

興味深い副作用として、**旧スコアと新スコアの流通混乱**もすでに見える。AA公式の現行v4.2ページではClaude Fable 5.1の総合点は57だが、BenchLMの9月4日付ミラーは65.7を掲げつつArtificial Analysis Index 2026としている。日本語ブログが旧版との混同を警告したのは、この種の二次サイト・記事による数字の持ち越しが実際に起きているため重要である。 37

ステークホルダーと代表的発言

Artificial Analysis は測定設計者として最大の当事者である。private化について公式記事は **“40% of our Index weighting is now private, held-out test sets”** と説明し、v5ではこの割合をさらに増やす方針を示している。つまりv4.2は完成形ではなく、v5への移行版という位置付けである。 15

Anthropic、OpenAI、Meta、SpaceXAI/xAI、Moonshot/Kimi、Z.AI、Googleなどのモデル開発企業は、ランキング、マーケティング、モデル最適化のインセンティブを直接受ける。v4.2初期結果ではClaude Fable 5.1が首位、GPT-6 Astraが次点となり、Astraは旧版で見えにくかった世代差が+4ptとして現れた。 38

Surge AI は新たに10%を占めるGDP.pdfの作成者である。GDP.pdfは、単純なOCRや表QAではなく、専門文書に散在する情報を組み合わせて回答することを狙って設計されており、公開datasetとharnessを提供している。 39

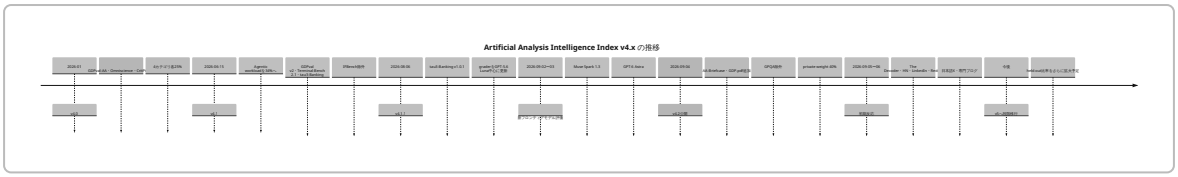
企業のAI導入・調達担当者にとってはポジティブな改定である。LinkedInでMaciej Wierzbinski氏は **“This more aligns with use cases seen in production.”** と評価した。一方、John Kelleher氏は「最高のbenchmarkを作ったのか、それとも現在のLLM市場をユーザーが感じる姿に合わせたのか」と問い、測定対象の定義そのものを問題提起している。 30

AIガバナンス担当者から見れば、Chungsam Lee氏の指摘が重要である。同氏は現実的・privateな評価を改善としながら、agent評価には最終成功率だけでなく、決定、権限、実行、失敗原因、責任までのprovenanceが必要だと論じた。これはv4.2の「capability評価」と、実運用で必要な「reliability/governance評価」が別物であることを示す。 30

研究者・オープンモデル開発者には複雑な影響がある。private testは大手ラボによるbenchmark gamingを難しくする一方、同じ問題を自分でrunして再現することも難しくする。Redditの初期反応には **“v4.2 feels like a reaction to poor press.”** という疑念もあり、今後はrevision policy、旧新スコアの橋渡し、grader変更の監査可能性が信頼形成の鍵になる。 31

リリース前後のタイムライン

v4.2を単独で見るより、2026年を通じてAAが「より実務的・agenticな評価」へ段階的に移行した流れとして見る方が正確である。v4.0は1月、v4.1は6月15日、v4.1.1は8月6日、そして9月4日にv4.2が公開された。直前の9月2〜3日にはMeta Muse Spark 1.3やGPT-6 Astraなど重要なフロンティアモデルの評価が相次いでいた。 [40]



日付	出来事	意義
2026年1月	v4.0	GDPval-AA、AA-Omniscience、CritPtを導入し、Agents/Coding/General/Scienceを各25%に。 [7]
6月15日	v4.1	Agentsを34%へ引き上げ、より難しいagentic tasksへ移行。IFBenchを飽和で除外。 [41]
8月6日	v4.1.1	graderとτ³-Bankingを修正。順位への影響は小規模。 [42]
9月3日	GPT-6 AstraのAA評価	v4.1.1では前世代との差が小さく見え、後の議論の背景となる。 [27]
9月4日	**v4.2公開**	Briefcase/GDP.pdf追加、GPQA削除、private 40%、grading修正、weight全面再配分。 [15]
9月5日	The Decoder報道、HN議論	Astral旧評価への懐疑と改定タイミングが論点化。HNも約140 points規模。 [43]
9月5〜6日	LinkedIn、Reddit、X、日本語ブログ	realism/private化を歓迎する声と、指標の恣意性・改定タイミングへの懐疑が併存。 [44]
将来	v5	AAは追加のincremental releasesとprivate比率のさらなる拡大を予告。具体的リリース日は未公表。 [15]

総合評価と未解決点

総合評価は「方向性は強く支持できるが、単一の“Intelligence”値としての権威は慎重に扱うべき」である。

最大の利点は、公開・飽和した学術テストから、**まだ難しく、業務に近く、過適合しにくい課題へ評価を更新する速度**にある。GPQAを永続的に神聖視せず外す、privateデータを増やす、graderの既知エラーやsandbox failureを修正する、PDF・スプレッドシート・業務成果物を取り込むという判断は、フロンティアモデルの測定として合理的である。 [9]

一方、重要なリスクは四つある。第一に、**再現性**。40% privateはgamingを防ぐが完全独立検証を難しくする。第二に、**measurement drift**。評価・grader・Elo anchor・weightが変わるたび、時系列の点数にはモデル進歩と「物差しの進歩」が混在する。第三に、**構成概念妥当性**。Agents 30%、General 30%などの重みは自然法則ではなく編集判断であり、「総合知能」の唯一の正解ではない。第四に、**範囲不足**。主指数は英語・テキスト中心で、日本語、音声、画像、安全性などは別評価である。 [45]

特に公式ドキュメントには、公開直後らしい**具体的な不整合**も残っている。最新methodologyとversion historyは明確に **Agents 30% / Coding 20% / General 30% / Scientific Reasoning 20%** としているのに対し、現行のIntelligence Index評価ページFAQはなお「4カテゴリが各25%」と記載している。v4.2の正しい計算には前者を使うべきだが、透明性を重視するベンチマークとしては修正すべき文書上の欠陥である。 [46]

今後v5で特に確認すべきなのは、**private setを増やしながらどう外部監査可能性を担保するか、旧版と新版を同一モデルで凍結再実行してmeasurement driftを定量化するか、agentの長期状態・権限逸脱・失敗原因まで評価するか、そして英語以外をheadline intelligenceへ統合するか**である。AA-Briefcaseが「multi-week」と称しながらタスク間のモデル状態を継承しない現在の設計も、真のlong-horizon autonomyを評価するv5では改善余地が大きい。 [47]

したがって、日本企業・研究機関にとっての最も安全な使い方は、**v4.2総合点をモデル選定の一次フィルターにしつつ、用途別の構成benchmark、コスト・時間、日本語Multilingual Index、そして自社のprivate evalを併用すること**である。v4.2は従来より「現場で役に立つモデル」を見つける物差しに近づいたが、「AIの知能そのもの」を一つの数字で決定したわけではない。この区別こそが今回の改定を正しく読む鍵である。 [48]

---

[1] [2] [9] [11] [13] [15] [38] [45] Announcing Artificial Analysis Intelligence Index v4.2 | Artificial Analysis
https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-2

[3] [27] [43] 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] [6] [7] [8] [12] [14] [16] [17] [20] [21] [23] [26] [40] [46] [47] [48] Intelligence Benchmarking | Artificial Analysis
https://artificialanalysis.ai/methodology/intelligence-benchmarking

[5] [42] Launching v4.1.1 of the Artificial Analysis Intelligence Index | Artificial Analysis
https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-1-1

[10] AIベンチマークの読み方 — Agents' Last Exam・ARC-AGI-3 ...
https://ainewsagent.jp/?utm_source=chatgpt.com

[18] [39] GDP.pdf Benchmark: Can Frontier Models Master the ...
https://surgehq.ai/blog/gdp-pdf-can-100b-ai-models-master-the-documents-that-run-the-world?utm_source=chatgpt.com

[19] [24] [41] Artificial Analysis Intelligence Index v4.1: a shift toward agentic workloads | Artificial Analysis
https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-1

[22] My issue with Artificial Analysis's 'intelligence index'
https://www.reddit.com/r/LocalLLaMA/comments/1vhoyw1/my_issue_with_artificial_analysiss_intelligence/?utm_source=chatgpt.com

[25] Consolidated TEXT: 32024R1689 — EN — 27.07.2026
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A02024R1689-20260727

[28] AA Intelligence Index v4.2: Fable 5.1 Leads, Astra Efficient ...
https://www.explainx.ai/blog/artificial-analysis-intelligence-index-v4-2-september-2026?utm_source=chatgpt.com

[29] Artificial Analysis Intelligence Index - Benchmarks
https://benchlm.ai/benchmarks/artificialanalysis?utm_source=chatgpt.com

[30] [44] Announcing Artificial Analysis Intelligence Index v4.2 | Artificial Analysis | 11 comments
https://www.linkedin.com/posts/artificial-analysis_announcing-artificial-analysis-intelligence-activity-7501770733681848320-gKCP

31 New Artificial Analysis scores but Qwen 3.8 27b Still holds up : r/LocalLLaMA
https://www.reddit.com/r/LocalLLaMA/comments/1w7lxeb/new_artificial_analysis_scores_but_qwen_38_27b/

32 Anthropic's Claude Fable 5.1 Tops New AI・Hacker News - Zeli
https://zeli.app/story/49571632?utm_source=chatgpt.com

33 医知創造ラボ 〜脳神経内科医がAIで紡ぐ最新医療情報〜
https://blog.ichisouzo-lab.com/

34 茶谷 和弘 / K.Chatani (@chataclaw) / X
https://x.com/chataclaw?lang=bg&utm_source=chatgpt.com

35 「テストの点数」から「稼ぐ力」へ――AI知能指数 v4.0 がIT技術 ...
https://qiita.com/mhamadajp/items/ff6b198ccac0e87819c9?utm_source=chatgpt.com

36 GDP.pdf
https://epoch.ai/benchmarks/gdp-pdf?utm_source=chatgpt.com

37 Artificial Analysis Intelligence Index v4.2 | Artificial Analysis
https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index