Grok の新モデルの特徴と初期評価

対象モデル：Grok 4.7　／　調査基準日：2026 年 9 月 22 日

GPT-6 Astra

Grok 4.7 は、長時間のコーディングと資料分析・文書作成を強化し、比較的低いトークン単価で提供する業務向けモデルと評価できる。独立評価でも改善は確認されているが、トークン消費の増加、日本語品質、評価条件による成績の差まで含めて判断する必要がある。[1][2]

本資料は、公式発表、公式モデルカード、独立評価機関の測定、利用者の初期報告を調査した結果をまとめたものである。公式資料の日付表記は 9 月 21 日。本資料中の価格・提供状況・評価値は調査時点の情報であり、公開直後の初期評価と、その評価に基づく考察を区別して記載する。[1]

改良の中心は、難しい仕事を継続し、成果物を確認する能力である。従来版 Grok 4.6 より大きい基盤モデルを用い、長時間かかる課題を重視した学習を行ったと説明されている。Cursor の公式文書も、複雑なコード作成、長い文脈の管理、自己検証を主要な改善点に挙げている。[1][3]

表 1　基本仕様と利用上の意味

項目	確認できた内容	利用上の意味
主な対象	コーディング、エージェント業務、知識労働	複数工程を伴う作業を重視
入出力	テキスト・画像入力、テキスト出力	図を含む資料の読解にも利用可能
コンテキスト	最大 50 万トークン	大量の資料やコードを扱える
推論設定	low／medium／high／xhigh 既定は high	課題の難度に応じて推論量を調整
外部ツール	Web 検索、X 検索、コード実行、関数呼び出し	情報取得と処理を組み合わせられる
提供形態	API などを通じたクローズドモデル	公開重みをダウンロードして運用する方式ではない

出典：[3][4][5]。Cursor では通常のコンテキスト枠が 25.6 万、最大 50 万トークンと記載されており、利用サービス側の設定も関係する。

利用経路によって、提供状況は異なる。公式モデルカードには、API、Grok Build、Cursor、Word・PowerPoint・Excel 用 Grok アドインなどが記載されている。一方、一般向け Web・モバイルアプリ・X 内 Grok への導入は「後日」とされている。公開されたモデルと、普段使う Grok 画面で選べるモデルは、確認を分ける必要がある。GitHub Copilot では 9 月 21 日に提供開始が発表されたが、段階的な展開である。[6][7]

公式ベンチマークでは、前モデルからの改善が広く見られる。以下は公式発表ページに掲載された比較値である。特定の評価で強さを示す一方、すべての領域で競合を上回る結果ではない。[1]

表 2　公式発表ページに掲載されたベンチマーク比較

評価対象とベンチマーク	Grok 4.7	Grok 4.6	GPT-5.6 Sol	Claude Fable 5.1
長時間のコード作成 CursorBench 4.0	46.3%	40.4%	41.7%	51.8%
ソフトウェア開発 DeepSWE v1.1	71.0%	65.2%	72.7%	70.0%
電気工学 EEBench	64.0%	53.0%	39.4%	56.4%
複数時間の事務作業 AA-Briefcase v1.1	1,657	1,546	1,487	1,678
端末操作を伴う作業 Terminal-Bench 4.0	38.0%	20.3%	37.3%	57.9%
法務業務 Harvey Legal Agent Benchmark	19.6%	15.8%	2.5%	6.7%
臨床推論 HealthBench Professional	56.7%	48.5%	60.5%	62.1%

出典：[1]。原則として Grok 4.7 は xHigh、4.6 は High、競合は Max 設定。DeepSWE の Grok 4.7 だけは High。AA-Briefcase は Elo 評価であり、百分率ではない。EEBench の資料間の数値差は後掲の表 6 参照。

電気工学や一部の法務評価では強さを示す一方、コード作成・端末操作・臨床推論などでは上回る競合がある。また、4.7 と 4.6 の推論設定が異なるため、改善幅のすべてを基盤モデルの進歩に帰することはできない。

ただし、同じ High 設定でも、モデルカード上の CursorBench は 4.7 が 43.9%、4.6 の公式掲載値は 40.4%である。改善が推論量の増加だけによる、と断定することも適切ではない。[1][6]

法務の 19.6%は、指標の定義を確認して読む必要がある。採用された Harvey の方式は、各課題のすべての評価条件を満たして初めて合格とし、二つの採点モデルによる課題合格率を平均する。「法律問題の正答率 19.6%」という意味ではない。部分的には多くの条件を満たしていても、重要な条件を一つ欠けば課題全体は不合格となる。専門業務の完成度を考えるうえで、意味のある厳しい指標である。[8]

独立評価で明確な改善が見られるのは、成果物を作る知識業務と、Grok Build を使うコーディングである。Artificial Analysis は、次の結果を公表している。[2]

表 3　Artificial Analysis による独立評価

評価	Grok 4.6	Grok 4.7
Intelligence Index	44	46
AA-Briefcase 長時間の知識業務	1,546 Elo	1,657 Elo

評価	Grok 4.6	Grok 4.7
GDPval-AA 実務成果物の作成	1,605 Elo	1,695 Elo
Coding Agent Index Grok Build との組み合わせ	47	56

出典：[2]。4.7 は xhigh。比較する 4.6 は、上 3 行が high、Coding Agent Index が xhigh。Coding Agent Index はモデルと実行環境の組み合わせを評価している。

改善の内訳は均一ではない。AA-Briefcase では分析品質が 1,690 から 1,994 Elo へ上昇した一方、見せ方の品質は 1,519 から 1,499 だった。長文推論の AA-LCR は 3.7 ポイント、業務自動化の AutomationBench-AA は 1.1 ポイント低下している。資料の分析と、読みやすい完成品への仕上げは個別に確認する必要がある。[2]

報道で見かける「4 位」には、主要 AI 開発企業として上位 4 社に入ったという評価と、各社固有の実行環境を使った Coding Agent Index で Grok 4.7＋Grok Build が 4 位という評価がある。後者の上位は Claude Fable 5.1、GPT-6 Astra、Claude Opus 5 である。「あらゆる能力を総合して世界 4 位の単体モデル」と読むのは正確ではない。[2]

料金は魅力的だが、実費はトークン消費まで見ないと判断できない。標準 API の価格は Grok 4.6 と同じである。[9]

表 4　標準 API のトークン料金

入力長	入力 100 万トークン	キャッシュ入力 100 万トークン	出力 100 万トークン
20 万トークン未満	2 ドル	0.50 ドル	6 ドル
20 万トークン以上	4 ドル	1 ドル	12 ドル

出典：[9]。通貨は米ドル。長文料金の閾値に達すると、超過部分だけでなく、そのリクエスト全体に長文料金が適用される。検索などのツール利用料金も別途発生する。

高速版の Grok 4.7 Fast は、同じモデルを高速な基盤で動かすものである。通常の短文料金は入力 4 ドル・出力 12 ドルで、提供先は Cursor と Grok Build に限定され、公開 API では利用できない。Cursor の長文課金の閾値は 25.6 万トークンで、API の 20 万とは異なる。[3][4]

Artificial Analysis では、Intelligence Index の 1 課題当たり出力トークン数が、同じ xhigh 設定で 4.6 の約 3.8 万から 4.7 の約 8.1 万へ増えた。単価が同じでも、使用量が約 2.1 倍なら、出力部分の費用も増える。[2]

調査時点の AA モデルページでは、high と xhigh の総合指数はどちらも丸め値で 46 だが、評価課題当たりの費用は high が 2.73 ドル、xhigh が 3.74 ドルだった。これはベンチマーク上の換算値で、一般業務の標準料金ではない。それでも、まず high で試し、難しい課題だけ xhigh にする運用を検討する根拠にはなる。[5][10]

宣伝文句の「2 倍速・半額」は比較対象モデルに対する主張である。Grok 4.6 からの変化については、公式発表は価格・速度の据え置きと説明している。[1]

利用者の初期評判は、用途によって分かれている。公開直後のため、以下は個別の使用報告として読むのが適切である。環境、設定、課題数がそろった比較ではなく、好評・不評の割合を示すものでもない。

表 5　利用者の初期報告

報告	内容	読み取れる範囲
Hacker News 画像から Web ページを作る比較	改善と価格面を評価。ただし完成度やアニメーションは Astra を支持	初稿作成では有望でも、仕上げに差が残るという事例
Hacker News コーディングとエージェント利用	遅さと費用増を指摘し、主力モデルの代替になるか判断を保留	長い推論が作業効率を下げる場合がある
Reddit 4.6 との比較	同じ extra high 設定で、4.7 の 3D 表現や修正のしやすさを評価	改善を体感した報告もある
日本語の試用記事	20〜30 分ほど試し、日本語の硬さと大量文脈での理解に不満	日本語品質には追加検証が必要

出典：表の上から順に[11]、[12]、[13]、[14]。各投稿者の観察・評価を要約した。

メディアの論調にも慎重さがある。The Decoder は低価格を認めつつ Claude や GPT-6 との性能差を強調し、Decrypt も最上位モデルへの追随という見方を示している。こうした論評と、特定の仕事で使いやすいという利用者の感想は両立する。採用判断では、自分の仕事に近い評価を優先するのが妥当である。[15][16]

今回確認した範囲では、Grok 4.7 を対象に、日本語の専門文書作成・特許翻訳・日本の特許実務を十分な件数で比較した独立評価は見つからなかった。日本語については、まだ個人の短時間の使用感が中心である。

情報の正確さには改善の兆候があるが、根拠確認は引き続き必要である。AA-Omniscience では、4.6 high から 4.7 xhigh への比較で、ハルシネーション指標は 34%から 29%へ改善した一方、正答率は 48%から 47%とほぼ横ばいだった。[2]

この 29%は、全回答の 29%が誤りという意味ではない。AA は「誤答と回答保留を合わせた中で、誤答した割合」と定義している。知識不足に直面した際の応答の仕方を測る指標である。[17]

補助的な調査として、TrueStandard は、30 の命題を同日に 3 回ずつ試し、架空 DOI の割合が 4.6 の 16／128 件から 4.7 の 6／129 件へ減ったと報告している。ただし、小規模な検索なしの試験で、別日の再現確認は未了である。一般的な正確性や他社モデルに対する優位性まで示す結果ではない。[18]

安全対策について、SpaceXAI は新しい防御機構と拒否判断の改善を説明している。HackerBench で危険性のあるデュアルユース要求を通した割合は 3.3%とするが、これは同社が公表した特定評価の結果である。[1]

紹介記事で取り上げられる LatchBio の 62.4%は、モデルカードでは BioSecBench の Refusal 項目に対応する。生物学全般の正答率や、あらゆる利用場面での安全性を表す数字ではない。[6]

公開直後の資料には、整合性を確認すべき点も残っている。発表ページ、API 文書、モデルカード、報道記事を比較すると、以下の相違が確認できる。

表 6　資料間で記載が異なる項目

項目	確認した記載	現時点での扱い
EEBench	発表ページは 64.0%、モデルカードは 66.0%	評価時点・条件の差を確認する必要がある
知識と学習データの期限	API 概要は知識カットオフを 2026 年 5 月、カードは事前学習データを 6 月、追加学習データを 8 月までと記載	指標の意味が異なる可能性があり、単一の期限にまとめない
パラメータ数	一部報道は 2.1 兆とするが、確認した公式モデルカードには具体的な数の記載がない	確定仕様として採用しない

出典：[1][4][6][16]。本資料の公式比較表は、発表ページの掲載値に統一し、異なる資料の数値を混ぜていない。

知財実務では、成果物作成とデータ処理から適性を検証したい。以下は公開評価に基づく考察であり、Grok 4.7 を用いた特許実務の実証結果ではない。試す価値が高いのは、複数資料から比較表や報告書を作る作業と、特許データを加工・集計するコードの作成である。今回改善が確認された知識業務・コーディングとの対応が比較的明確だからである。[2]

クレーム解釈、進歩性の判断、FTO 分析への適性は、一般的な法務ベンチマークだけでは判断できない。評価するなら、文章の流暢さに加え、次の項目を重視したい。

・クレームの構成要件を漏らさず抽出できるか。

・引用文献の該当箇所と、説明・対比内容が一致するか。

・根拠不足や判断不能を適切に示せるか。

・日本語の修正時間と専門家の確認時間を含めて、作業時間が減るか。

・検索、再試行、長文課金まで含めた費用が見合うか。

未公開の発明情報を扱う場合には、利用経路のデータ条件も判断材料になる。公式 API は、明示的な許可なしに入力・出力を学習に使わないと説明しているが、標準では監査目的で 30 日保存する。ゼロデータ保持の選択肢もある一方、Files・Collections などの機能に制限が生じる。この API の条件を、一般向け Grok や他社経由のサービスにそのまま当てはめることはできない。[19]

参考文献

本文中の角括弧内の番号は、以下の文献番号に対応する。URL の最終閲覧日はすべて 2026 年 9 月 22 日。動的な評価ページ・料金表・提供状況は、その後更新される可能性がある。

[1] SpaceXAI. Introducing Grok 4.7. 2026 年 9 月 21 日.
https://x.ai/news/grok-4-7
参照内容：公式発表。改良点、比較表、安全評価、価格・提供経路。

[2] Artificial Analysis. Benchmarking Grok 4.7. 2026 年 9 月 21 日.
https://artificialanalysis.ai/articles/benchmarking-grok-4-7
参照内容：独立評価。知識業務、Coding Agent Index、トークン消費、AA-Omniscience の結果。

[3] Cursor. Grok 4.7 — Cursor Docs. 公開日記載なし.
https://cursor.com/docs/models/grok-4-7
参照内容：モデル仕様、推論設定、標準・長文コンテキスト、料金。

[4] SpaceXAI. Grok 4.7 — SpaceXAI Docs. 2026 年 9 月 21 日更新.
https://docs.x.ai/developers/grok-4-7
参照内容：公式 API 概要。入出力、コンテキスト、知識カットオフ、Fast 版の提供条件。

[5] Artificial Analysis. Grok 4.7 (xhigh) — Intelligence, Performance & Price Analysis. 継続更新.
https://artificialanalysis.ai/models/grok-4-7
参照内容：xhigh 設定のモデル評価ページ。数値は閲覧時点。

[6] SpaceXAI. Model Card: Grok 4.7. 2026 年 9 月 21 日版.
https://media.x.ai/v1/website/card4p7-3a96f40b.pdf
参照内容：公式モデルカード、全 30 ページ。主な参照箇所は紙面ページ 4、5、11、12、19。

[7] GitHub. Grok 4.7 is now available in GitHub Copilot. 2026 年 9 月 21 日.
https://github.blog/changelog/2026-09-21-grok-4-7-is-now-available-in-github-copilot/
参照内容：公式変更履歴。対応プラン・環境、段階的な展開。

[8] Vals AI. Harvey's Legal Agent Benchmark. 2026 年 9 月 21 日更新表示.
https://www.vals.ai/benchmarks/hlab
参照内容：評価方法。課題全体の合格率と個別条件の合格率の区別、二つの LLM 判定者。

[9] SpaceXAI. API Pricing — SpaceXAI Docs. 2026 年 9 月 21 日更新.
https://docs.x.ai/developers/pricing
参照内容：標準・長文 API 料金、ツール課金、Fast 版の提供・料金条件。

[10] Artificial Analysis. Grok 4.7 (high) — Intelligence, Performance & Price Analysis. 継続更新.
https://artificialanalysis.ai/models/grok-4-7-high
参照内容：high 設定のモデル評価ページ。数値は閲覧時点。

[11] jjcm ほか／Hacker News. It's definitely gotten better at image->html workflows. 投稿日時は取得時に相対時刻で表示.
https://news.ycombinator.com/item?id=49792331
参照内容：画像から Web ページを作る比較についての個人投稿と返信。統制試験ではない。

[12] mchusma ほか／Hacker News. Initial impressions, Grok 4.6 for me just didn't really hack it for any usecase I tried. 投稿日時は取得時に相対時刻で表示.
https://news.ycombinator.com/item?id=49793859
参照内容：Grok 4.7 の速度、費用、コーディング・エージェント利用に関する初期所感。

[13] Beneficial-Day7238 ほか／Reddit r/cursor. Grok 4.6 vs 4.7. 投稿日時は取得時に相対時刻で表示.
https://www.reddit.com/r/cursor/comments/1wmlaji/grok_46_vs_47/
参照内容：両モデルを extra high 設定で試した個人報告と返信。

[14] 奥村 龍晃／note．Grok 4.7 リリース！OpenAI GPT-6 Sol/Luna 発売間近！？Jev 待機リスト解除！スキマ時間に読む AI 速報．2026 年 9 月 22 日．
https://note.com/redcord/n/n02f1fb8759bd
参照内容：Grok 4.7 の日本語品質に関する短時間の試用所感を参照。他項目の未確認情報は採用していない。

[15] Matthias Bastian／The Decoder．xAI launches Grok 4.7 at bargain prices, but benchmarks reveal a wide gap to Claude and GPT-6．2026 年 9 月 21 日．
https://the-decoder.com/xai-launches-grok-4-7-at-bargain-prices-but-benchmarks-reveal-a-wide-gap-to-claude-and-gpt-6/
参照内容：価格と競合との差を論じた記事。メディアの評価・論調として参照。

[16] Jose Antonio Lanz／Decrypt．xAI Launches Grok 4.7. It's Bigger, But Late to the AI Frontier Party．2026 年 9 月 21 日．
https://decrypt.co/378824/xai-launches-grok-4-7
参照内容：初期報道・論評。アプリ提供状況やパラメータ数などは公式資料との相違を留保。

[17] Artificial Analysis．AA-Omniscience: Knowledge and Hallucination Benchmark．2025 年 11 月 16 日．
https://artificialanalysis.ai/articles/aa-omniscience-knowledge-hallucination-benchmark
参照内容：正答率、ハルシネーション率、AA-Omniscience Index の定義。

[18] TrueStandard．Does Grok 4.7 fabricate fewer citations than Grok 4.6?．2026 年 9 月 22 日．
https://truestandard.ai/releases/grok-4-7
参照内容：30 命題を同日に 3 回測定した DOI 検証の報告。小規模な補助的証拠として扱う。

[19] SpaceXAI．FAQ — API Security．公開日記載なし．
https://docs.x.ai/developers/faq/security
参照内容：API データの学習利用、標準 30 日保持、Zero Data Retention の条件。