Gemini 4 Carbon 調査報告

一般提供の内容と性能評価および初期の評判

調査基準日 2026 年 11 月 11 日

対象読者 知財部門 研究開発部門 経営企画部門

GPT-6 Astra

Google は米国時間 2026 年 11 月 10 日、Gemini 4 Carbon の提供対象を開発者、企業、研究者、一般ユーザーへ拡大した。本報告は、その提供条件、基本仕様、第三者ベンチマーク、初期利用者の反応を整理し、知財実務への適用可能性を検討する。[1][2]

現時点では、複雑な資料を読み、ツールを使いながら長い作業を進める能力に強みがあるモデルと評価できる。ただし、用途によって競合との優劣が変わるため、待ち時間と課題全体の費用を含めて選ぶ必要がある。総合ランキングだけで採用を決めることは適切でない。

調査方法と範囲　本報告は公開資料に基づく調査であり、Carbon を実行した独自の比較実験ではない。Google の発表、第三者による測定、個人の感想、本報告の推論を区別した。動的なランキングや価格は調査時点の掲載値である。本文中の［番号］は文末の参考文献に対応し、番号をクリックすると該当書誌へ移動できる。

1 一般提供の範囲と利用条件

今回の一般提供は利用資格の拡大であり、無料版や AI Pro 契約だけで Gemini アプリの Carbon を使えるようになったことを意味しない。一般ユーザー向けのアプリでは、Google AI Ultra 契約者に順次提供される。一方、有料 API は Ultra 契約とは別の利用経路である。[2][3][4]

表 1 利用経路ごとの提供条件

利用経路	2026 年 11 月 11 日時点の条件
Gemini アプリ	Google AI Ultra 契約者に順次提供
Google 検索の AI モード	Google AI Ultra 契約者に順次提供
Gemini API と AI Studio	有料 API 利用者向け　無料ティアでは利用不可
Google Antigravity	Ultra 契約者向けの提供を案内
企業向け	Gemini for business などを通じて提供
サイバー防御向け特別アクセス	Fairwind Program を通じた提供を継続

出典 [1][2][3][4]。実際の表示・利用枠は、提供経路、契約、順次展開の状況に左右される。研究者には選考制の利用支援プログラムも案内されている。[1]

2 基本仕様と設計上の特徴

Google は、日常的な開発や知識業務の多くには Argon を位置付け、Carbon を複雑なデバッグや答えが定まっていない科学研究などに向けている。Carbon は、より多くの思考トークンを使い、仮説の検討と検証を繰り返す設計である。[1]

表 2 基本仕様

項目	内容
API モデル名	gemini-4-carbon
入力	テキスト　画像　音声　動画　PDF
出力	テキスト
コンテキスト	入出力を合わせて最大 1,048,576 トークン
通常の同期処理の出力上限	65,536 トークン
長文生成	バックグラウンド実行で、入力分を差し引いた統合コンテキストの範囲まで拡張
推論設定	Low　Medium　High　推奨初期設定は Medium
ツール連携	コード実行　Computer Use　ファイル検索　関数呼び出しなど

出典 [5][6]。API の仕様を示しており、アプリの画面や利用枠が同じであることを意味しない。

100 万トークンの入力に加えて 100 万トークンを出力できる仕様ではない。思考トークンも出力上限を消費するため、長い資料を大量に投入すると、推論や回答に使える余地が減る。長文生成では、入力、思考、回答の配分を設計する必要がある。[6]

3 複数エージェントによる長時間作業

一般提供と同時に、Antigravity では「/swarm」のプレビューが発表された。これは Carbon というモデル自体と区別すべき、開発環境側の機能である。調整役、実装担当、検証担当などのエージェントが作業を分担し、共有掲示板で連絡しながら進める。[7]

コード変更は独立した作業領域で行われ、検証役がテストや変更内容を点検する。期待される用途は、資料の読解、問題の特定、修正、動作確認という複数の工程を連続して進める作業である。[7]

ただし、複数のエージェントを動かせば常に品質が上がるとは限らない。本報告では、成果を評価する際に、モデル単体の能力に加えて、役割分担、利用可能なツール、停止条件、検証方法を含むシステム全体を確認すべきと考える。

4 第三者ベンチマークによる性能評価

10 月 28 日の Google 発表では、Artificial Analysis の総合指数 v5.0 で Carbon は 59 点、首位とされた。今回確認した v5.0.1 では Carbon High が 60 点、Claude Fable 5.2 Max が 65 点である。発表時点の順位を現在の順位として紹介することは避ける必要がある。[8][9]

表 3 Carbon と Claude Fable 5.2 の評価項目別比較

評価項目	Carbon High	Fable 5.2 Max	出典
AA 総合指数	60	65	[9]
Terminal-Bench 4.0	65.2%	59.6%	[9]
FrontierCode-AA 2.0	40.6%	48.8%	[9]

評価項目	Carbon High	Fable 5.2 Max	出典
Terminal-Bench-Science 0.1	64.8%	73.3%	[9]
CursorBench 4.0	60.4%	60.0%	[10]
Zapier AutomationBench	53.88%	51.45%	[11]
MKA Climb WSR	54%	57%	[12]

注　いずれも値が高い方が良い。Fable の評価には、評価元が明記する既定のフォールバックを含むものがある。各行は異なる試験であり、行をまたいで百分率の大小を比較できない。AA の AutomationBench-AA と、Zapier の AutomationBench も、実行・採点条件を確認して区別する。

Carbon は、ターミナルを使う複合的な作業や、業務ツールをまたぐ処理で良好な成績を示す。一方、別の実践的開発評価や科学研究作業では Fable が上回る。「コーディングなら Carbon が全面的に上」とも、「総合指数の首位モデルが全用途で上」ともいえない。[9][10][11]

CursorBench の 0.4 ポイント差は、掲載順位としては首位でも、それだけで実務上の決定的優位を意味しない。試験課題、設定、実行環境、再現性を確認し、自社の作業で差が出るかを検証する必要がある。

Zapier の試験は返答文の見栄えではなく、操作後のデータが所定の状態になったかを採点する。Carbon の強さを裏付ける一方、掲載スコアは約 54%であり、この結果を無人で業務を任せられる保証として扱うことはできない。[11]

5 図面や表を含む資料と科学研究への適性

知財・製造業にとって注目すべき点は、図面や表を含む資料から成果物を作る能力である。MKA は、建築図面、工学的な模式図、スキャン画像などを読み、報告書、表計算、プレゼンテーションなどを作成する、10 業種・200 課題の評価である。[12]

MKA の課題は Google DeepMind が開発し、Artificial Analysis が自らの実行環境で評価している。採点は基準表に沿った AI 判定を使う。54%という値は、難しい評価項目の加重充足率であり、「54%の案件を完全処理した」という意味ではない。外部実行による比較として有用だが、課題設計を含めて Google と無関係な試験とは区別すべきである。[12]

Artificial Analysis の MKA 解説には、1 課題当たりの費用として Carbon が 6.29 ドル、Fable 5.2 が 31.97 ドルと掲載されている。この評価では、Carbon は上位に近い品質を約 5 分の 1 の費用で示す。ただし、業種別・成果物別で首位モデルが変わり、図面・文書処理で一律に最高という結果ではない。[13]

科学研究では、Google は実験ノートと生データから異常を見つけ、混合不足や測定装置の試料持ち越しなどを指摘した事例を紹介している。材料開発や実験データのレビューに関心を持つ理由になるが、これらは Google の事例報告であり、広範な独立再現試験とは分けて評価する必要がある。[1]

6 サイバーセキュリティ性能と評価条件

Carbon については、脆弱性の発見、悪用可能性の検証、修正の各段階で高い性能が公表されている。Google 発表の主要な値を表 4 に示す。[8][14]

表 4 公表されたサイバーセキュリティ評価

評価	主な対象	Carbon
SEC-bench Pro	脆弱性の発見	88.2%
ExploitGym	悪用可能性の検証	54.8%
CWE-bench	脆弱性の修正	71.0%

出典 [8][14]。一般向け製品の利用条件とは区別して読む必要がある。

10 月の発表には、これらをガードレールなしで評価したとの記載がある。モデルカードでは、サイバー評価値は Fairwind Program 経由の性能として示されている。一般ユーザー向けのアプリや API で、同じ条件・挙動が得られることを示す値ではない。[8][14]

安全性報告書も、モデル固有の挙動を測る評価と、本番環境で追加される防御措置を区別している。性能評価と、安全機構を含む製品体験を分けて読む必要がある。[15]

7 事実性とハルシネーションの評価

Artificial Analysis の AA-Omniscience では、Carbon Medium のハルシネーション指標は約 4%、High は約 11%と掲載されている。事実性の観点では好材料だが、指標の定義を正確に理解する必要がある。[16]

この指標は「全回答のうち誤りが 4%」という意味ではない。正解できなかった応答群の中で、分からないと認めたり回答を控えたりせず、誤答した割合を測る。具体的な分母は、誤答、部分回答、未回答の合計である。「正答率 96%」や「特許調査の誤判定率 4%」への読み替えはできない。[16]

不確かな事項を断定しにくい傾向は、根拠確認を重視する実務には有益と考えられる。一方、Medium の方がこの指標では良好であり、推論設定を High にすれば、あらゆる品質指標が改善するわけではない。

8 料金と待ち時間を含む実用性

料金は導入価格と通常価格を分けて判断する必要がある。標準 API では思考トークンも出力料金の対象であり、検索などのツール利用には別の料金条件がある。[3]

表 5 標準 API 料金 100 万トークン当たり

項目	2026 年 11 月 30 日まで	12 月 1 日以降
入力	2 ドル	4 ドル
出力と思考トークン	10 ドル	20 ドル
キャッシュされた入力	0.10 ドル	0.20 ドル

出典 [3]。標準料金を記載。Batch、Flex、Priority には別料金が設定されている。

例えば、キャッシュなしで入力 10 万トークン、思考を含む出力 2 万トークンなら、標準 API の単純計算は導入期間中 0.40 ドル、12 月以降 0.80 ドルとなる。さらに思考トークンを 20 万使えば、合計はそれぞれ 2.40 ドル、4.80 ドルとなる。これは表 5 に基づく試算で、ツール料金などは含まない。表示された回答の長さだけでは請求額は分からない。

表 6 推論設定ごとの AA 評価結果

設定	総合指数	1 課題の費用	1 課題の時間
Carbon Low	54	7.19 ドル	約 23 分
Carbon Medium	58	6.59 ドル	約 24 分
Carbon High	60	7.76 ドル	約 26 分
Fable 5.2 High	62	9.08 ドル	約 17 分
Fable 5.2 Max	65	26.54 ドル	約 45 分

出典 [9]。AA の難しい評価課題、価格設定、時間算定方法に基づく値であり、通常のチャット 1 回の費用・時間ではない。Carbon の費用には通常単価が使用されている。時間は閲覧時の値を丸めて表示した。

注目点は、Low が最安ではなく Medium が最安になっていることである。推論を弱めても、余計な試行や作業の長期化が起きれば、課題全体では安くならない可能性がある。ただし、この表だけでは原因を確定できない。

応答速度について、AA は Carbon High の最初の回答トークンまで約 66 秒と測定している。資料を渡して待つ用途と、対話しながら細かく修正する用途では、体感評価が大きく変わると考えられる。[17]

9 公開直後の評判とレビュー

公開直後の評判には、能力への期待と使い勝手への不満が併存する。ただし、実使用例、個人の感想、公開情報の解釈を混同してはならない。

開発者 Simon Willison は、推論設定別の描画例と、Antigravity で作成した Three.js アニメーションを公開した。単独 HTML ファイルへのまとめ直しには追加指示を使っている。成果物を伴う実使用例として参考になるが、モデル全体の優劣を判定する比較試験ではない。[18]

Hacker News では、利用者 5555watch が、自分の論文草稿について、他モデルが見落とした興味深い点を Carbon が指摘したと報告している。本人も 1 例の体験であると明記しており、肯定的な兆候として扱うのが適切である。[19]

同じ議論では、Ultra 限定の利用資格や、個人アカウントと Workspace アカウントのモデル表示、提供状況の分かりにくさへの不満が見られる。これらは主として提供方法・製品体験への評価である。Argon に関する不満も混じっており、そのまま Carbon の欠点として扱うべきではない。[19]

eesel AI のレビューは、難しい推論や調査を高く評価する一方、応答を待つ用途への評価を低くしている。ただし、同記事は独自ベンチマークを実施していないと明示している。本報告では同記事を評論として扱い、性能数値は評価元の資料を優先した。[20]

現段階で「利用者の大多数が高評価」「実務で安定性が実証済み」とは言えない。確認できた情報は、継続利用の評判よりも発売直後の試用結果やアクセスに関する議論が中心である。

10 日本語の知財実務への適用可能性

以下は公開性能からの本報告の推論であり、日本語の特許実務で実証された成績ではない。Carbon を評価する場面として、複雑な資料の精読と、既存案の欠点を探す作業に優先度があると考える。

表 7 知財実務での試用対象と確認項目

業務	試す価値がある使い方	評価時の確認項目
明細書と図面の読解	構成要素、符号、実施例、作用効果の対応付け	本文と図面の区別、根拠箇所の正確さ
クレーム対比	請求項の分解と複数文献の開示箇所の整理	文献にない構成を補っていないか
拒絶理由対応	引用発明認定、相違点、補正案の弱点の点検	反論に不利な記載も拾えるか
無効資料調査	候補文献の精読、証拠不足、追加検索事項の抽出	関連文献の見落としと開示の過大評価
IP ランドスケープ	特許、論文、企業資料の横断整理	集計の再現性と結論ごとの出典
材料と製造研究	図表や実験データの矛盾と説明不足の探索	原データと推論の区別

表 7 は本報告による用途仮説であり、各業務での性能保証や法的判断の代替を意味しない。

検索の網羅性や日本法に即した判断精度は、一般的な知識業務・法務ベンチマークからは保証されない。今回確認した資料の範囲では、日本語特許公報の読解、先行技術の見落とし率、FTO 判断について、十分な独立比較評価は確認できなかった。

モデルカードに記載された知識カットオフは 2026 年 3 月である。最新の公報、審査基準、判決を扱う場合は、原文や検索結果を与える運用が必要となる。同カードは、ハルシネーション、遅延、タイムアウトも既知の限界として挙げている。[14]

機密情報の扱いでは、有料 API の料金表に製品改善への利用なしと記載されている。ただし、その記載だけで保存、ログ、外部ツールへの送信条件まで判断することはできない。契約する経路の条件を個別に確認する必要がある。Carbon は独自モデルで、モデルウェイトは一般公開されておらず、ローカル LLM として導入する選択肢には含まれない。[3][17]

11 導入判断のための検証方法

導入を判断する際は、Carbon Medium を起点に、High と既存モデルを同じ案件で比較する方法が適切である。Google も 3 段階の推論設定を評価し、必要な品質を満たす設定を選ぶよう案内している。[6]

例えば公開済み案件 20 件程度を用い、重要事項の見落とし、根拠引用の正確さ、文献にない内容の補完、人が修正した時間、処理時間と総費用を記録する。これは小規模な予備評価案であり、20 件だけで全業務への一般化ができるという趣旨ではない。

採用判断では、回答の見栄えや総合ランキングに加え、専門家が確認・修正する時間を、どの工程でどれだけ減らせるかを中心に据える。資料読解、反証探索、対比表作成などの工程ごとに、得意・不得意と費用対効果を確認することが望ましい。

参考文献

全資料の閲覧日 2026 年 11 月 11 日。公開日が明記されていない動的ページは、その旨を記した。本文の数値は調査時点の掲載値であり、リンク先は更新される場合がある。Google 資料は提供者の公式情報、各評価機関のページは当該評価の一次資料、掲示板とレビューは初期の反応を把握する資料として用いた。

[1] Koray Kavukcuoglu／Google. Gemini 4 Carbon is now available. The Keyword. 2026 年 11 月 10 日.
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-carbon-now-available/
参照内容　一般提供の発表、製品の位置付け、研究利用の事例。　閲覧日 2026 年 11 月 11 日。

[2] Koray Kavukcuoglu／Google. Gemini 4 Carbon の提供を開始. Google Japan Blog. 2026 年 11 月 11 日.
https://blog.google/intl/ja-jp/company-news/technology/gemini-4-carbon/
参照内容　日本語の公式発表と利用経路。　閲覧日 2026 年 11 月 11 日。

[3] Google. Gemini API pricing. Google AI Studio Documentation. 更新型ページ　公開日記載なし.
https://aistudio.google.com/docs/pricing
参照内容　Gemini 4 Carbon 欄。標準料金、導入期間、思考トークン課金、データ利用の表示。　閲覧日 2026 年 11 月 11 日。

[4] Google. Gemini 4 Carbon is now available for developers. The Keyword. 2026 年 11 月 10 日.
https://blog.google/innovation-and-ai/technology/developers-tools/gemini-4-carbon-developers/
参照内容　開発者向けの提供経路。　閲覧日 2026 年 11 月 11 日。

[5] Google. Gemini 4 Carbon. Google AI Studio Documentation. 更新型ページ　公開日記載なし.
https://aistudio.google.com/docs/models/gemini-4-carbon
参照内容　モデルコード、対応入力、コンテキスト、出力上限、ツール対応。　閲覧日 2026 年 11 月 11 日。

[6] Google. What's new in Gemini 4 Carbon. Google AI Studio Documentation. 更新型ページ　公開日記載なし.
https://aistudio.google.com/docs/whats-new-gemini-4-carbon
参照内容　推論設定、長文生成、入力と出力の統合上限、評価・移行時の留意点。　閲覧日 2026 年 11 月 11 日。

[7] The Antigravity Team／Google. Gemini 4 Carbon and /swarm. Google Antigravity Blog. 2026 年 11 月 10 日.
https://antigravity.google/blog/gemini-4-carbon-in-google-antigravity/
参照内容　複数エージェントの協調、作業領域、検証担当の構成。　閲覧日 2026 年 11 月 11 日。

[8] Koray Kavukcuoglu／Google. Gemini 4 Carbon: our next frontier model. The Keyword. 2026 年 10 月 28 日.
https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-4-carbon/
参照内容　初回発表、当時の AA 指数、サイバー評価の値と条件。　閲覧日 2026 年 11 月 11 日。

[9] Artificial Analysis. Claude Fable 5.2 vs Gemini 4 Carbon: Release Comparison. Artificial Analysis. 更新型ページ　公開日記載なし.
https://artificialanalysis.ai/models/releases/comparisons/claude-fable-5-2-vs-gemini-4-carbon
参照内容　AA 指数 v5.0.1、個別評価、推論設定別の費用と時間。Fable の既定フォールバックを含む表記に注意。　閲覧日 2026 年 11 月 11 日。

[10] Cursor. CursorBench 4.0. Cursor. 更新型ページ　公開日記載なし.
https://cursor.com/cursorbench
参照内容　実際の Cursor セッションに由来する複数ファイルの開発課題。　閲覧日 2026 年 11 月 11 日。

[11] Zapier. AutomationBench AI benchmark leaderboard. Zapier. 更新型ページ　参照時の表示版 1.0.6.
https://zapier.com/benchmarks
参照内容　操作後の環境状態を採点する業務自動化評価。掲載値とフォールバック条件。　閲覧日 2026 年 11 月 11 日。

[12] Artificial Analysis. Multimodal Knowledge Agent (MKA) Benchmark. Artificial Analysis. 更新型ページ　公開日記載なし.
https://artificialanalysis.ai/evaluations/mka
参照内容　200 課題、10 業種、Climb WSR、実行・採点方法とランキング。　閲覧日 2026 年 11 月 11 日。

[13] Artificial Analysis. Multimodal Knowledge Agent (MKA) benchmark: testing knowledge work capabilities with complex visual files. Artificial Analysis. 2026 年 11 月公開.
https://artificialanalysis.ai/articles/google-deepmind-multimodal-knowledge-agent
参照内容　MKA の解説、モデル別費用、業種別・成果物別の結果。　閲覧日 2026 年 11 月 11 日。

[14] Google DeepMind. Gemini 4 Carbon Model Card. Google DeepMind. 2026 年 11 月 10 日.
https://deepmind.google/models/model-cards/gemini-4-carbon/
参照内容　仕様、評価条件、Fairwind 経由のサイバー評価、知識カットオフと既知の限界。　閲覧日 2026 年 11 月 11 日。

[15] Google DeepMind. Gemini 4 Carbon Safety & Security Report. Google DeepMind. 2026 年 11 月　94 頁.
https://storage.googleapis.com/deepmind-media/gemini/gemini_4_carbon_safety_report.pdf
参照内容　特に Executive Summary。モデル評価と本番の追加防御措置の区別。　閲覧日 2026 年 11 月 11 日。

[16] Artificial Analysis. AA-Omniscience v1.1: Knowledge and Hallucination Benchmark. Artificial Analysis. 更新型ページ　公開日記載なし.
https://artificialanalysis.ai/evaluations/omniscience
参照内容　ハルシネーション指標の定義と Carbon の設定別結果。　閲覧日 2026 年 11 月 11 日。

[17] Artificial Analysis. Gemini 4 Carbon (High) Intelligence, Performance & Price Analysis. Artificial Analysis. 更新型ページ　公開日記載なし.
https://artificialanalysis.ai/models/gemini-4-carbon
参照内容　応答開始までの時間、課題費用、モデルの公開形態など。　閲覧日 2026 年 11 月 11 日。

[18] Simon Willison. Release: llm-gemini 0.36. Simon Willison’s Weblog. 2026 年 11 月 10 日.
https://simonwillison.net/2026/Nov/10/llm-gemini/
参照内容　設定別の描画と Antigravity でのアニメーション制作の実使用例。　閲覧日 2026 年 11 月 11 日。

[19] Hacker News 投稿者. Gemini 4 Carbon is now available. Hacker News. 2026 年 11 月 10 日から 11 日の議論.
https://news.ycombinator.com/item?id=50442303
参照内容　5555watch の論文草稿に関する報告、提供条件への反応など。個人の自己申告であり代表性は未確認。　閲覧日 2026 年 11 月 11 日。

[20] eesel AI. Gemini 4 Carbon review: the most honest default setting at the top. eesel AI Blog. 2026 年 11 月公開.
https://www.eesel.ai/blog/gemini-4-carbon-review
参照内容　初期レビュー。独自ベンチマークを実施していないと明記。本報告では評論として参照。　閲覧日 2026 年 11 月 11 日。