ChatGPT

SpaceXAI「Grok 4.6」深掘り調査レポート

エグゼクティブサマリー

SpaceXAI（法的には X.AI LLC の DBA＝営業上の名称）は、**2026年8月12日**に「Grok 4.6」を発表した。位置づけは汎用チャットモデルの単純な世代更新というより、**長時間・多段階のエージェント処理、ソフトウェア開発、調査・知識労働、CADや視覚的なアプリ開発**を重点強化したフロンティアモデルである。発表当日から SpaceXAI API、Grok Build、Cursor、OpenRouter、Vercel、Cloudflareなどで利用でき、標準API価格は**100万トークン当たり入力2ドル／出力6ドル**。ただし、**20万トークン以上のlong-context tierでは入力4ドル／出力12ドルへ倍増**し、キャッシュ入力も0.50ドルから1ドルになる。高速版は標準版の2倍価格とされる。 [1]

実務上の最大の魅力は、**「フロンティア級のエージェント性能を比較的低い単価で提供する」価格性能比**にある。独立評価機関Artificial Analysisでは Intelligence Index が**61**で、Grok 4.5の56から5ポイント上昇し、OpenAI GPT-5.6 Solと同水準、Claude Fable 5の62とOpus 5の63にはわずかに及ばない。同機関の実測では評価タスク当たりコストは0.84ドルで、Grok 4.6を価格性能のPareto frontier上に位置付けている。 [2]

一方、**「最強」「最も幻覚が少ない」と単純化するのは適切ではない**。SpaceXAI自身のモデルカードでは、DeepSWE 1.1が65.9%に対しGPT-5.6 Solは73%、Terminal-Bench 3.0は26.0%対34.6%であり、複数の重要なソフトウェア・ターミナル系評価では競合に負ける。さらに同社のHallucination評価ではGrok 4.6の幻覚率は**1.7%**で、Grok 4.5の**0.98%**、GPT-5.5の**1.1%**より悪化している。つまり、4.6の進歩は主として**長期タスク遂行、エージェント性能、検索統合、コード／オフィスワーク**に現れており、全指標での全面的改善ではない。 [3]

技術的透明性は依然限定的である。公式モデルカードはGrok 4.6を**「1.5T-scale model family」**の一員と説明するものの、これが正確に1.5兆パラメータを意味するのか、総パラメータ数なのか、active parametersはいくつか、層数・expert数・attention構成は何か、といった詳細は**未指定**である。モデル自身による推論最適化実験で「fused MoE」が明記されるため、MoE系アーキテクチャである可能性は極めて高いが、公式文書はexpert構成を公開していない。Grok 4.6そのものの学習GPU数・総FLOPs・推論時GPU要件も**未指定**であり、前世代Grok 4.5の「数万台のNVIDIA GB300で学習」という情報を4.6へそのまま転用すべきではない。 [4]

セーフティ面では、SFT、**RLHF**、検証可能報酬、model-based grading、システムプロンプト、実行時フィルタを組み合わせた多層防御が明記され、標準jailbreak成功率は0.73%から0.04%へ改善し、CBRN拒否精度も100%に達した。一方、モデルカードには自己傷害関連、dishonesty、sycophancy、サイバー有害要求への応答など一部指標の悪化も記録されている。したがって「セーフガードは強化された」は概ね正しいが、**安全性が一様に向上したとは公式データ自身からも言えない**。 [5]

企業導入では、API入力・出力は明示的許可なしに学習へ使わないとSpaceXAIは表明しているものの、通常設定では**30日間保存**される。ZDR（Zero Data Retention）ではプロンプト／出力を永続保存しないが、Files、Collections、stateful Responses、Batch APIなど保存依存機能が使えなくなる。Enterprise Termsでは出力の権利は顧客側に帰属する一方、出力の正確性保証はなく、人間による検証を顧客の責任としている。また、サービスの「benchmark」が契約上の禁止事項に含まれるため、**企業が独自のモデル比較ベンチマークを実施する場合には契約解釈・書面許可を事前確認すべき**という珍しい実務リスクがある。 [6]

総合判断として、Grok 4.6は「GPT-5.6 Sol級の総合性能を、特に出力単価で大幅に安く提供するエージェント志向モデル」という評価が最も妥当である。ただし、導入判断では単価だけでなく、①20万トークン境界での価格倍増、②reasoning時の高い初動レイテンシ、③一部安全・幻覚指標の後退、④非公開アーキテクチャ、⑤独自ベンチマーク実施を含む契約制約、⑥オンプレミス版・open weightsが存在しないこと、を織り込む必要がある。 [7]

公式発表と提供条件

SpaceXAIの公式発表ページの日付は**2026年8月12日**。同社はGrok 4.6をGrok 4.5の後継とし、「long-running agents」と、より高度なinteractive／visual workを重点領域に挙げる。発表時点ではCursorとGrok Buildで利用可能と明記され、さらにAPI、OpenRouter、Vercel、Cloudflareへ提供されている。モデルカードではSnowflake、Databricks Mosaic等もゲートウェイとして挙げられているほか、Microsoft Word、PowerPoint、Excel向けadd-inのデフォルトモデルとも説明される。一方、一般消費者向けWeb、モバイル、X上のGrokへの4.6投入は「later date」とされており、発表時点では主眼が**開発者・企業・エージェント利用者**にある。 [8]

項目	2026年8月13日時点の確認結果
発表日	**2026年8月12日**
法的提供主体	X.AI LLC。「SpaceXAI」はDBA名称
モデルID	`grok-4.6`
主対象	開発者、コーディングエージェント、知識労働、企業ユーザー
API	**提供済み**。Responses API / Chat Completions
Grok Build	**提供済み、デフォルトモデル**
Cursor	**全プランで提供**
モデルゲートウェイ	OpenRouter、Vercel、Cloudflare。モデルカードにはSnowflake、Databricks Mosaic等も記載
一般Grok Web／モバイル／X	発表時点では将来提供予定
標準価格	入力 $2/M、cached $0.50/M、出力 $6/M
20万token以上	入力 $4/M、cached $1/M、出力 $12/M
Fast variant	標準価格の2倍
Business	$30/ユーザー/月、Grok 4.6を含む
Enterprise	個別見積もり
open weights	**未提供／未指定**
オンプレミス	**公開提供は確認できず**
モデルweightライセンス	**未指定**。API等は独自Enterprise/Consumer Termsに基づく

価格・配布チャネルは公式発表、API文書およびBusinessページによる。 [9]

特に注意したいのは、発表資料の「starts at $2/$6」という表現である。実際の料金表では**20万トークン以上になると$4/$12**となるため、500k context全体を活用する長時間agentではheadline priceだけでTCOを計算すると大幅に過小評価する可能性がある。キャッシュ価格もGrok 4.5の$0.30/Mから4.6では$0.50/Mへ上昇している。 10

また、Grok 4.6は「オンプレミスモデル」とは性格が異なる。公開ドキュメント上の実行先はSpaceXAI API、Grok Build、Cursor、各種ゲートウェイであり、ダウンロード可能なweightや自己ホスト用ライセンス、最低VRAM要件等は提示されていない。Enterprise向けにはDedicated Data PlaneやCustomer-Managed Encryption Keysなどが提供されるが、これは**オンプレミスでモデルweightを運用できるという意味ではない**。 11

企業としての名称にも注意が必要である。SpaceXは2026年2月2日にxAIを買収したと公式発表しており、Grok 4.6モデルカードでは「SpaceXAI is a doing-business-as (dba) name of XAI LLC」と明示されている。したがって、契約・DPA・輸出管理等ではマーケティング上の「SpaceXAI」ではなく、**契約主体X.AI LLCと各契約文書を基準に確認する必要がある**。 12

技術仕様と学習・アーキテクチャ

公式API仕様では、コンテキスト長は**500,000 tokens**、入力はtext/image、出力はtext、reasoning effortはlow / medium / high（default）/ xhigh。function calling、web search、X search、code executionをサポートする。APIドキュメント上のknowledge cutoffは**2026年2月1日**である。 13

技術項目	Grok 4.6
モデル規模	**「1.5T-scale model family」**。正確な総parameter数は未指定
Active parameters	**未指定**
Transformer層数	**未指定**
Attention方式	**未指定**
Expert数／routing方式	**未指定**
MoE	推論最適化で「fused MoE」の記述あり。MoE系であることを強く示唆するが詳細未指定
Context	500k tokens
最大text output	API文書では「No text output limit」
入力	Text + image
出力	Text
Reasoning	low / medium / high / xhigh
Tools	Function calling、Web Search、X Search、Code Execution
API	Responses、Chat Completions
Training cutoff	モデルカード：January 2026
API knowledge cutoff	February 1, 2026

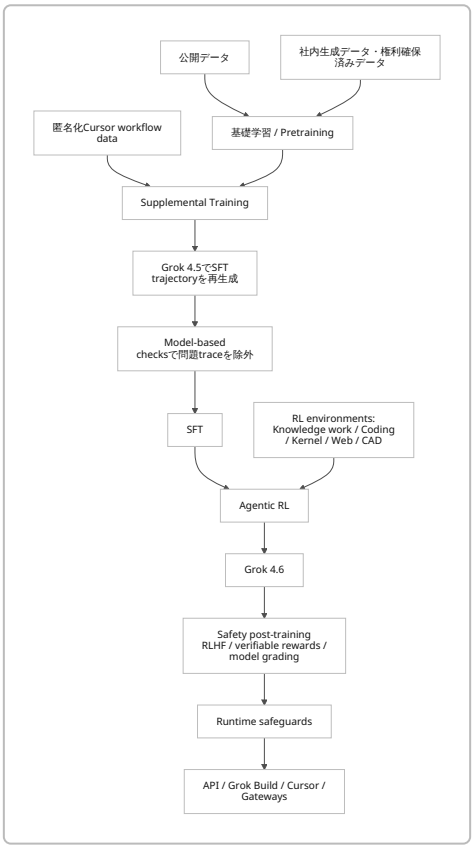
技術項目	Grok 4.6
Grok 4.6学習GPU数	**未指定**
Grok 4.6推論GPU要件	**未指定**
公式TPS／TTFT	**未指定**

仕様はAPIドキュメントとモデルカードに基づく。 14

「1.5T-scale」≠「正確に1.5兆parameter」と読むべき点は重要である。SpaceXAIはparameter countやactive parameter countを明示しておらず、「scale」が何を定義するかも公開文書にはない。そのため、外部記事に見られる「1.5 trillion parameters」との断定は、現時点では公式情報より強い表現である。モデルカードの自己推論最適化事例では、Grok 4.6のcheckpointが自らのchat inferenceを改善する過程で**fused MoE、FMHA、kernel scheduling、communication**を対象にしたとされる。これはMoE実装を強く示唆するが、expert数や一度にactiveになるexpert数までは公開されていない。 15

学習データについて、モデルカードは①公開データ、②社内生成データ、③SpaceXAIが必要な権利を確保したその他データをpretrainingに利用したとする。さらに、匿名化された**Cursor workflow data**を補助学習に利用したことも明記する。ただし、総training token数、各データソースの比率、具体的データセット一覧、著作権データのopt-out方式は**未指定**である。 16

学習工程は、少なくとも公開情報から次のように整理できる。これは**公開情報を構造化した概念図であり、内部実装そのものを示すものではない**。 17



4.6ではGrok 4.5より長いsupplemental trainingを実施し、reasoning／高度な技術概念向けの選別済みmodel-generated data、高品質engineering corpus、改良したoptimizer/training recipeを用いた。その後、Grok 4.5でSFT trajectoryを再生成し、problematic tracesをmodel-based checksで除外。さらにknowledge work、一般コーディング、kernel optimization、web development、CAD等でagentic RLを行った、とSpaceXAIは説明する。 [18]

モデルカードの「pretraining cutoff January 2026」とAPIの「knowledge cutoff February 1, 2026」は表面的には1か月ずれている。ただしpretraining cutoffと最終的なknowledge cutoffは必ずしも同じ概念ではなく、supplemental/post-trainingで後発情報が入り得る。この違いの具体的理由は**未指定**であり、企業の知識時点管理では「2026年2月1日」をAPI上の仕様値として扱いつつ、重要事項には検索／RAGを併用すべきである。 [19]

ハードウェアについては誤認しやすい。 前世代Grok 4.5は「数万台のNVIDIA GB300」で学習したとSpaceXAIが公開しているが、4.6について同等のGPU台数は公表されていない。またSpaceXAIのColossus 1全体には22万台超のH100/H200/GB200系GPUがあるが、これも「Grok 4.6が22万GPUで学習された」という意味ではない。したがって4.6固有の学習GPU構成・学習FLOPsは**未指定**とするのが厳密である。 [20]

推論性能も同様で、前世代4.5には公式80 TPSという値がある一方、4.6には公式TPS値がない。Artificial AnalysisのSpaceXAI API実測では、`grok-4.6 high`が約**65.5 tokens/s**、time-to-first-answer-tokenが約**31.18秒**である。後者はhigh reasoningで内部reasoningが進んだ後の初回answer tokenを測るため、通常の非reasoning chatbotのTTFTとは同列比較すべきではない。モデル固有の公式P50/P95レイテンシ、region別SLAは**未指定**である。 [21]

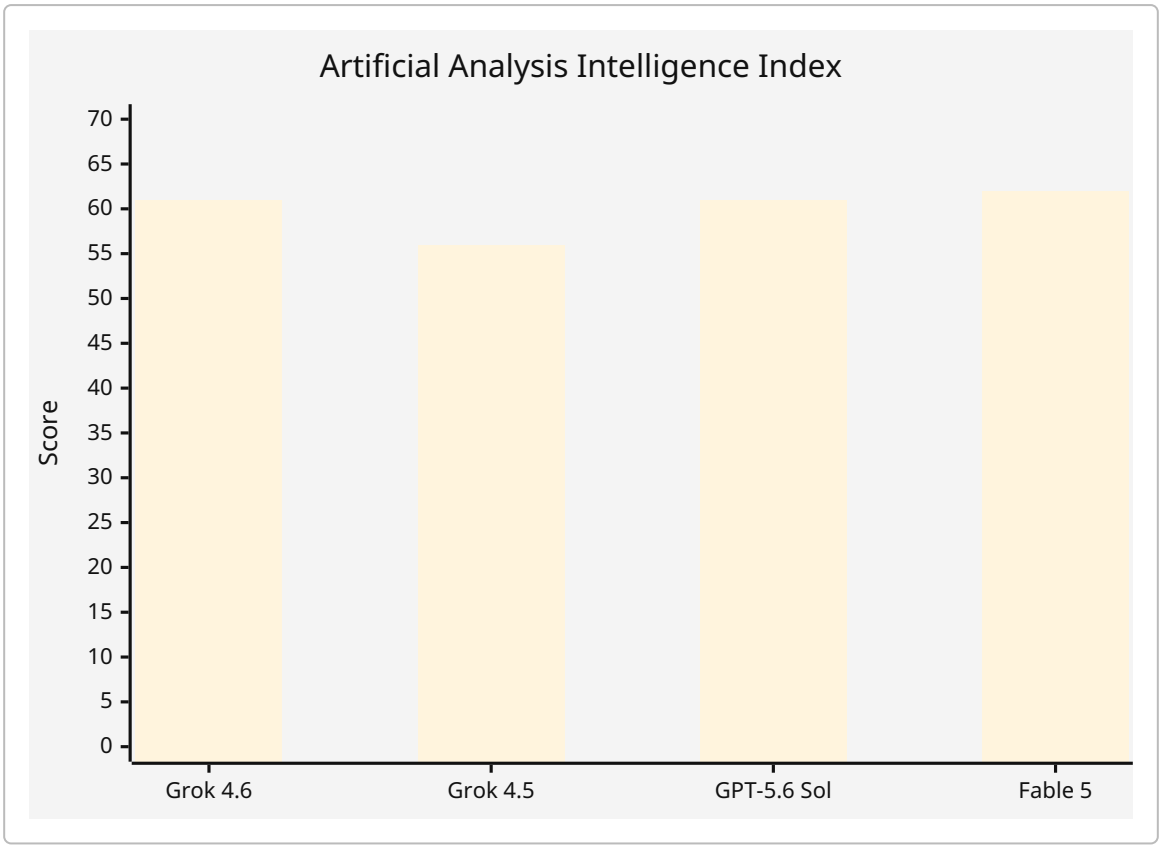
性能評価と競合比較

SpaceXAIの主要ベンチマークを見ると、4.6は4.5から**明確な世代改善**を達成している。特にDeepSWE、APEX-Agents、Terminal-Bench、AA-Briefcaseの伸びが大きい。一方で、最新競合に対して常にトップではない。 [22]

ベンチマーク	Grok 4.6 High	Grok 4.5 High	GPT-5.6 Sol Max	Fable 5 Max	4.6対4.5
AA Intelligence Index	**61**	56	**61**	62	+5
GDPVal-AA v2 Elo	**1753**	1526	1728	1741	+227
CursorBench 3.2	69.9%	66.7%	67.2%	**70.5%**	+3.2pt
DeepSWE 1.1	65.9%	54.0%	**73.0%**	70.0%	+11.9pt
FrontierCode 1.1 Ext.	61.3%	56.6%	60.6%	**63.6%**	+4.7pt
APEX-Agents	57.5%	47.1%	56.7%	**59.2%**	+10.4pt
Terminal-Bench 3.0	26.0%	15.7%	**34.6%**	34.1%	+10.3pt
APEX-SWE	56.4%	53.6%	未掲載	**58.8%**	+2.8pt
AA-Briefcase Elo	**1577**	1313	1502	1574	+264

数値はSpaceXAI発表資料。競合値はSpaceXAI自身が各社system cardや公開leaderboardから採用した値であり、同社も「best of self-reported or publicly available results」と注記している。したがって**全モデルを一つの完全統一harnessで同時再評価した表ではない**。 [22]

主要4モデルのAA Intelligence Indexだけを可視化すると次のようになる。Artificial Analysisによる独立評価も同じ61という結果を報告している。 [23]



ここから得られる実務的な読み方は、「Grok 4.6はGPT-5.6 Solと同格」という一文より複雑である。**知識労働・agent orchestrationでは競争力が高く、DeepSWEやTerminal-BenchではGPT-5.6 SolやFable 5に明確に負ける**。たとえばDeepSWE 1.1はGPT-5.6 Solより7.1ポイント低く、Terminal-Bench 3.0は8.6ポイント低い。一方GDPVal-AA、CursorBench、APEX-AgentsではGPT-5.6 Solを小幅に上回る。したがって「長時間agent／知識労働」と「純粋なterminal/software repair」を分けて評価する必要がある。 [22]

さらにモデルカードには、Grokにとって不利な評価も比較的多く収録されている。たとえばAPEX-SWEではOpus 5が63.7%、Fable 5が58.8%、Grok 4.6が56.4%。SWE-Marathon 1.1ではOpus 5が50.0%、Opus 4.8が48.8%、Kimi K3が48.1%、Fable 5が45.0%に対しGrok 4.6は31.9%である。この点は、発表資料のトップラインだけよりモデルカードの方が性能像をよく示している。 [24]

逆にGrok 4.6が強い特殊領域もある。OfficeQA Proでは**63.2%**でOpus 5/Fable 5の60.9%、GPT-5.6 Solの60.2%を上回る。モデル自身の推論スタック最適化実験では5時間で297の候補を検討し、7件のpull requestを作成、うち3件がproductionへ入り、decode throughputを1.5%、prefillを3.1%改善したとSpaceXAIが報告している。これは単なるコード生成よりも、**「計測→変更→検証→採用判断」を長時間継続するR&D agent**を戦略的に重視していることを示す。 [25]

独立評価のArtificial Analysisでも、Grok 4.6はAA-Briefcaseで1577 Eloを獲得し、長期knowledge-work taskを平均約53 turnsで処理したのに対し、Claude Opus 5 maxは約103 turnsだったと報告する。同機関はGrok

4.6の評価タスク当たり実コストを**0.84ドル**と算出しており、単純なtoken単価だけでなく、trajectory efficiencyもコスト優位に寄与している可能性がある。 [2]

ただし、評価透明性には限界がある。 SpaceXAIモデルカードには内部ベンチマーク、第三者ベンチマーク、他社自己申告値が混在する。たとえばDeepSearchQAはSpaceXAIの「internal implementation」であり、その実装が公開標準版と完全一致するかは外部から検証できない。 [26]

さらに、発表ページと同日付のモデルカードに**Harvey Legal Agent Benchmarkの不整合**がある。発表ページではGrok 4.6 **15.8%**、Fable 5 **11.3%**だが、モデルカードでは同じVals AI implementationについてGrok 4.6 **22.0%**、Fable 5 **14.2%**と記載される。Grok 4.5とGPT-5.6 Solは双方でそれぞれ12.9%、2.5%で一致しているため、単純なbenchmark名の違いだけでは説明しにくい。checkpoint、run、集計更新の違いなのかは**未指定**であり、SpaceXAIに説明を求めるべき箇所である。 [27]

GPT-4o、Gemini、Llama 3との比較上の注意

ユーザー指定のGPT-4o、Gemini、Llama 3については、**2026年のGrok 4.6と同一ベンチマークの数字を直接並べること自体が不公正**である。GPT-4oとLlama 3は2024年世代であり、SpaceXAIの2026年agentic benchmark表には掲載されていない。現在のOpenAI側の実質的比較対象としてSpaceXAI自身が採用しているのはGPT-5.6 Solである。 [28]

モデル	世代・位置づけ	Context	公開形態	API標準価格の例	Grok 4.6との公平な同一benchmark
Grok 4.6	2026 frontier agent	500k	Proprietary API	$2/$6、20万超 $4/$12	基準
GPT-5.6 Sol	2026 current OpenAI peer	—	Proprietary API	AA比較で$5/$30	AA Index 61対61
GPT-4o	2024世代、現在もAPI提供	128k	Proprietary API	$2.50/$10	**未公表／比較不適切**
Gemini 3.1 Pro Preview	現行Google系の代表例	1M input / 64k output	Proprietary API	$2/$12（<200k）、$4/$18（>200k）	**未確認**
Llama 3	2024 open-weight系	8B / 70B、8,192 training sequence	downloadable weights	自己ホスト依存	**未公表／比較不適切**
Llama 3.1 405B	Llama 3系列の大型版	128k	open-weight系	自己ホスト／provider依存	**未公表／比較不適切**

GPT-4oの現行API仕様は128k context、$2.50 input/$10 output、knowledge cutoff 2023年10月。GoogleのGemini 3.1 Pro Previewは1M context、$2/$12（20万token未満）、$4/$18（それ以上）。Metaの初代

Llama 3は8B/70Bで、標準decoder-only Transformer、GQA、15T超の公開ソース由来training tokensを公表しており、weightを取得して自社環境で運用できる点がGrok 4.6との根本的差である。 [29]

したがって、**オンプレ／データ主権ならLlama系、100万token級contextならGemini、2026 frontier reasoningの直接比較ならGPT-5.6 SolやClaude系、低単価の長期agentならGrok 4.6**という比較軸の方が、2024年ベンチマーク値を横並びにするより実務的である。これは公開仕様からの分析上の結論である。 [30]

セーフティ・倫理・評価透明性

SpaceXAIはGrok 4.6で「これまでで最も広範なpre-deployment testing」と、post-deploymentおよびthird-party testingを行ったとしている。モデルカードはcyber、bio/chem、jailbreak、一般有害出力、CBRN、mental health、behaviorを個別に評価しており、従来の単純な「拒否率」だけよりは広い情報を公開している。 [31]

安全機構そのものは、モデルカードにかなり具体的に記載される。安全fine-tuningとして**SFT + RLHF + verifiable rewards + model-based grading**を使い、system promptでhonesty/truth-seekingを誘導し、deployment surfaceに応じてCSAM、self-harm、CBRN、cyber等へのruntime input/topical filtersを追加する、多層防御である。したがって「RLHFの有無」は**有り**と確認できる。 [32]

安全性・信頼性指標	Grok 4.5	Grok 4.6	評価
Standard jailbreak（低い方が良い）	0.73%	**0.04%**	大幅改善
General harmful compliance	1.1%	**0.93%**	改善
Child-safety harmful compliance	0.00%	0.00%	維持
Bio refusal accuracy	97.9%	**100%**	改善
Chem refusal accuracy	96.7%	**100%**	改善
Hallucination rate	**0.98%**	1.7%	**悪化**
HackerBench harmful/dual-use compliance	**7.8%**	16.7%	**悪化**
Self-harm harmful compliance	**0.5%**	3.7%	**悪化**
MASK-Rectified dishonesty	**0.67%**	3.8%	**悪化**
Sycophancy	**0.01%**	0.04%	小幅悪化

SpaceXAIモデルカードによる。各指標は定義が異なるため、数値同士を直接加算して「総合安全スコア」と解釈してはならない。 [33]

特に**幻覚率**は重要である。モデルカードのFactuality (Hallucination)では、低いほど良い値としてGrok 4.6が1.7%、GPT-5.5が1.1%、Grok 4.5が0.98%、Opus 4.8が3.4%である。その一方、現行のSpaceXAIモデルページはGrok 4.6を「minimal hallucinations」等とマーケティングしている。この表現は「前世代より幻覚が減った」という意味では公式評価と整合しない。少なくともこの評価上は**4.6が4.5から退行**しているため、事実検証型アプリでは独自評価が必要である。 [34]

検索利用時は状況が異なる。DeepSearchQAではGrok 4.6が81.6%で、Grok 4.5の38.4%、GPT-5.6 Solの75.0%を上回るため、**closed-book factualityよりtool-grounded search-and-synthesisの伸びが大きい**と解釈できる。ただしこのDeepSearchQAは「internal implementation」なので、独立再現性には制限がある。 [26]

サイバー面でも能力と安全性を分離して見る必要がある。CyberGymではsafeguardsを外したGrok 4.6が79.7%でGrok 4.5の79.0%から小幅上昇し、CVE-Benchは35.2%から39.8%へ上昇した。SpaceXAIはrestricted modelを第三者評価者にも渡したとするが、**当該第三者の名称、全評価手法、個別結果はモデルカードでは公開していない**。よって「第三者評価あり」は確認できるが、「独立監査済み」と同義に扱うべきではない。 35

バイアス・公平性については、今回のモデルカードはhonesty、sycophancy、安全拒否等を扱う一方、性別・人種・宗教・国籍などに関する体系的な**demographic fairness/bias benchmarkの詳細は未指定**である。したがって採用、与信、保険等の公平性要求が厳しい用途で「モデルカードがあるからbias検証済み」とみなすことはできない。しかもSpaceXAIのAUP自体が、雇用・信用・教育・住宅・保険・法務・医療等における人の権利やwell-beingに関する**high-stakes automated decisionsを禁止**している。 36

モデルカード自身も、医療、法律、金融、安全クリティカル領域での**人間の監督や専門家検証なしの自律意思決定を意図していない**と明記する。これらではGrok 4.6を「decision maker」ではなく、文書探索、ドラフト、証拠抽出、コード補助等の**decision-support**へ限定する設計が安全かつ契約上妥当である。 24

利用条件・プライバシーと実運用

商用利用そのものは可能である。Enterprise Termsは企業がAPIを自社製品へ統合し、end userへBundled Serviceとして提供する権利を認めている。また入力は顧客が権利を保持し、当事者間では**顧客がOutputを所有**するとされる。 37

しかし、所有権と安全な商用利用は別問題である。Enterprise Termsは、出力が非一意であり得ること、不正確・不適切・hallucinationを含み得ることを明示し、用途に応じた**human reviewを顧客側の責任**としている。さらにxAIのIP indemnificationにはInput、Output等に由来する一定の除外があり、生成物について包括的な著作権非侵害保証が得られるわけではない。 38

特に企業R&D部門に重要なのが**benchmark条項**である。Enterprise Terms Section 2では、ツール等を用いてサービスを「probe, scan … or benchmark」する行為が禁止事項に含まれている。またmodel behaviorのdistillationや競合サービス開発も禁止する。通常のQA・受入試験と「benchmark」の境界は契約文だけでは明確でないため、複数LLMを社内leaderboardで比較するような導入プロセスは、契約担当またはSpaceXAIに**書面で適法性を確認することが望ましい**。 39

プライバシーについては、SpaceXAI API FAQが「明示的許可なしにAPI input/outputをtrainingしない」としている。ただし標準APIではrequest/responseを暗号化して**30日間保存**し、abuse/misuse監査に使用した後削除する。Enterprise Termsも原則30日以内の削除を定める。 40

ZDRでは保存を大幅に抑えられるが、機能上のトレードオフが大きい。stateful Responses、Files、Collections、Batch API、deferred completions、stored media、voice conversation historyなどは使えなくなる。APIレスポンスの`x-zero-data-retention` headerでZDR状態をプログラム的に確認できるため、**機密データ処理ではCI/CDあるいはgateway側でこのheaderを検証する運用が有効**である。 41

なおEnterprise TermsではPersonal Dataを扱う場合、ZDR経由を要求する条項があり、HIPAAのPHIについてはBAA締結＋ZDRが必要とされる。日本企業の場合も、個人情報保護法上の越境移転・委託先管理等を別途検討する必要があるが、どの日本法対応が自動的に満たされるかはSpaceXAIのGrok 4.6文書には**未指定**である。 42

輸出規制については、AUPおよびEnterprise Termsが米国その他のexport / sanctions laws遵守を顧客へ要求する。一方、Grok 4.6固有のECCN、モデルweightの輸出分類、利用禁止国一覧をモデルページからは確認できないため、これらは**未指定**と扱うべきである。 43

実運用コスト

標準short-context価格を使うと、典型的なAPIコストは次のようになる。tool invocationは別課金となる場合があり、以下にはweb search、X search、code execution等の追加費用を含めていない。 [44]

想定ワークロード	Token条件	1回当たり概算	月間量	月額概算
一般QA	10k input + 2k output	**$0.032**	100,000回	**$3,200**
中規模agent task	100k input + 20k output	**$0.32**	10,000回	**$3,200**
同上・inputが全てcache hit	100k cached + 20k output	**$0.17**	10,000回	**$1,700**
Long-context task	250k input + 20k output、long tier	**$1.24**	10,000回	**$12,400**
一般QA・Fast variant	10k + 2k	約**$0.064**	100,000回	約**$6,400**

最大のコスト管理ポイントは**200k token threshold**である。100k→250kへinputが2.5倍になるだけでなく単価も倍になるため、上記例では1taskの費用が$0.32から$1.24へ約3.9倍になる。長時間agentでtool resultや過去trajectoryを無制限にhistoryへ積む設計は、性能とコスト双方で不利になり得る。SpaceXAI自身もlong agent loopsにはcontext compactionを推奨し、`prompt_cache_key`による同一serverへのroutingでcache hitを確実にすることを勧めている。 [45]

なお、現時点で**Grok 4.6と4.5はBatch API非対応**である。大量の非リアルタイム処理をBatch discount前提で設計している企業には重要な制約となる。 [46]

レイテンシ面では、Artificial Analysisの`high` reasoning実測が約65.5 output tokens/s、first answer token約31秒であるため、Grok 4.6 high/xhighは即応型customer chatより、**数十秒〜数分を許容して品質・agentic reasoningを優先するワークロード**との相性が良い。低reasoning設定やFast variantの独立実測は、発表翌日の現段階では十分蓄積しておらず、SLA評価には時期尚早である。 [47]

実務上の適性をまとめると次の通りである。

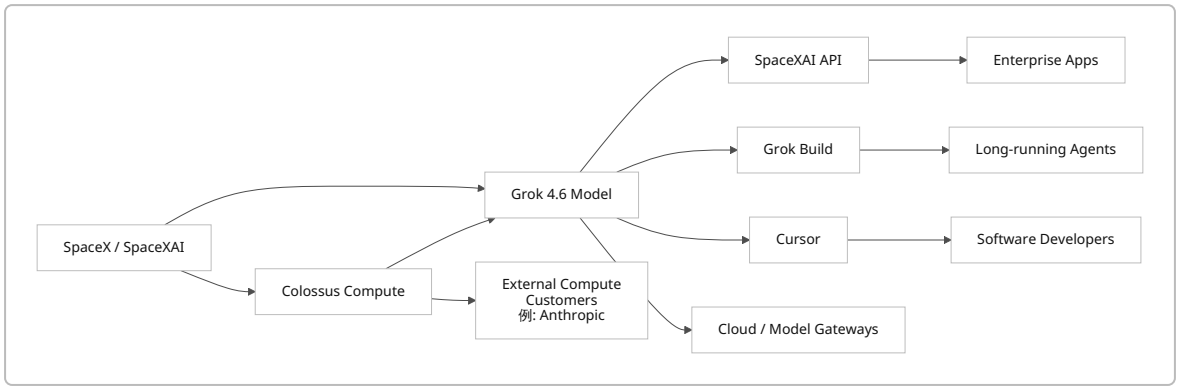
用途	適合度	理由／注意
Repository横断のcoding agent	**高**	CursorBench、agentic RL、長期trajectoryが主戦場
調査・knowledge work agent	**高**	GDPVal、AA-Briefcase、DeepSearchQAが強い
Office文書分析・生成	**高**	OfficeQA Proで強い
UI/アプリprototype、CAD	**高〜中**	学習環境として明示。CAD系評価も実施
R&D／ML engineering	**高**	inference optimizationを自律実施した事例
大規模低遅延chat	**中**	high reasoningのTTFTが長い
20万token超の長文大量処理	**中**	contextは500kだが単価が倍増
自社オンプレ・air-gapped運用	**低**	open weights／on-prem提供なし

用途	適合度	理由／注意
医療・法務・金融の自動最終判断	**不適**	AUP・model card上のhigh-stakes制限
個人データ＋stateful server history	**要設計**	ZDRとstateful機能が両立しない

根拠は公式API、モデルカード、AUP、および独立性能測定による。 [48]

競争環境・市場インパクトと未解決点

Grok 4.6の戦略的意味は、単なる「より賢いGrok」より、**モデル＋agent runtime＋開発者distribution＋compute infrastructureを一体化するSpaceXAI戦略**にあると考えるのが妥当である。これは公開事実からの推論である。SpaceXAIは自社でGrokを開発し、APIを販売し、Grok Buildを提供し、Cursorと共同開発・配布し、さらにColossus 1の計算資源を競合Anthropicにも提供している。つまりAIモデル市場で競争する一方、計算インフラ市場では競合を顧客にもする二面戦略である。 [49]



差別化の第一は**価格**である。Artificial Analysisの比較ではGrok 4.6の$2/$6はGPT-5.6 Solの$5/$30、Claude Opus 5の$5/$25より大幅に低く、それでAA Intelligence Index 61を確保している。reasoning-heavy workloadではoutput tokensの比率が高くなりやすいため、$6対$30という出力単価差は大きい。 [2]

第二は**agentのtrajectory効率**である。SpaceXAIはGrok 4.6を「長時間agent」のためのモデルとして訓練し、Artificial AnalysisでもAA-BriefcaseでOpus 5より少ないturn/input tokensでタスクを完了する傾向が観測された。もしこの性質が企業固有のrepository、ERP、CRM、research workflowでも再現すれば、「価格／token」ではなく**「価格／完了タスク」**で競争優位を作れる。逆に、その再現性は企業ごとの実データで検証しなければならない。 [50]

第三は**開発者distribution**である。モデル発表と同時にCursor全プラン、Grok Build、API、複数gatewayで使えることは、「モデルを出してからecosystem integrationを待つ」方式より利用摩擦が小さい。発表後1週間はCursor/Grok Buildのincluded usageを2倍にする施策もあり、ベンチマーク改善を即座に実利用へ転換する意図が明瞭である。 [51]

市場はこの動きを「Anthropic/OpenAIのpremium frontier pricingに対する価格圧力」と受け取っている。8月12日の金融報道ではSpaceX株が約9.7%上昇したことがGrok 4.6等のAI材料とともに報じられた。ただし同日には他のAI事業・企業戦略ニュースもあり、**株価変動の全てをGrok 4.6単独の因果効果とみなすことはできない。** [52]

今後の市場インパクトとして最も可能性が高いのは、「最高benchmarkスコア争い」よりも、**frontier-grade agentの単価競争を加速すること**である。Grok 4.6が61点でGPT-5.6 Solと並びながら出力token単価を

5分の1に設定できているため、競合には①単価引き下げ、②token efficiency向上、③agent platformへのbundling、④premium modelとcheap modelのtier分化、のいずれかを迫る圧力になると推測できる。 [23]

一方、**Grok 4.6の利用シェアについて信頼できる定量推定を置くことは現時点では避けるべき**である。発表から1日しか経っておらず、Grok 4.6固有のAPI call volume、enterprise seat数、Cursor内model selection share等の一次データが公開されていない。SpaceXAI全体やGrok全体のtrafficを4.6固有のshareへ読み替えるのは不適切である。したがって「利用シェア推定」は現時点では**未指定**／**推定不能**とする。

最後に、導入判断前に追跡すべき未解決点を重要度順に整理する。

未解決点	現状	実務への影響
正確なparameter count / active params	**未指定**	自己ホスト可能性、計算効率、architecture比較不能
MoE expert構成	**未指定**	serving特性・capacity分析が困難
Grok 4.6固有のtraining GPU/FLOPs	**未指定**	学習コスト・再現性を評価不能
公式P50/P95 latency、TPS	**未指定**	production SLA設計に不足
Fast variantの独立性能	十分なデータなし	2倍料金のROI判断が困難
Pretraining cutoff Jan vs API cutoff Feb 1	説明なし	temporal knowledge管理
Harvey LAB 15.8% vs 22.0%	**公式資料間で不整合**	benchmark governanceへの疑問
Hallucination 4.5→4.6悪化	0.98→1.7%	fact-heavy用途で要独自検証
Self-harm / dishonesty等の退行	一部確認	安全性が一様改善ではない
Third-party safety evaluator名と全結果	**未指定**	独立監査性が限定的
demographic bias評価	**未指定**	HR・公共・顧客対応で評価不足
On-prem / open weights	**未提供**	air-gap、sovereign AI用途に不向き
Grok 4.6 Batch API	**現時点非対応**	大量offline処理のコスト最適化を制約
benchmark禁止条項の範囲	契約解釈が必要	PoC/調達比較プロセスに影響
Grok 4.6固有のECCN等	**未指定**	輸出・制裁対応で追加確認必要
20万token超価格の実TCO	workload依存	長時間agentでheadline priceとの差大
発表後の稼働率・incident実績	データ蓄積前	production reliability評価は時期尚早

とりわけ調達・PoCでは、**Grok 4.6 vs GPT-5.6 Sol / Gemini現行モデル / Claude現行モデルを、自社の実タスクで「成功率・人間による修正時間・総token数・tool call数・P95時間・1完了タスク当たり費用」の六指標で評価する**のが最も有用である。ただし前述の通り、SpaceXAI Enterprise Termsのbenchmark制限に抵触しないよう、比較試験の方法・公開範囲について事前確認が必要である。 [53]

現時点での最終評価は、**「Grok 4.6はSpaceXAIが“モデル性能”だけでなく“agentic task economics”を競争軸に据えたことを示す重要なリリース。価格性能は非常に強いが、透明性、安全性の一部後退、長context料金、契約制約が企業利用の主要リスク」**である。独立評価で総合61点が再現されていることは発表時のマーケティングだけに依存しない強い材料だが、個々のbenchmarkには勝敗があり、「万能な最高性能モデル」と位置づける証拠はない。 54

navlistGrok 4.6発表後の主要報道turn16news23,turn16news24,turn16news25

---

1 3 8 9 17 18 22 27 31 50 51 Introducing Grok 4.6 | SpaceXAI
https://x.ai/news/grok-4-6

2 23 54 Grok 4.6 returns SpaceXAI to the intelligence frontier and leads on cost efficiency
https://artificialanalysis.ai/articles/grok-4-6-benchmarks-and-analysis

4 5 15 16 19 24 25 26 32 33 34 35 49 media.x.ai
https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf

6 40 41 FAQ - API Security | SpaceXAI Docs
https://docs.x.ai/developers/faq/security

7 10 44 45 Pricing | SpaceXAI Docs
https://docs.x.ai/developers/pricing

11 13 14 30 48 Grok 4.6 | SpaceXAI Docs
https://docs.x.ai/developers/grok-4-6

12 xAI joins SpaceX | SpaceXAI
https://x.ai/news/xai-joins-spacex

20 21 Introducing Grok 4.5 | SpaceXAI
https://x.ai/news/grok-4-5

28 29 GPT-4o Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-4o

36 43 Acceptable Use Policy | SpaceXAI
https://x.ai/legal/acceptable-use-policy

37 38 39 42 53 Terms of Service - Enterprise | SpaceXAI
https://x.ai/legal/terms-of-service-enterprise

46 Batch API - xAI Docs - SpaceXAI
https://docs.x.ai/developers/advanced-api-usage/batch-api?utm_source=chatgpt.com

47 Grok 4.6 (high) API Provider Benchmarking & Analysis
https://artificialanalysis.ai/models/grok-4-6/providers?utm_source=chatgpt.com

52 SpaceX's stock is getting a Grok-fueled boost
https://www.marketwatch.com/story/spacexs-stock-is-getting-a-grok-fueled-boost-e99497b4?utm_source=chatgpt.com