SpaceXAI Grok 4.7の推論アーキテクチャと現場運用の実効性検証

Gemini 3.8 Flash

SpaceX傘下の人工知能開発部門SpaceXAI（旧xAI）が2026年9月21日（日本時間9月22日）に公開した「Grok 4.7」は、最先端フロンティアモデルの知能限界そのものを押し上げる存在ではなく、確立された高度な推論・コーディング能力を前世代と同一の低価格で大量供給する「量産型ワークホース」である[1]。前世代「Grok 4.6」のトークン単価（入力100万トークンあたり2ドル、出力6ドル）を据え置きながら、基盤モデルを公称2.1兆パラメータへと約40%スケールアップし、長時間タスクに焦点を当てた強化学習パイプラインを組み込んでいる[1]。

技術的な位置付けとして、Grok 4.7はソフトウェア開発実務を評価するDeepSWE v1.1（71.0%）や対話型開発環境でのCursorBench 4.0（46.3%）において米OpenAIのGPT-5.6 Sol（41.7%）を明確に上回り、電気工学（EEBench：64.0%）や法務分析（Harvey Legal：19.6%）といった専門分野の静的タスクでは業界最高水準のスコアを記録している[1]。さらに、同社が2026年6月に600億ドルで買収した開発環境「Cursor」をはじめ、マイクロソフトの「GitHub Copilot」など主要開発基盤へ即日配備され、実務ワークフローへの直接的な浸透を図っている[2]。

しかし、コマンドライン環境での完全自律実行を評価するTerminal-Bench 4.0（38.0%）では米AnthropicのClaude Fable 5.1（57.9%）や米OpenAIのGPT-6 Astra（60%）に大きく水をあけられており、総合知能指数（Artificial Analysis Intelligence Index：46）でも競合フロンティアの「53」に対して明確な性能差が残る[1]。加えて、複雑な思考プロセスに伴う出力トークンの過剰消費（1タスク平均約81,000トークン）により、長大な自律セッションでは名目上の低単価ほど実効費用が下がらない構造的課題も確認されている[3]。エンジニアリング組織における意思決定の結論として、本モデルは「完全自律型エージェント」としてではなく、人間がループ内に介在する「対話型高スループット・コパイロット」として配置した際に最も確実な投資対効果を発揮する[3]。

2.1兆パラメータ基盤と長時間強化学習のメカニズム

Grok 4.7の基盤設計における最大の更新は、パラメータ規模の大幅な拡張と、数時間にわたる多段階タスクの完遂を目的とした事後強化学習（RL）の再設計にある[1]。

基盤モデルのパラメータ数は、前世代Grok 4.6の1.5兆から約40%拡大された2.1兆（2.1T）に達する[4]。事前学習およびドメイン適合の過程では、SpaceXが保有する航空宇宙工学、熱流体力学、軌道計算などの実務エンジニアリングデータが大規模に投入されたほか、買収したAnysphereから得られた匿名化済みのCursor実開発ワークフローデータが補完的に学習された[2]。これにより、単一関数の補完にとどまらず、リポジトリ全体の広域な依存関係を把握し、物理的・論理的整合性を検証する基礎推論能力が強化された[1]。

事後学習においては、完了までに長時間を要する高難度タスクの比率を高めた長期間の強化学習が実行された[1]。従来の推論モデルが直面していた最大のボトルネックは、数十ステップに及ぶ作業の初期段階で生じた微細なエラーが後続処理で増幅され、最終的に致命的な破綻を招く現象であった[11]。Grok 4.7では、ファイル変更の適用や単体テストの成否をモデル自身が中間検証し、誤った仮説を自律的に破棄・再試行する「自己検証（Self-Verification）」能力の獲得に学習リソースが集中配分されている[1]。

さらに、SpaceXAIの常時稼働型エージェント基盤「Grok Bot」の実行ハーネスをモデル訓練レベルでネイティブに学習させた点が特徴的である[1]。通常の外部ツール連携モデルでは、システムプロンプトとして大量のJSONスキーマやツール記述子をコンテキストへ都度注入する必要があり、これがトークン消費の肥大化と状態追跡のノイズとなっていた[11]。Grok 4.7は、ファイル走査、差分編集、シェルコマンド実行、ブラウザ操作といった標準的エージェントツール群を重み内部で直接解釈するため、推論実行時のプロンプトオーバーヘッドを削減し、長時間の対話セッションにおけるコンテキスト汚染を構造的に軽減している[11]。

仕様項目	技術的諸元および運用プロファイル
モデル識別名	grok-4.7 [cite: 5, 13]
パラメータ規模	2.1兆パラメータ（前世代1.5T比で約40%拡大）[4]
コンテキスト長	標準256,000トークン / 最大500,000トークン[5]
ナレッジカットオフ	2026年5月[5]
モダリティ	テキスト・画像入力（最大20MiB、JPEG/PNG対応）／テキスト出力[5]
推論制御（**Reasoning Effort**）	low, medium, high（デフォルト）, xhigh（完全オフは不可）[5]
暗号化推論返却	Responses APIにおいて reasoning.encrypted_content を常時自動返却[5]
非対応パラメータ	presencePenalty, frequencyPenalty, stop（指定時はHTTP 400エラー）[5]

開発運用の観点において注意すべきは、推論APIのプロトコル仕様である。Responses API（POST /v1/responses）を利用する際、Grok 4.7はリクエストパラメータの設定にかかわらず、推論トレースを暗号化文字列 reasoning.encrypted_content として必ずレスポンスに含めて返却する[5]。ステートレスなHTTP環境で複数ターンの対話を構築する場合、クライアント側はこの暗号化オブジェクトを改変せずに次リクエストの入力ペイロードへ戻す必要があり、これを怠ると思考コンテキストが失われ、推論品質の大幅な低下やキャッシュミスを招く設計となっている[5]。

主要ベンチマークの定量的検証とフロンティアモデルとの性能

境界

SpaceXAIの公式データおよび第三者評価機関Artificial Analysisの検証結果を総合すると、Grok 4.7は特定ドメインにおける突出した強みと、汎用自律処理における明確な弱点が同居する極端な性能プロファイルを示している[1]。

評価領域・ベンチマーク指標	Grok 4.7 (xHigh)	Grok 4.6 (High)	GPT-5.6 Sol (Max)	Claude Fable 5.1 (Max)	GPT-6 Astra (Max)
CursorBench 4.0（IDE内コーディング）[1]	**46.3%**	40.4%	41.7%	**51.8%**	未公表
DeepSWE v1.1（実務ソフトウェア工学）[1]	**71.0%**	65.2%	**72.7%**	70.0%	未公表
EEBench（電気工学・回路設計）[1]	**64.0%**	53.0%	39.4%	56.4%	未公表
AA Briefcase v1.1（総合オフィス業務・Elo）[1]	**1,657**	1,546	1,487	**1,678**	未公表
Terminal-Bench 4.0（CLI自律実行）[1]	**38.0%**	20.3%	37.3%	**57.9%**	60.0%[3]
Harvey Legal（法務エージェント業務）[1]	**19.6%**	15.8%	2.5%	6.7%	未公表
HealthBench Prof.（臨	**56.7%**	48.5%	60.5%	**62.1%**	未公表

床医学推論）[1]					
GDPval（専門職業務品質・Elo）[1]	**1,695**	1,605	未公表	**1,735**	1,542
AA Intelligence Index v4.3.2 [cite: 3, 8]	**46**	未計測	未公表	**53**	**53**
入力単価（**/1M tokens**）[cite: 1]	**$2.00**	$2.00	$4.00	$10.00	未公表
出力単価（**/1M tokens**）[cite: 1]	**$6.00**	$6.00	$20.00	$50.00	未公表

ソフトウェア開発領域において、Grok 4.7は日常的な実装業務に対して極めて高い競争力を提示している。Cursorエディタ内での実際のコーディング支援を模したCursorBench 4.0では46.3%を達成し、OpenAIのGPT-5.6 Sol（41.7%）を凌駕した[1]。GitHubの実課題を解決するDeepSWE v1.1においても71.0%を記録し、Claude Fable 5.1（70.0%）をわずかに上回る水準へ到達している[1]。特に電気工学評価（EEBench）では64.0%をマークし、競合全モデルを大幅に引き離した[1]。SpaceXの航空宇宙・ハードウェア開発資産に由来するデータセットの注入が、厳密な数理・物理制約を伴う工学計算タスクにおいて直接的な成果を挙げている[1]。

専門職の定型業務においても強力な結果が確認されている。法務エージェントタスクを評価するHarvey Legal Agent Benchmarkにおいて、Grok 4.7は19.6%を記録した[1]。これはClaude Fable 5.1（6.7%）の約3倍、GPT-5.6 Sol（2.5%）の約8倍に相当し、長大な契約条項の照合や法令リサーチの精度において明確な比較優位を示している[1]。複数週にわたるオフィスドキュメント作成やデータ分析を評価するAA Briefcase v1.1でもスコア1,657（前世代から111ポイント向上）を獲得し、GPT-6 Astra（1,542）を上回る実務処理能力を示した[1]。

しかし、自律型CLI操作を評価するTerminal-Bench 4.0は、本モデルの構造的限界を浮き彫りにしている。Grok 4.7のスコアは38.0%にとどまり、Claude Fable 5.1の57.9%やGPT-6 Astraの60%に対して約20ポイントの開きが存在する[1]。人間が介在する対話環境では高い成功率を示すものの、シェル環境内で数十回のコマンド実行を連鎖させ、エラー出力を解析しながら自律的に環境状態を復旧させるタスクでは、途中で前提条件を取り違えてループに陥る傾向が顕著である[3]。

この自律タスクの弱さは、総合指標であるArtificial Analysis Intelligence Index（v4.3.2）の総合スコ

アにも反映されている。同指数においてGrok 4.7は「46」にとどまり、Claude Fable 5.1およびGPT-6 Astraの「53」に対して明確な下位に位置する[3]。単一のコーディングや定型ナレッジワークではフロンティア級の出力を示すものの、多段階の自律エージェント推論を包括した総合知能としては、依然としてトップラボの最上位モデルに対して一段の隔たりが存在する[3]。

料金体系の構造とエコシステム配備の全貌

SpaceXAIは、Grok 4.7の展開において前世代Grok 4.6と同一の価格設定を維持し、フロンティア市場における価格破壊を戦略の前面に押し出している[1]。

標準APIにおける利用料金は、入力100万トークンあたり2.00ドル、出力100万トークンあたり6.00ドルである[5]。Claude Fable 5.1（入力10ドル／出力50ドル）と比較して入力で5分の1、出力で約8分の1という設定であり、GPT-5.6 Sol（入力4ドル／出力20ドル）に対しても半額以下の水準を保つ[1]。また、プロンプトキャッシュがヒットした入力トークンは0.50ドル（75%割引）で処理される[5]。

利用ティア・エンドポイント	入力単価（**/1M tokens**）	キャッシュ入力単価	出力単価（**/1M tokens**）	適用条件・運用仕様
標準短文（**< 200k tokens**）[cite: 5]	$2.00	$0.50	$6.00	グローバルエンドポイント標準枠
標準長文（**≥ 200k tokens**）[cite: 5]	$4.00	$1.00	$12.00	プロンプトが20万トークン以上で全量に適用
US地域限定（**< 200k tokens**）[cite: 5]	$2.20	$0.55	$6.60	全推論を米国内リージョンで完結（10%割増）
US地域限定（**≥ 200k tokens**）[cite: 5]	$4.40	$1.10	$13.20	米国内データ主権要件対応（10%割増）
Grok 4.7 Fast（短文）[cite: 5, 13]	$4.00	$1.00	$12.00	Cursor / Grok Build専用、生成速度2倍
Grok 4.7 Fast（長文）[cite: 5, 13]	$6.00	$1.50	$18.00	20万トークン以上の長文セッション

インフラアーキテクチャの設計上、留意すべきは長文コンテキストにおける価格倍増トリガーである。リクエスト内のプロンプト長が200,000トークンに達した時点で、当該リクエストの全トークンに対して長文レート（入力4.00ドル／出力12.00ドル）が適用される[5]。さらに、金融機関や官公庁などのデータガバナンス要件に対応するため、すべての推論処理を米国内インフラに限定するUS Regional Endpoint（https://us.api.x.ai/v1）が開設されており、標準単価の10%増しで提供されている[5]。
スループット性能に関して、推論エフォート「high」設定時におけるGrok 4.7の出力速度は毎秒55トークン（55 t/s）、最初の回答トークンが返却されるまでのレイテンシ（Time to First Answer Token）は0.84秒を記録している[8]。「xhigh」設定時では内部の思考ステップ増加に伴い出力速度は39.3 t/s、レイテンシは0.88秒となる[8]。加えて、待ち時間の削減を最優先する開発者向けに、通常インフラの2倍の生成速度を提供する「Grok 4.7 Fast」がCursorおよびGrok Build環境限定で提供されている[5]。
エコシステムへの展開は、開発者接点を最短で制圧する方針のもとで設計された。SpaceXが買収したCursorでは全プランの標準モデルプールに即日組み込まれ、マイクロソフトのGitHub Copilotにおいても VS Code、Visual Studio、JetBrains、Xcode、CLI環境向けに即日利用可能となった[7]。さらに、Vercel AI Gateway、OpenRouter、Cloudflareなどの主要ルーティング層に即時統合され、Vercelでは提供開始直後に40%割引（入力1.20ドル／出力3.60ドル）の大規模プロモーションが展開された[5]。エンドユーザー向けには、X（旧Twitter）のWeb/モバイルアプリ、スタンドアローンのGrokアプリ、さらにはテスラ車載システムの音声対話ナビゲーションへも順次バックエンドとして配備されている[2]。

防御的セーフガードスタックと高リスク領域の適合性

Grokシリーズは従来、規制的ガードレールの少ない自由度の高いアライメントを売りにしてきたが、Grok 4.7では安全評価機構が完全に再設計された[1]。過剰拒否（False Refusal）による開発生産性の低下を抑えつつ、破滅的な二重用途（Dual-use）リスクを厳格に遮断するキャリブレーションが施されている[1]。
バイオセキュリティ領域において、Grok 4.7はLatchBioのバイオセーフティベンチマークで62.4%の適合率を記録し、主要フロンティアモデルの中で首位を獲得した[1]。病原体の兵器化プロトコルや危険な合成手順に関する不正プロンプトを確実に拒絶する一方、創薬支援、タンパク質フォールディング解析、正常な生物学研究に関する問い合わせには十全な推論能力を提供するバランスを確立している[1]。
サイバーセキュリティ領域では、社内の悪意あるサイバータスク評価「HackerBench v0.3」において、危険な二重用途プロンプトの誤通過率をわずか3.3%に抑え込んだ[1]。ゼロデイ脆弱性の悪用コード自動生成やマルウェア難読化といった攻撃的意図の96.7%を確実に遮断する[1]。その一方で、ペネトレーションテストのスクリプト作成やパッチ適用、脆弱性修正といった正当な防御的セキュリティ業務を誤って拒否する事態を最小限に抑えている[1]。SpaceXAIはこの防御的性能をテコに、選定された一部のセキュリティパートナー企業に対し、Grok 4.7のレッドチーム評価機能を招待制で先行提供している[1]。

開発現場における初期評価と実運用上の構造的トレードオフ

現場のエンジニアリング組織における初期運用結果は、短時間セッションにおける高い費用対効果を称賛する声と、長大な自律エージェント運用における経済性・信頼性の乖離を指摘するシビアな見

解に二極化している[3]。

高スループット対話環境での実効的ROI

CursorやGitHub Copilotの現場開発者からは、インラインコード生成や単一機能のテスト作成、小規模なリファクタリングにおいて、圧倒的な速度とコスト効率が高く評価されている[3]。競合フロンティアモデルを使用する場合に発生していたAPIトークン費用の懸念が大幅に緩和されたため、開発者が躊躇なく大量のコード生成や多段階の推論を実行できる環境が整った[1]。実例として、Blenderで構築された3Dシーンからインタラクティブなthree.js環境を構築するタスクにおいて、Claude Opus 5が約2.05ドルのAPIコストを要したのに対し、Grok 4.7は約0.20ドル（10分の1）のコストで迅速にコードを出力した事例が報告されている[4]。

トークン過消費（Token Bloat）による実質コストの上昇

一方で、クラウドインフラアーキテクトからは、名目トークン単価の低さだけで実効コストを判断することに対する強い警告が発せられている[3]。Artificial Analysisのタスク別消費データによると、Grok 4.7（xHigh）はIntelligence Indexの1タスクを完了するにあたり、平均して約81,000の出力トークンを消費している[4]。これはGrok 4.6（High）の約36,000トークン、GPT-6 Astra（Max）の約27,000トークンと比較して約2.2倍から3倍のトークン消費量に達する[4]。

この現象は、Grok 4.7が解を導出する過程で内部の思考トレースを極めて冗長に出力し、かつ自己検証の各ステップを詳細に記述する推論アルゴリズムに起因する[4]。その結果、Artificial Analysisが計測したタスクあたりの加重平均実効コスト（Weighted Cost per Task）は、Grok 4.7（High）で2.73ドル、Grok 4.7（xHigh）で3.74ドルに達する[8]。名目上のトークン単価はClaude Fable 5.1の数分の一であっても、消費トークン数が2〜3倍に膨張するため、1つの複雑なプルリクエストを完遂する実質コストの差は大幅に縮小する[3]。プロンプトが200,000トークンを超えて長文倍額レートが適用される大規模コードベースの処理では、このトークン過消費が総請求額をさらに押し上げる要因となる[5]。

自律エージェントの壁と状態維持の破綻

技術系メディア『The New Stack』などの詳細な検証では、長時間自律タスクにおける「エージェント・スタミナの限界」が報告されている[11]。SpaceXAIは数時間に及ぶタスクを前提にモデルを訓練したとしているが、Terminal-Bench（38.0%）の低スコアが示す通り、複数ステップの自律実行タスクでは依然として過半数が失敗に終わる[1]。

失敗のメカニズムとして、実行途中で遭遇したビルドエラーや環境不整合に対してモデルが不適切な修正を試み、その失敗履歴がコンテキスト内に蓄積されることでコンテキスト長を急速に圧迫し、最終的に指示の根本を見失って無限ループに陥る「エラーの連鎖」が挙げられている[11]。人間が介在して適宜プロンプトで軌道修正を行う対話環境では極めて有能であるものの、夜間にバックグラウンドで完全自律稼働させて複数の不具合を修正させる用途では、Claude Fable 5.1のような高度な文脈保持力を持つモデルに対して信頼性で明確に劣る[3]。

Cursorエコシステムの統合動向と将来ロードマップ

SpaceXによるCursor（Anysphere）買収以降、Cursorコミュニティ内部では、旧来の独自モデルであった「Composer」系統（Composer 2.5等）の開発継続性に対する懸念が広がっている[2]。Reddit等では、イーロン・マスク氏の意向によってCursorの基盤モデルがすべてGrokブランドへ吸収・一本化されるのではないかという議論が続いている[15]。

マスク氏自身は、Grok 4.7の性能位置付けについて「Claude Opus 5.0と同等の水準であり、Opus 5.1には及ばない。マルチモーダル性能にはまだ改善の余地がある」と客観的な限界を公言している

[4]。その上で、同社はすでに2.5兆パラメータ規模の次期モデル「Grok 4.8」の事前学習を完了して強化学習に移行しており、将来的には「Grok 4.9」でAstra/Fable級へ到達し、「Grok 5」で業界完全首位を獲得するという長期開発計画を提示している[6]。
総括として、Grok 4.7はフロンティアモデルの最高知能を塗り替えた革新ではなく、日常の開発実務で要求される高いコーディング・専門業務推論を、圧倒的な低単価と主要IDEへの直接統合によってコモディティ化した実利的なシステムである[3]。完全自律型CLI環境における状態喪失や思考トークンの過剰消費といった明確な制約が存在するため、エンジニアリング組織は本モデルを「完全自律型エージェント」として過信することなく、人間が介在する反復的コーディング支援やバッチ処理パイプラインの高速推論エンジンとして適材適所で組み込むことが、最大の経済的・技術的リターンを獲得する要件となる[2]。

引用文献

1. Introducing Grok 4.7 - SpaceXAI, https://x.ai/news/grok-4-7
2. SpaceXAI、「Grok 4.7」発表 コーディングと知識労働向けで同社, https://www.itmedia.co.jp/news/article/2609/22/2000001668/
3. Grok 4.7 Pricing vs GPT-6 Astra, Claude Fable 5.1 - shattered.io, https://shattered.io/grok-4-7-launch-pricing-benchmarks-gpt-6-2026/
4. Musk's Grok 4.7 Arrives After Months of Delays, Undercutting Rivals, https://finance.biggo.com/news/348590e8-33a0-4266-94b5-43d11611cd84
5. Grok 4.7 | SpaceXAI Docs, https://docs.x.ai/developers/grok-4-7
6. xAI Releases Grok 4.7 AI Model - Android Headlines, https://www.androidheadlines.com/2026/09/grok-4-7-ai-launch-coding-upgrades-pricing.html
7. Grok 4.7 is now available in GitHub Copilot, https://github.blog/changelog/2026-09-21-grok-4-7-is-now-available-in-github-copilot/
8. Grok 4.7: Release Intelligence, Performance & Price, https://artificialanalysis.ai/models/releases/grok-4-7
9. SpaceX傘下xAI、コーディング特化「Grok 4.7」を公開...価格, https://finance.biggo.jp/news/3d270193-e716-4ea1-a541-3eb1eddb6e44
10. Model Card: Grok 4.7 - SpaceXAI, https://media.x.ai/v1/website/4p7card-5eccc980.pdf
11. Grok 4.7 was built to work for hours. It still fails most of the time., https://thenewstack.io/grok-4-7-agent-stamina/
12. 「Grok 4.7」登場、長時間タスクも安定処理 価格はSolの半分以下, https://www.watch.impress.co.jp/docs/news/2142588.html
13. Grok 4.7 | Cursor Docs, https://cursor.com/docs/models/grok-4-7
14. Grok 4.7's 80% Discount Squeezes Coding Rivals [2026] - shattered.io, https://shattered.io/grok-4-7-price-war-coding-market-2026/
15. Introducing Grok 4.7 : r/cursor - Reddit, https://www.reddit.com/r/cursor/comments/1wmgn3m/introducing_grok_47/
16. After agreeing with Anthropic CEO Dario Amodei on slowing pace of AI, Sam Altman and Elon Musk teased next powerful model days later,

https://timesofindia.indiatimes.com/technology/tech-news/after-agreeing-with-anthropic-ceo-dario-amodei-on-slowing-pace-of-ai-sam-altman-and-elon-musk-teased-next-powerful-model-days-later/articleshow/134362820.cms