生成 AI 3 モデルの個別評価と知財実務

GPT-6.2 Astra ／ Claude Fable 5.2 ／ Gemini 4 Carbon

GPT-6.2 Astra

Artificial Analysis（AA）の総合指数は Fable が 65、Astra が 62、Carbon が 60 だった。課題別には、Astra は科学ツールによる検証、Fable は専門的な難問と複数資料の統合、Carbon は業務システムの操作と不確かな事実への誤答抑制で強みを示した。[1][2][3][5][7]

各モデルの最高推論設定で比較した。Astra は Max、Fable は Max・Default Fallback、Carbon は High。総合指数 v5.0.1 は知識労働・コーディング・科学・一般知識／推論を各 25%で集計する。[2][3][4]

表 1　総合指数と 9 つの構成評価

評価項目	評価対象	Astra	Fable	Carbon
Intelligence Index v5.0.1	総合指数	62	**65**	60
AA-Briefcase v1.1	複数工程の知識労働（Elo）	1,643	**1,785**	1,611
AutomationBench-AA	業務用 SaaS の自動化	77.2%	76.4%	**77.8%**
GDP.pdf	実務 PDF の全要件達成	**29.8%**	24.2%	28.4%
FrontierCode-AA 2.0	既存リポジトリの改修	45.4%	**48.8%**	40.6%
Terminal-Bench 4.0	端末を使った課題遂行	64.1%	59.6%	**65.2%**
Terminal-Bench-Science 0.1	端末での科学研究作業	**78.1%**	73.3%	64.8%
CritPt v2（難問）	研究水準の物理的推論	71.5%	**76.5%**	69.0%
HLE-Diamond	専門分野の難問への解答	72.2%	**80.2%**	69.1%
AA-Omniscience v1.1	外部検索なしの知識正答率	68.6%	**79.1%**	58.3%
同　ハルシネーション率	非正答のうち誤答した割合 ↓	約 59%	約 59%	**約 11%**

出典　個別得点は[1][2][3]、知識正答率とハルシネーション率は[1][5][6][7]。太字は 3 モデル中の最高成績。Elo は相対評価値。AA-Omniscience は一つの評価を 2 行で示した。

ハルシネーション率の分母は、誤答・部分回答・回答見送りの合計である。約 59%という値は、回答全体の 59%に誤りがあるという意味にはならない。[5]

モデルごとの得意分野と課題

GPT-6.2 Astra　科学ツールを使う検証と原資料の照合

科学研究の実行を測る Terminal-Bench-Science では 78.1％で、Fable を 4.8 ポイント、Carbon を 13.3 ポイント上回った。一方、物理の難問を解く CritPt では Fable が 5.0 ポイント上に立つ。**計算を実行し、観測値と仮説が整合するかを確かめる工程**に、Astra の強みが表れている。[1][2][3]

実務 PDF の全要件達成率も 29.8％で 3 モデル中首位だった。明細書の実施例から配合量・測定条件・物性値を対応付ける作業で試す根拠になる。ただし、全要件を満たしたのは 3 割弱である。外部検索なしの知識正答率は 68.6％で、非正答時の誤答率は約 59％だった。判例番号や公報番号は原典と照合したい。[1][2][3][5][6][9]

Claude Fable 5.2　専門的な難問と複数資料の統合

知識正答率は 79.1％、HLE-Diamond は 80.2％、研究水準の物理問題は 76.5％で、いずれも 3 モデル中首位だった。AA-Briefcase でも Astra に 142 Elo、Carbon に 174 Elo の差を付けた。この評価は、数千の入力ファイルを含む 4 つの業務プロジェクトで成果物を採点する。**技術・市場・競合資料を読み合わせ、経営会議で検討する戦略案を組み立てる工程**と接点がある。[2][3][7][8]

一方、長い実務 PDF を基に全要件を満たす GDP.pdf は 24.2％で 3 モデル中最下位だった。コーディングでも、既存リポジトリの改修は 48.8％で首位、端末を操作する課題は 59.6％で最下位となった。業務を選ぶ際は、文書やコードという対象物に加え、成果物の採点条件に注目したい。[2][3][9]

Gemini 4 Carbon　業務システムの操作と誤答の抑制

業務用 SaaS の自動化は 77.8％、端末による課題遂行は 65.2％で、3 モデル中首位だった。ただし Astra との差はそれぞれ 0.6、1.1 ポイントと小さい。既存リポジトリの改修は 40.6％、科学研究の実行は 64.8％で、どちらも 3 モデル中最下位だった。[2][3]

より明瞭な差が出たのは、不確かな事実に直面した場面である。知識正答率は 58.3％だが、非正答時の誤答率は約 11％にとどまった。**正答数と誤った断定の少なさを別々に評価する**と、Carbon の適性が見える。特許公報の記載を台帳に転記し、確認できない項目を保留する用途が試行候補になる。[3][5][7]

知財実務では何を任せるか

業種別指数では Fable が 6 分野すべてで 3 モデル中首位だった。Carbon は法務で 62、財務・会計で 63 と、いずれも Astra を上回る。工学では Fable が 73、Astra が 67、Carbon が 65 の順だった。[10]

表 2　業種別指数の比較 [10]

モデル	財務・会計	戦略・業務運営	法務	医療・ヘルスケア	工学	経済
Astra	61	63	59	57	67	61
Fable	**65**	**67**	**63**	**65**	**73**	**67**
Carbon	63	61	62	57	65	60

法務指数と案件資料に基づく成果物で順位が逆転

法務指数は法律知識・推論・長文処理などの合成指標である。Harvey LAB-AA v1.1 は 24 分野・120 の非公開課題で案件資料に基づく成果物を採点する。重大なハルシネーションが一つでもあれば、その課題の得点はゼロになる。[11][12]

表 3　法務指数と案件課題の評価結果 [10][11][12]

指標	Astra	Fable	Carbon
法務指数	59	**63**	62
案件課題の全要件達成率（重大な誤りなし）	**15.0%**	7.2%	未確認
1 課題当たりの重大なハルシネーション	**0.06 件**	0.75 件	未確認

案件資料に基づく成果物では Astra が Fable を上回った。Carbon の同一設定の公開値は確認できなかった。Astra の 15.0%は通常版の成績である。[12]

特許実務への当てはめ

Astra は、明細書・引例・拒絶理由通知書を突き合わせる対比表や実施例の数値照合。**Fable** は、複数文献からの論点抽出や反論候補の展開、知財戦略案の起草。**Carbon** は、新着特許の定期監視（SDI）での管理表登録や担当者への振り分け、根拠を確認できない項目の保留が試行候補となる。[2][3][5][8][9][12]

日本語の特許実務への適性は用途仮説である。自社案件では、重要論点の抽出率、引用段落と結論の一致率、根拠がないのに記入した件数を分けて評価したい。

参考文献

閲覧日はすべて 2026 年 11 月 18 日（日本時間）。[6][7]は AA のデータを再掲した資料で、知識評価の小数値の照合に用いた。各 URL は掲載元にリンクする。

[1] Artificial Analysis　OpenAI has the #2 model in the Intelligence Index with the release of GPT-6.2 Astra and GPT-6.2 Sol. 2026 年 11 月 17 日。発表時の総合順位・科学評価・誤答率。
https://artificialanalysis.ai/articles/openai-has-the-2-model-in-the-intelligence-index-with-the-release-of-gpt-6-2-astra-and-gpt-6-2-sol/

[2] Artificial Analysis　Release Comparison Tool. Astra（Max）と Fable（Max）の個別評価。
https://artificialanalysis.ai/models/comparisons

[3] Artificial Analysis　Gemini 4 Carbon (High) vs GPT-6 Astra (Medium). Carbon（High）列の個別得点を参照。
https://artificialanalysis.ai/models/comparisons/gemini-4-carbon-vs-gpt-6-astra-medium

[4] Artificial Analysis　Announcing Artificial Analysis Intelligence Index v5. 2026 年 10 月 27 日。総合指数の評価項目と配点。
https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v5/

[5] Artificial Analysis　AA-Omniscience v1.1: Knowledge and Hallucination Benchmark. 正答率・ハルシネーション率の定義。
https://artificialanalysis.ai/evaluations/omniscience

[6] Village LLM Atlas　GPT-6.2 Astra benchmark scores. AA データの 2026 年 11 月 17 日時点の記録。知識正答率を参照。
https://atlas.village.dev/models/gpt-6-2-astra

[7] Klu　AA-Omniscience v1.1: Knowledge and Hallucination Results. Fable と Carbon の AA 評価値。2026 年 11 月 14 日集計。
https://klu.ai/benchmarks/aa-omniscience

[8] Artificial Analysis　AA-Briefcase v1.1: Agentic Knowledge Work Benchmark. 業務プロジェクトと採点方法。
https://artificialanalysis.ai/evaluations/aa-briefcase

[9] Artificial Analysis　GDP.pdf Benchmark. 専門文書の入力方法と全要件達成率の定義。
https://artificialanalysis.ai/evaluations/gdp-pdf

[10] Artificial Analysis　AI Model & API Providers Analysis. 「Capability Index Comparison」掲載の 6 業種の指数値。
https://artificialanalysis.ai/

[11] Artificial Analysis　Best AI for Legal: LLM Leaderboard. 法務指数の構成・配点。
https://artificialanalysis.ai/indexes/legal

[12] Artificial Analysis　Harvey LAB-AA v1.1 Benchmark. 評価条件・Leaderboard 掲載の全要件達成率および重大な誤りの件数。
https://artificialanalysis.ai/evaluations/harvey-lab-aa