ChatGPT

Google Gemini 4 Argon 総合調査報告

調査基準時点：2026年10月14日（日本時間）／グローバル一般提供日：2026年10月13日（米国時間）

エグゼクティブサマリー

Googleは2026年9月30日（米国時間、日本時間10月1日）に「Gemini 4 Argon」を発表し、当初は高度なサイバー能力を理由に、審査済み防御組織向け「Fairwind Program」から限定展開した。その後、**10月13日（米国時間）にグローバル提供へ移行**した。現在の主要な提供面は、開発者向けのGemini API／Google AI Studio／Gemini platform、企業向けGemini for business、Google Antigravity、個人向けGeminiアプリ、Google SearchのAI Modeである。GeminiアプリとAI ModeではGoogle AI Pro/Ultra、Antigravityでは少なくともローンチ告知上はUltraが主要対象となっている。日本語のGoogle AIプランページではProへのArgonアクセスも「rolling out」とされており、**同日ロールアウト中のため、国・アカウント・製品面ごとにアクセス条件が完全には揃っていない**。 [1]

結論を先に述べると、Argonの最大の特徴は単純な「チャット知能の更新」ではなく、**長時間エージェント、実コードベース改修、法務・財務、サイバー防御、科学研究、複雑なマルチモーダル資料を一つの仕事として最後まで処理するモデル**へ重点が移ったことにある。Googleは1,048,576トークンを入力・思考・出力で共有する「unified」コンテキストを導入し、通常の同期応答は最大65,536トークン、長時間実行では64K単位の継続処理を組み合わせ最大約100万トークンまで生成できる仕組みを用意した。 [2]

技術的には、Googleが公開したモデルカードで**大規模な疎なMixture-of-Experts（MoE）Transformer、ネイティブ・マルチモーダル**であることが確認できる一方、総パラメータ数、アクティブパラメータ数、専門家数、学習計算量、TPU世代・クラスタ規模、量子化方式、推論精度、推論用GPU/TPU構成は非公開である。したがって「1Tパラメータ級」などの推測値を事実として扱うべきではない。Googleが確認しているのは、TPU/TPU PodとJAX、ML Pathwaysによる学習である。 [3]

性能面はかなり強いが、Googleのマーケティング表現には留保が必要である。同一のGoogleモデルカード評価では、Argonは19項目中**13項目で単独首位、1項目で同点首位、5項目で競合に敗れている**。特にVals Index、DeepSWE、法務、長文GraphWalks、LVBenchなどで強い一方、FrontierSWE v2、Terminal-bench 4.0、Terminal-Bench ScienceではGPT-6 AstraやClaude Opus 5.5が上回る。つまり「全方位で最強」ではなく、**長文・専門業務・エージェント実務に強く、ターミナル型コーディング／科学エージェントでは競合に残る優位領域がある**というのがより正確である。 [4]

第三者評価も概ね「トップ層への復帰」を支持する。Artificial AnalysisではArgon Highが同機関のIntelligence Indexで53を記録し、AutomationBench-AAでは78%、Terminal Bench 4では約57%、独自のOmniscience評価では**hallucination rate 15%**だった。ただし同評価で正答率そのものは50%であり、低hallucination率の一部は「誤答を作る代わりに答えない」挙動による。したがって「Argonがハルシネーションを解決した」というコミュニティ上の主張は過大である。 [5]

速度についても「Flash級」と考えるべきではない。Artificial AnalysisのArgon High測定では、出力開始後の生成速度はVertex経由で約116.5 tokens/s、AI Studioで約99.5 tokens/sと高速だった一方、約10K入力を使った最初の回答トークンまでの時間はAI Studio約44.7秒、Vertex約45.6秒で、思考時間を含めるとかなり長い。**高スループットだが高い思考レイテンシを許容する「deep work」用モデル**という位置づけが妥当である。 [6]

価格競争力は非常に強い。APIの導入価格は2026年11月30日まで100万トークン当たり入力$2、出力$10、キャッシュ入力は95%引き。12月1日以降は$4/$20になる。Free API Tierはない。これでも2026年の最上位エージェント型モデルとしては攻撃的な価格設定であり、Googleがモデル性能だけでなく**「長時間エージェントを経済的に大量稼働させるプラットフォーム」**として勝負していることがうかがえる。 7

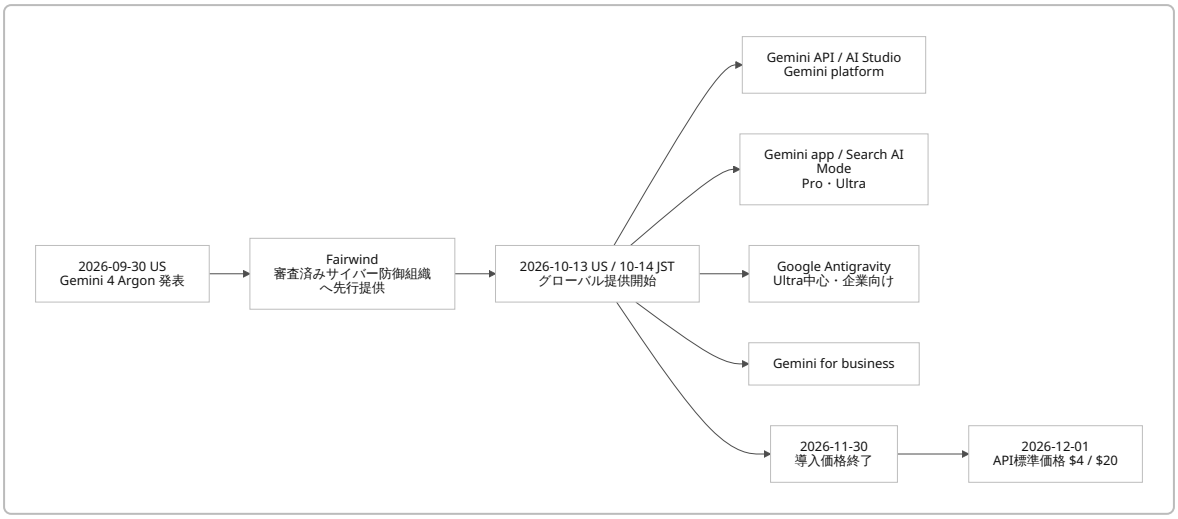
安全性では、Gray SwanのIndirect Prompt Injection評価で攻撃成功率0.7%、エージェントが制約を迂回しようとする率を2%まで下げたとGoogleは報告している。一方、サイバー能力自体が強いため、一般モデルには攻撃用途の制限を入れつつ、Fairwindの審査済み防御組織にはサイバーガードレールを外した版を提供する二層構造を採用した。これはArgonの技術的強みであると同時に、**アクセス統制、誤用、国家・重要インフラ用途を巡る評判リスク**でもある。 8

発表、製品展開、Googleエコシステムでの位置づけ

Argonは「Gemini 3.xの単純な微調整」ではない。Googleのモデルカードは、既存モデルのfine-tuneやmodificationではなく、新しいフロンティアモデルだと明記している。9月30日の最初の発表では、ソフトウェアエンジニアリング、法務・金融などの企業知識労働、サイバー防御を主用途に据え、Fairwind参加者から段階提供した。日本語版Google Blogも10月1日に同内容を公開している。 9

10月13日の第二段階が今回の「グローバルリリース」である。Google自身は、開発者・企業向けにはAI StudioおよびGemini platformのAPI、Ultra加入者にはAntigravity、Pro/Ultra加入者にはGeminiアプリとSearch AI Mode、と明記した。Google Cloud側でもモデルIDを `gemini-4-argon` として企業向けAPIに投入している。 8

この二段階展開と価格変更予定を時系列にすると次のようになる。 10



各製品面の違いは重要である。

製品面	Argonの提供状況	主な対象・特徴
Gemini API / AI Studio	有料APIで提供	`gemini-4-argon`、SDK day-one対応、Free Tierなし
Gemini platform / Google Cloud	提供開始	企業API、Gemini for business、エンタープライズ統制

製品面	Argonの提供状況	主な対象・特徴
Geminiアプリ	Pro / Ultraへロールアウト	長時間タスク、より直接的な回答、タスク実行
Search AI Mode	Pro / Ultra	複雑・多段階検索、会話的な指示追従
Google Antigravity	ローンチ告知ではUltra、企業契約でも提供	`/goal`、長時間コーディング、マルチエージェント
Android Studio	開発者向け展開	Googleの開発ツール群との統合
Fairwind	審査制	一部防御組織にはサイバーガードレールなし
Robotics	API/Antigravity経由	物理推論、Real2Sim、ロボット制御研究

Google公式発表、Cloud、Robotics開発者ブログによる。 [11]

Geminiアプリ側では、Argonは従来モデルより「hedgingを減らす」「無用な繰り返しや定型句を減らす」よう調整され、長いタスクでは進行状況を示す設計が強調されている。またGemini環境は40以上のアプリとの接続を拡大しており、macOS版ではAntigravity型の作業環境との連携も進められている。 [12]

Searchでは、ArgonはAI Modeに投入され、「ニュアンスのある多段階タスク」「ユーザー意図の理解」「往復会話での命令追従」の改善が公式に強調されている。つまりArgonは単なるWeb検索回答モデルではなく、Searchを**検索＋継続的タスク実行**へ拡張するための基盤モデルとして位置づけられている。 [13]

一方で、アクセス表記にはローンチ当日の食い違いがある。10月13日の中央発表はAntigravityをUltra向けと記述する一方、日本語Google AIプランページはProについてもargonのAntigravityアクセスを「expanded access」、Argon全体を「rolling out」と表示している。同ページはArgonがUltraで利用可能、Proでは一部国へ段階展開中とも明記しているため、**「世界リリース済み」と「全Proユーザーが即時利用可能」は同義ではない**。 [14]

技術仕様、アーキテクチャ、モダリティ、計算基盤

Googleの公開情報から確認できる技術仕様は以下である。特に重要なのは、**パラメータ規模を公開していない**点である。モデルカードは大規模sparse MoE Transformerとしかしておらず、Artificial Analysisもパラメータ数を「非公開」と扱っている。 [15]

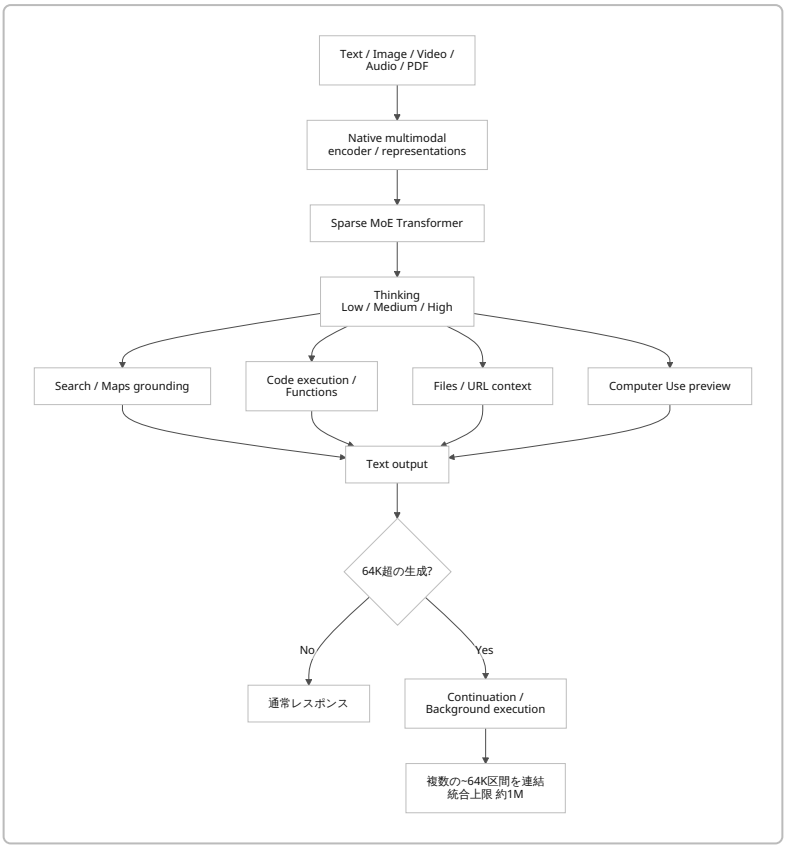
比較軸	Gemini 4 Argonの公開仕様
モデル種別	フロンティア推論モデル、旧モデルのfine-tuneではない
アーキテクチャ	大規模 sparse Mixture-of-Experts Transformer
パラメータ数	**非公開**
Active parameters / experts数	**非公開**
入力	Text / Image / Video / Audio / PDF
ネイティブ学習モダリティ	Text / vision / audioを含むnative multimodal

比較軸	Gemini 4 Argonの公開仕様
API出力	Text
unified token limit	1,048,576 tokens
通常最大出力	65,536 tokens（同期系）
Long Decode	最大約1,048,576、64K単位の継続処理
Thinking	Low / Medium / High
Function calling	対応
Code execution	対応
Structured output	対応
Google Search grounding	対応
Maps grounding	対応
URL context	対応
File Search	対応
Computer Use	Preview
Live API	Argonでは非対応
画像生成・音声生成	非対応
Explicit context cache	非対応
Implicit cache	対応
Batch / Flex / Priority	対応
学習ハードウェア	Google TPU / TPU Pods
学習ソフトウェア	JAX、ML Pathways
量子化方式	**非公開**
学習時数値精度	**非公開**
推論TPU/GPU型番	**非公開**
Edge/on-device weights	**公開されていない**

Google API仕様、モデルカードによる。 16

「100万出力トークン」は少し注意が必要である。Argonでは入力・思考・出力が独立して各100万なのではなく、**1,048,576トークンの統合予算を共有する**。標準的な一回の同期生成は65,536トークンまでで、それ以上はbackground executionとcontinuation protocolを通して複数の約64K区間を連結する。したがって「一回の通常HTTPレスポンスで100万トークンが一気に流れる」と理解するのは正確ではない。 17

概念的な実行経路は次のように整理できる。 18



学習データについてGoogleは、公開Web文書、テキスト、コード、画像、音声、動画からなる大規模かつ多様な事前学習データを使用したと説明している。より詳細なモデルカードでは、公開データ、クローリングデータ、商用ライセンスデータ、規約・プライバシー設定に従ったGoogle製品・サービス由来のユーザーデータ、Googleが取得・生成したデータ、ワークフォース由来データ、synthetic dataなどを挙げている。後学習にはinstruction tuning、reinforcement learning、人間の選好データが使われた。 19

前処理ではdeduplication、`robots.txt`の尊重、品質フィルタ、安全フィルタ、ポルノ・暴力・CSAMなどの除外をGoogleは挙げる。一方、**総学習トークン数、具体的Webサイト構成比、学習データの完全一覧、知識カットオフ、言語別構成比は公開されていない**。この透明性不足は、著作権・データガバナンスを評価する際の重要な限界である。 20

また、ArgonがGoogleの内部事例でA100上のGemma 4 E2Bを約37,900 tokens/sから108,000 tokens/s超へ高速化したという実績があるが、これは**Argon自身がA100上で108K tokens/sで動くという意味ではない**。ArgonがvLLM/Gemma向け最適化コードを生成した結果であり、Argon自身の推論ハードウェアは公開されていない。 21

旧世代・競合モデルとの技術的位置づけ

ユーザー指定のGPT-4o、GPT-4o-mini、Claude 3/4、Llama 3は2024〜2025年世代であり、2026年10月のArgonにとっては「同時代競合」ではなく**歴史的基準点**である。現在の直接競合としてGoogleがモデルカードで採用しているのはGPT-6 Astra、Claude Fable 5.1、Claude Opus 5.5であり、Mistralでは2026年10月6日にMistral Large 4が登場している。 22

比較で見るべき軸は、**①公開／非公開、②総・activeパラメータ、③dense/MoE、④入力モダリティ、⑤コンテキスト、⑥最大出力、⑦ツール利用、⑧エージェント能力、⑨推論レイテンシ、⑩生成スループット、⑪API価格、⑫オンプレ／edge可否、⑬安全性、⑭同一ハーネスでの実務ベンチマーク**である。

モデル	世代・性格	パラメータ/構造	コンテキスト等	位置づけ
Gemini 4 Argon	2026 frontier	sparse MoE、規模非公開	unified 1M、long decode 〜1M	hosted、長時間エージェント重視
Gemini 3.1 Pro	2026旧世代	非公開	約1M入力、64K級出力	Argon前の高性能Gemini
GPT-6 Astra	2026競合	Google比較では規模非公開	—	Argonの直接比較相手
Claude Fable 5.1	2026競合	非公開	1M context、128K max output	長時間agentic coding/knowledge work
Mistral Large 4	2026競合	granular MoE、**1.05T総／52B active**、1.6B vision encoder	1M	open-weight系、Argonより構造公開度が高い
GPT-4o	2024歴史基準	非公開	multimodal	当時GPT-4 Turbo比2倍高速・半額とされた
GPT-4o-mini	2024小型基準	非公開	text/vision	MMLU 82%、当初$0.15/$0.60
Llama 3.1 405B	2024 open-weight基準	405B	128K	重み公開、8言語
Llama 4 Scout	2025 open baseline	17B active、16 experts	10M	単一H100にも適合とMetaが発表
Claude 3/4	2024–25歴史基準	非公開	世代・モデルごとに異なる	2026では後継5.xが直接競合

Argon/Gemini 3はGoogle、GPT-4o系はOpenAI、ClaudeはAnthropic、MistralはMistral AI、LlamaはMetaの一次資料に基づく。 [23]

この表で最も特徴的なのは、ArgonよりMistral Large 4の方がパラメータ透明性で大幅に上回る点である。Mistralは1.05T総／52B activeまで公開する一方、GoogleはArgonについて総数すら公開していない。対してArgonの優位点は、Search、Maps、Computer Use、code execution、Antigravityなど**Googleの巨大な実行環境とモデルを一体化できること**にある。これは純粋なモデル重量比較では捉えにくい競争優位である。 [24]

ベンチマーク、実測性能、レイテンシ、安全性

Googleモデルカードの主比較は、古典的なMMLUだけを並べる方式ではなく、ソフトウェア開発、法務、金融、科学、コンピュータ操作、長文、サイバーセキュリティなど実務型評価を中心としている。2026年10月の同一表では次の結果となる。数値は原則として高いほど良い。 [25]

Benchmark	Argon	GPT-6 Astra	Claude Fable 5.1	Claude Opus 5.5	判定
Vals Index	**68.9**	63.1	65.8	67.0	Argon

Benchmark	Argon	GPT-6 Astra	Claude Fable 5.1	Claude Opus 5.5	判定
AutomationBench	**51.3**	41.4	31.4	42.5	Argon
Vals Finance Agent v2	**65.4**	53.5	58.9	58.6	Argon
Harvey Legal Agent	**19.6**	5.4	6.7	3.8	Argon
DeepSWE v1.1	**77.9**	74.1	67.4	74.2	Argon
FrontierSWE v2	55.0	**65.5**	56.3	62.3	Astra
Vibe Code Bench	**91.9**	89.6	90.3	90.3	Argon
Terminal-bench 4.0	57.4	58.2	57.9	**66.4**	Opus
PostTrainBench	45.3	44.3	40.2	**49.3**	Opus
Terminal-Bench Science 0.1	57.6	**68.1**	52.6	63.3	Astra
LABBench2	**88.8**	85.4	68.6	73.1	Argon
RiemannBench	**76.0**	72.0	65.6	69.6	Argon
GraphWalks ≤128K F1	**99.7**	98.7	91.4	90.6	Argon
GraphWalks 256K–1M F1	**84.2**	71.8	65.0	66.8	Argon
Agent's Last Exam	**39.5**	34.2	—	38.2	Argon
OSWorld-2 partial	69.2	**72.6**	—	—	Astra
Chartography	**71.6**	71.0	46.2	66.3	Argon
LVBench	**91.7**	87.5	79.7	83.7	Argon
CWE-bench v1	**68.0**	**68.0**	58.0	67.0	同点

Googleの2026年10月モデルカード値。集計するとArgonは**13単独首位＋1同点首位／19評価**であり、日本のITmediaも「14項目で首位（同点含む）」と整理している。 4

この結果は強いが、弱点も明瞭である。FrontierSWE v2ではArgon 55.0に対しAstra 65.5、Terminal-bench 4.0ではArgon 57.4に対しOpus 5.5が66.4、Terminal-Bench ScienceではAstra 68.1に対してArgon 57.6である。したがって、**リポジトリ全体のソフトウェア能力とターミナル操作は同じものではなく、ArgonのDeepSWE優位を「すべてのcoding benchmarkで最強」と解釈するのは誤り**である。 26

Gemini 3.8 Flashと同名ベンチマークを比較すると、世代差の方向も見える。ただしGoogleのモデルカード間でハーネスや設定が完全に同一である保証はないため、「厳密なA/B」ではなく参考値として読むべきである。 27

共通評価	Gemini 3.8 Flash	Gemini 4 Argon	差
DeepSWE v1.1	73.7	**77.9**	+4.2 pt
Vals Finance Agent v2	61.4	**65.4**	+4.0

共通評価	Gemini 3.8 Flash	Gemini 4 Argon	差
Harvey Legal	10.0	**19.6**	+9.6
Terminal-bench 4.0	19.1	**57.4**	**+38.3**
LVBench	87.8*	**91.7**	+3.9
LABBench2	86.2	**88.8**	+2.6

*3.8 Flashはagentic設定。 [27]

特にTerminal-bench 4.0の改善が大きい。これはGoogleがGemini 4を「返答モデル」よりも**環境内で操作を継続するエージェント**として重点的に最適化した、という製品設計と整合的である。ただし、絶対値では前述の通りOpus 5.5がなお上回る。 [28]

古典的な「標準NLP/ML」については重要な欠測がある。Gemini 3.1 ProではGPQA Diamond 94.3、Humanity's Last Exam（toolなし）44.4、MMMU-Pro 80.5などが公開されていたが、Argonの公式ローンチ表にはMMLU、GPQA、MMMU-Proなどの直接対応値がない。したがって**Argon対Gemini 3.1 ProのGPQA/MMLU上の世代改善率を現時点で定量化することはできない**。GPT-4o-miniのMMLU 82%など2024年の数字をArgonの2026年エージェント評価と直接並べて優劣を断定することも評価条件が違いすぎる。 [29]

第三者ベンチマークとハルシネーション

独立系Artificial Analysisの公開直後評価は、Googleの主張を部分的に裏付ける一方、より複雑な姿を示している。Argon Highは同機関のIntelligence Indexで53を記録し、AutomationBench-AAでは78%でClaude Sonnet 5.5系の約71%を上回った。一方Terminal Bench 4では約57%で、Claude Sonnet 5.5約64%、Opus 5.5約60%、GPT-6 Astra約59%に届かなかった。これはGoogleモデルカードに見える「ビジネス・長期ワークフローには強いが、一部ターミナル作業は競合優位」という傾向と一致する。 [30]

Valsの独立ランキングでもArgonはVals Index 68.90%で首位となり、Claude Sonnet 5.5 67.04%、Opus 5.5 66.97%などを上回った。Vals Indexはfinance、coding、legal、taxなど経済価値の高い仕事を加重評価するため、Argonが特に専門業務を狙っていることとの相性が良い。 [31]

ハルシネーションではArtificial Analysis Omniscience評価の約15%が注目されたが、これは「事実精度85%」を意味しない。同じ評価でArgonのaccuracyは約50%であり、誤った内容を断言せず回答を控える傾向が低hallucination率に寄与している。GPT-6 AstraよりハルシネーションはArgonの方が正答率は高いというトレードオフも観測された。 [30]

したがって、Redditなどで見られる「Gemini 4がhallucinationを解決した」という表現は不適切である。むしろ**Argonは“知らないことを作る”頻度をかなり抑えたが、その代わり慎重に回答を差し控える場合があるモデル**と理解するのが妥当である。コミュニティでも15%を「solved」と呼ぶことへの反論が出ている。 [32]

レイテンシとスループット

Artificial AnalysisのArgon Highプロバイダー比較では以下が観測された。これはモデル自体とサービング層の双方を含む測定であり、地域、入力長、負荷、thinking levelで変動する。 [6]

指標	Google Vertex	Google AI Studio
生成開始後の出力速度	**116.5 tokens/s**	99.5 tokens/s
First answer tokenまで（約10K input、thinking含む）	約45.55秒	**約44.70秒**
解釈	高い生成throughput	わずかに短い初動

この性質から、Argonは低遅延チャット、音声会話、UIの即時autocompleteより、**30秒〜数分以上かけても正確に完遂したい長い仕事**に適している。Live APIをサポートしていないことも、この用途設計と一致する。[33]

Safety・robustness

GoogleはArgonについて、Frontier Safety Frameworkに沿ったred teamingを実施し、モデルカードではGemini 3.1 Proよりマルチモーダル／多言語安全性を改善し、不必要な拒否を減らしたとしている。画像→テキスト安全性では3.1 Pro比で不安全応答指標を約9.29ポイント改善し、unjustified refusalsも約9.60ポイント低下したと報告している。[34]

よりエージェント固有の評価では、Googleは制約を無断迂回するshortcut attemptが2%、Gray Swan Indirect Prompt Injectionで15回の適応攻撃後も成功率0.7%、外部bio red teamが危険な下位タスクを引き出せた割合が1%未満だったと報告する。また危険要求へのポリシー違反は以前のGeminiに比べほぼ半減したとしている。ただしこれらの多くはGoogleまたはGoogle協力先によるローンチ評価であり、公開後の継続的な第三者追試が必要である。[35]

独立評価、開発者反応、企業導入、評判

主要メディアの評価は「Googleがfrontier競争のトップ層に戻った」という点では概ね一致するが、「Googleのベンチマーク表だけで最強と決めるのは早い」という慎重論も強い。Reutersは世界展開、企業・開発者中心の提供、競争力のある価格を報じる一方、ArgonがDeepMindの遅延や経営陣の変化を経た重要なローンチである点を強調した。The VergeもGoogleの低価格と広い製品統合を評価しつつ、Free/下位プランへの提供は不透明と報じた。Ars Technicaは、10月13日のモデル自体は9月30日に公表されたものと基本的に同じであり、初期性能主張の多くがベンダー提示ベンチマークである点に注意を向けている。[36]

日本では窓の杜が10月14日午前1時31分にグローバル一般提供を報じ、GeminiアプリではPro/Ultraが対象と整理した。ITmediaは初回発表時の19評価を分析し、Argonが14項目で首位または同点首位である一方、Astra／Opusが上回る評価もあると指摘した。PC Watchも日本語で100万トークンと専門業務型ベンチマークを詳しく報じている。[37]

開発者コミュニティについては、**一般公開からまだ数時間程度しか経っていないため、長期運用評価としては信頼できるサンプルが存在しない**。公開前のGoogle AI Developers ForumではArgonへの早期アクセスを求める投稿があり、公開当日には「AntigravityがUltra限定なのか」という不満がすぐに出た。Pro利用者からアクセス拡大を求める投稿もあり、性能そのもの以上に**料金階層・モデル入れ替え・利用制限の分かりにくさ**が初日の評判リスクになっている。[38]

Redditでは、公開前からArgonの到着を待つ投稿が非常に多く、Antigravityクライアント内に `gemini-4-argon-low/medium/high` の文字列が見つかったとの分析も共有された。一方で「frontier」という呼称に対し一部ランキングの順位が低いことを問題視する投稿もあり、コミュニティ評価は熱狂と懐疑の両方が強い。これらは再現性のあるbenchmarkではなく、**世論・開発者心理の観測値**として扱うべきである。[39]

GitHubでは公開前からサードパーティツールがArgon対応を求めるissueを作成していたが、10月14日時点では大規模な本番運用から得られたissue群はまだ形成されていない。また「Gemini 4 Argon Desktop」を名乗り「Official free download」と称するGoogle管理外に見えるリポジトリも検索結果に出現している。これはマルウェアだと断定できないが、**Google公式配布物である証拠はなく、公式API・Antigravity以外から実行ファイルを入手するのは避けるべき**である。 [40]

Stack Overflowでは、本調査時点の公開検索でArgonそのものについて意味のある技術Q&Aの蓄積を確認できなかった。これは否定的評価ではなく、単純に一般API公開直後であるためと考えるのが自然である。Xでも現時点ではGoogle DeepMindやGoogle AI Studio関係者によるローンチ投稿が目立ち、独立開発者による数日〜数週間規模の運用検証はまだ不足している。 [41]

企業・研究現場での実績

Google社内では、Antigravity上のArgon導入後、**agentic code submissionが35%以上増え、rollbackが約3分の1減った**としている。ChromeではChromiumの約4,000万行コードベース内にある約600万行のlegacy領域から機能を分離するリファクタリングを行い、数カ月相当の作業を数週間で進めた。Agent Development KitではArgonが一晩自律動作して見つけた改善が実際にupstreamされ、長いsessionのin-memory preparationを40%、cold importを21%高速化した。 [42]

科学分野では、量子化学パッケージxTBを分子構造予測で最大2.3倍、conformer searchで1.8倍高速化し、量子計算ではAlphaQubit系の内部error-decoderカーネルのthroughputをほぼ倍増させたとGoogleは報告する。またarcsin量子サブルーチンでは、Argon生成・Lean形式検証済みの実装が公表済みCORDIC baseline比で100倍以上のspacetime cost改善を示したとしている。 [43]

企業外の初期事例では、Isomorphic LabsがPDFや特許を大量に統合し、従来数週間かかる情報抽出を数時間に短縮したとGoogleにコメントしている。ServiceNowは自社NowAI-Bench早期評価でGemini 3.8 Flash比およそ20%高いtask successを報告し、GitLabは長時間SDLC、Figmaは複雑なプロトタイピングでの利用を評価した。HubXは独自評価で最優先脆弱性の100%を検出し、検証済みclaimにおける架空脆弱性0%と報告した。いずれも有望だが、顧客コメントまたは共同評価であり、独立監査済み研究ではない。 [44]

Reutersは初期導入企業としてDeloitte、LexisNexis、Ambience Healthcare、Isomorphic Labsなども挙げている。Google CloudページにはAmplitude、Deloitte、楽天のリーダーがArgonの結果を共有した旨も掲示されているため、エンタープライズ評価は限定的な実験段階を越えつつある。ただし、**何千人規模の全社導入でROIが数四半期再現された、といった成熟事例はまだ存在しない**。 [45]

Roboticsも特徴的である。GoogleのERQA（Embodied Reasoning）評価ではArgonが87%を記録したとされ、Spotのmobile manipulation、Real2Sim、zero-shot arm control、simulationを使った自己改善などを実演した。別の実験では人間のcornhole動画から環境と全身運動を推論し、シミュレーションを構築してApptronik Apolloへ移し、実機で7回連続成功させたとしている。ただしこれは**Argonのon-device版がロボットに載っているという意味ではなく、クラウド側の汎用モデルをロボティクススタックの推論・コード生成に利用したデモ**である。 [46]

評判を分析すると、現時点では次のように整理できる。

評判軸	現時点の評価	根拠
フロンティア性能	**高い**	Google同一評価19項目中14項目首位/同点、Vals首位
長時間エージェント	**非常に高い期待**	unified 1M、ADK/Chrome実運用
コーディング	**強いが無敵ではない**	DeepSWE強、FrontierSWE/Terminalでは競合優位

評判軸	現時点の評価	根拠
法務・金融	**特に強い**	Vals Finance、Harvey Legal
ハルシネーション	**大幅改善、未解決**	AA約15%、accuracy約50%
応答速度	**用途依存**	100+ t/sだが初動約45秒
価格競争力	**非常に高い**	導入$2/$10
透明性	**弱い**	params、expert数、学習規模、推論HW非公開
開発者アクセス	**やや混乱**	Pro/Ultra/APIの表記差、初日Forum不満
サイバー安全性	**強力だが高リスク**	Fairwind、guardrail-free版、0.7% IPI
長期信頼性	**未確定**	公開初日で独立本番データ不足

これは一次資料、独立benchmark、主要メディア、公開コミュニティを統合した本調査上の分析であり、Google自身による格付けではない。 [47]

料金、利用枠、データプライバシー、法務・倫理

API価格体系はモデルの魅力の一つである。Google Developer APIではArgonにFree Tierがなく、標準、Batch、Flex、Priorityの複数クラスを提供する。価格は2026年11月30日までプロモーション設定で、12月1日に標準価格へ切り替わる。 [48]

APIクラス	11月30日まで 入力/出力 1M tokens	12月1日以降	備考
Standard	**$2 / $10**	$4 / $20	通常API
Batch	**$1 / $5**	$2 / $10	非同期、大量処理向け
Flex	**$1 / $5**	$2 / $10	柔軟なcapacity
Priority	**$3.60 / $18**	$7.20 / $36	優先サービング
Implicit cached input	約$0.10	約$0.20	標準入力比95%引き
Explicit cache	非対応	非対応	—
Search/Maps grounding	月5,000 queriesまで含む、その後$14/1,000	同体系	Gemini 3+共通枠

Google Gemini API価格表による。 [49]

APIのrequest/token rate limitについては単一の「Argonは毎分X件」という世界共通固定値ではなく、課金tier、project、capacityによって異なり、GoogleはAI Studio内で各projectのactive rate limitsを確認するよう案内している。大規模企業ではProvisioned Throughput等も選択できる。したがって、実運用ではモデル単価だけでなく**同時実行数、long decode占有時間、Priority capacity**を含めた負荷試験が必要である。 [50]

個人向け日本価格について、Google One日本語ページでは**Google AI Proが月額2,900円、Google AI Ultraが月額14,500円から**と案内され、ProではArgonアクセスを段階的に展開、Ultraではより高い利用枠を提供している。ProはGeminiの利用枠を基準比4倍、Ultraは最大20倍など、モデル単位の固定「1日X回」ではなく相対的な利用枠で表現されている。Ultraは150カ国以上、Proも150カ国以上で提供される一方、個別Argon機能は一部国へのrollout条件が付く。 51

ここで非常に紛らわしいのが、**Google AI Proの個人契約とGemini Developer APIのPaid Tierは別物**という点である。API価格ページはArgonについてFree Tierおよび「Google AI Pro plan」では利用不可と表示する一方、Google One側はProでGeminiアプリ／Search／一部AI Studio・Antigravityアクセスを拡大するとしている。したがってPro加入を「無制限のArgon API token entitlement」と解釈してはいけない。 52

プライバシーはConsumerとPaid APIで大きく違う

有料Gemini APIについては比較的明快である。GoogleのZero Data Retention説明は、Paid Servicesに入力したprompt、system instruction、cached content、画像・動画・文書などのfile、およびresponseを**Google製品改善のために使用しない**としている。Argon自体にFree API Tierがないため、通常のArgon Developer API利用ではこの有料サービスのデータポリシーが重要になる。 53

一方、個人向けGeminiアプリでは扱いが異なる。GoogleのGemini Apps Privacy Hubは、収集データの一部をGoogleまたは委託先の訓練済みhuman reviewerが確認することがあり、Googleのサービス／machine-learning technologies改善に使われる可能性があるため、「reviewerに見られたりGoogleによる改善利用を望まない機密情報を入力しない」よう明示的に警告している。ユーザーはKeep Activityを管理・削除できるが、企業秘密・医療・法務・個人情報を扱う組織では**Consumer GeminiとPaid Enterprise/APIを同じリスク水準として扱うべきではない**。 54

また、Argonの事前学習データ自体には公開Web、ライセンスコンテンツだけでなく、規約・privacy controlsに沿ったGoogle製品／サービス由来のuser dataが含まれ得るとモデルカードが記載する。データセットの完全一覧や比率が非公開であるため、データ由来を詳細に監査したい規制産業にとっては透明性上の課題が残る。 55

日本では個人情報保護委員会が生成AI利用について、事業者が個人情報をpromptとして入力する際には、特定した利用目的との整合性やサービス提供者によるデータ利用条件などを確認するよう注意喚起している。Argon固有の規制ではないが、顧客情報、従業員情報、要配慮個人情報を使う法務・金融・医療用途では引き続き重要である。 56

EUではAI ActのArticle 50の透明性義務が2026年8月2日から適用されており、対象となるinteractive/generative AIにはAIと対話していることの通知や、生成・操作されたコンテンツに関するmachine-readable markingなどの透明性要件が課される。ArgonのEU展開はモデル単体だけでなくGeminiアプリ、Search、企業エージェントなど具体的な利用形態ごとに適用関係を検討する必要がある。 57

倫理面で最も重要なのはサイバーのdual-useである。一般版ではpentestingやexploit生成など攻撃性の高い行為を制限する一方、Googleは政府、重要インフラ、セキュリティ組織を審査するFairwind参加者に、サイバーガードレールを外したArgonを提供している。Wizはそれを使い、政府系システムや医療インフラの重大脆弱性を発見・修正したと報告している。この成果は防御価値を示す一方、アクセス審査が破られた場合のモデル能力リスクも大きく、Google自身が初期一般公開を遅らせた理由でもある。 58

さらにGoogleは、エージェントがpermission errorなどに遭遇した際に無断回避しないよう学習させるだけでなく、Argonのstep-by-step reasoningとactionを読むautomated misalignment monitorを内部展開している。これは高度エージェントの安全策として重要である一方、企業側ではagent log、tool call、アクセス

権限、機密データの記録・保存ポリシーまで含めてガバナンス対象とする必要がある、というのが本調査からの実務的な含意である。 59

総合評価、推奨ユースケース、弱点と未確定事項

Gemini 4 Argonを一言で評価するなら、**「高速チャットをさらに賢くしたモデル」ではなく、「数十分〜数時間続く高度なデジタル仕事を任せるためのGoogle製フロンティア・エージェント基盤」**である。1M unified token budget、long decode、native multimodality、Google Search/Maps grounding、code execution、Computer Use、Antigravity、多数の企業ツールを組み合わせた設計を見ると、その方向性は明確である。 60

最も適した用途は、大規模コードベースの理解・移行・リファクタリング、長期間のagentic coding、複数資料を横断する法務・金融調査、巨大PDF・表・画像混在資料の統合、科学計算・アルゴリズム最適化、検証可能なサイバー防御、動画／3Dを伴うロボティクス研究、数十段階の企業業務自動化である。Chrome、ADK、Isomorphic Labs、ServiceNow、xTBなどの初期事例はこの方向を支持する。 61

反対に、**低遅延の会話UI、音声ボット、単純分類、数行のコード修正、大量の短文変換**ではArgonは過剰である可能性が高い。Argon Highは回答開始まで40秒台を要する独立測定があり、Live APIも非対応である。こうした用途ではFlash系、小型Mistral、Llama、GPT mini系などの方がコスト・体感速度に優れる場合がある。 62

オープンウェイト／オンプレが必須ならArgonは第一候補ではない。パラメータもweightsも公開されず、edge deployment targetも公表されていない。Mistral Large 4は1.05T総／52B activeという構成を公開し、Llama系もローカルGPU・edge向け展開を明示しているため、データ主権や推論インフラ制御ではそれらに優位性がある。 63

一方、Google Cloud、Search、Workspace系サービス、Maps、Android Studio、Antigravityなどを既に利用する組織にとっては、Argonの最大のmoatはモデル単体のbenchmarkではなく**Googleのデータ・検索・開発・エージェント実行基盤を一つのモデルで横断できること**にある。ここはMMLUの数ポイント差より実務ROIに影響する可能性が高い。Google社内のコードsubmission増加やServiceNowのtask success改善は、その仮説を支持する初期証拠ではあるが、まだ長期的・独立な検証ではない。 44

現時点で最も重要な**未確定事項**は、総／activeパラメータ数、MoE expert数、学習トークン数、正確なknowledge cutoff、学習計算量、TPU世代と規模、推論ハードウェア、quantization／数値精度、GPU/Edge用weights、国別Pro rollout、実際のアカウント別Argon利用回数上限である。またMMLU、GPQA、MMMU-Proなど従来型academic benchmarkのArgon値も公式ローンチ資料にはなく、第三者による大規模なcoding benchmark再現、長時間agent failure rate、本番SLA、concurrency、数百万リクエスト規模のlatency distributionもまだ不足している。 64

したがって現時点の最終評価は、**強み：長期推論、専門業務、長文、マルチモーダル、agentic execution、安全性改善、価格、Googleエコシステム統合。弱み：初動レイテンシ、完全ではないcoding首位、閉鎖的な技術情報、Consumer/API料金体系の複雑さ、公開直後ゆえの実運用証拠不足、dual-use cyber risk**となる。Googleが「最も難しい問題を解く」モデルと位置付けることには相応のデータがある一方、現時点で「すべてのLLMを全面的に上回る」と評価する証拠はない。 65

特に実務導入では、**Argonを全リクエストの既定モデルにするより、Flash／小型モデルで処理できない高価値・長時間・高複雑度タスクにルーティングする階層構成**が合理的である。Argonの$2/$10という導入価格は魅力的だが、long decodeでは大量のthinking/output tokenを消費し得るため、モデル単価だけでなく完了タスク当たりのコスト、平均時間、成功率、人間による手戻り率で評価すべきである。Anthropic自身もモ

デル選択では実際のprompt/dataを使った用途固有benchmarkを推奨しており、Argonについても同じ原則が当てはまる。 66

総じて、2026年10月14日時点のGemini 4 Argonは、**Googleがフロンティア AI競争で再び最上位集団に入ったことを示す有力なモデルであり、特に「長く働くAI」という軸では市場でも最も野心的な製品の一つ**である。ただし、一般公開からまだ極めて短時間しか経っておらず、Google／早期顧客の成功事例と第三者の長期本番評価を明確に区別して読む必要がある。今後の評価を決定づけるのは追加のベンチマークよりも、**数週間〜数カ月の実運用で、100万トークン級の長時間エージェントがどの程度安定して正しい仕事を完遂し、どれほどのコストと人間監督を必要とするか**である。 67

navlistGemini 4 Argonの主要な最新報道turn18news34,turn18news35,turn18news36

---

1 7 8 11 14 21 28 35 42 43 58 59 61 65 Gemini 4 Argon is now available for everyone
https://deepmind.google/blog/introducing-gemini-4-argon/

2 16 18 33 60 Gemini 4 Argon | Gemini API | Google AI for Developers
https://ai.google.dev/gemini-api/docs/models/gemini-4-argon

3 4 9 15 19 20 22 25 26 34 55 64 Gemini 4 Argon - Model Card — Google DeepMind
https://deepmind.google/models/model-cards/gemini-4-argon/

5 30 47 Gemini 4 Argon: Google is back as one of the top three labs in intelligence achieved | Artificial Analysis
https://artificialanalysis.ai/articles/gemini-4-argon-google-top-three-labs

6 62 Gemini 4 Argon (high): API Provider Performance Benchmarking & Price Analysis | Artificial Analysis
https://artificialanalysis.ai/models/gemini-4-argon/providers

10 Gemini 4 Argon：フロンティア知能の新時代
https://blog.google/intl/ja-jp/company-news/technology/gemini4argon/?utm_source=chatgpt.com

12 Gemini 4 Argon is now available in the Gemini app
https://blog.google/innovation-and-ai/products/gemini-app/gemini-4-argon-availability/

13 Google Search - A new kind of help
https://search.google/

17 What's new in Gemini 4 Argon | Gemini API | Google AI for Developers
https://ai.google.dev/gemini-api/docs/latest-model

23 29 Gemini 3.1 Pro - Model Card — Google DeepMind
https://deepmind.google/models/model-cards/gemini-3-1-pro/

24 63 Mistral Large 4
https://docs.mistral.ai/models/mistral-large-4-0?utm_source=chatgpt.com

27 Gemini 3.8 Flash - Model Card — Google DeepMind
https://deepmind.google/models/model-cards/gemini-3-8-flash/

31 Vals Index Leaderboard and Methodology
https://www.vals.ai/benchmarks/vals_index?utm_source=chatgpt.com

32 Gemini 4 Argon solved hallucinations. : r/singularity
https://www.reddit.com/r/singularity/comments/1wuj72j/gemini_4_argon_solved_hallucinations/?utm_source=chatgpt.com

36 45 67 Google launches flagship AI model Gemini 4 Argon for developers, businesses
https://www.reuters.com/business/google-launches-flagship-ai-model-gemini-4-argon-developers-businesses-2026-10-13/?utm_source=chatgpt.com

37 「Gemini 4 Argon」が一般提供開始 〜満を持して登場した最強 ...
https://forest.watch.impress.co.jp/docs/news/2147842.html?utm_source=chatgpt.com

38 Early Access Request: Gemini 4 Argon – Building an Agentic ...
https://discuss.ai.google.dev/t/early-access-request-gemini-4-argon-building-an-agentic-technical-engineering-knowledge-adaptive-assessment-platform/187425?utm_source=chatgpt.com

39 Antigravity 2.23.0 already contains "Gemini 4 Argon (Low/ ...
https://www.reddit.com/r/GeminiAI/comments/1x4g2e2/antigravity_2230_already_contains_gemini_4_argon/?utm_source=chatgpt.com

40 Gemini 4 Argon: role/model-tier integration when API exposes it
https://github.com/sleegme/native-gravity/issues/74?utm_source=chatgpt.com

41 Logan Kilpatrick (@OfficialLoganK) ...
https://x.com/OfficialLoganK?utm_source=chatgpt.com

44 Gemini 4 Argon available on Google Cloud | Google Cloud Blog
https://cloud.google.com/blog/products/ai-machine-learning/gemini-4-argon-available-on-google-cloud

46 Gemini 4 Argon for Robotics - Google Developers Blog
https://developers.googleblog.com/gemini-4-argon-for-robotics/

48 49 52 66 Gemini Developer API pricing | Gemini API | Google AI for Developers
https://ai.google.dev/gemini-api/docs/pricing

50 Rate limits | Gemini API - Google AI for Developers
https://ai.google.dev/gemini-api/docs/rate-limits?utm_source=chatgpt.com

51 Google AI のプラン（クラウド ストレージ付き）
https://one.google.com/intl/ja_jp/about/google-ai-plans/?utm_source=chatgpt.com

53 Zero data retention in the Gemini Developer API
https://ai.google.dev/gemini-api/docs/zdr?utm_source=chatgpt.com

54 Gemini Apps Privacy Hub
https://support.google.com/gemini/answer/13594961?hl=en&utm_source=chatgpt.com

56 生成AIサービスの利用に関する注意喚起等について
https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/?utm_source=chatgpt.com

57 Transparency obligations for AI providers and deployers
https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-transparency-obligations?utm_source=chatgpt.com