ChatGPT

xAI「Grok 4.20」徹底調査報告書

エグゼクティブサマリー

本調査の最重要結論は、**「Grok 4.20」について“公式に確定できる技術仕様は限定的”**であり、現時点（2026-02-23, JST）で**公式一次情報として明示的に確認できるのは、xAIがAPI側で「Grok 420（および420 Multi-Agent）」を“coming soon”として早期アクセス募集を開始している事実**が中心、という点である。[1]
一方で、xAIは **Grok 4（2025-07）→ Grok 4 Fast（2025-09）→ Grok 4.1（2025-11）→ Grok 4.1 Fast＋Agent Tools API（2025-11）** という系譜で、（a）大規模RL（強化学習）を中核とする能力拡張、（b）ツール使用の“ネイティブ化”とエージェント設計の製品化、（c）推論・非推論モード統合やトークン効率化による低コスト化、を段階的に推し進めてきた。[2]

また、xAIは**安全性に関してモデルカード（Grok 4 / Grok 4.1）とリスク管理枠組み（RMF）を公開**し、悪用（CBRN/サイバー等）・制御喪失（欺瞞等）を中核リスクとして、評価指標（例：Refusal、Jailbreak耐性、AgentHarm、AgentDojo、MASK、WMDP、VCT等）と緩和策（システムプロンプト、入力フィルタ等）を体系化している。[3]
ただし、**Grok 4.20固有の安全性評価（モデルカード相当）やベンチマーク結果、API上のモデル名、価格、コンテキスト長などは公式ニュース／公式ドキュメントからは確定困難**であり、ここは推定と事実を厳格に分離して扱う必要がある。[4]

性能面では、xAIはGrok 4 / 4 Fast / 4.1 / 4.1 Fastについて多数の定量情報を公開している（例：Grok 4は200,000 GPUクラスター「Colossus」による“RLを事前学習スケールで回す”方針、Grok 4 Heavyの並列テスト時算出、Grok 4.1のLMArenaでの上位、Grok 4.1 Fastのツール呼び出し系ベンチ（τ^2-bench、Function Calling等））。[5]
一方、「Grok 4.20」自体の実運用評判は、Hacker News[6]等で「リンクはあるが仕様が出ておらず議論しにくい」という指摘が見られ、情報の非対称性が採用判断のボトルネックになっている。[7]

結論として、**プロダクト組込み（API利用）という観点では、現時点で“確定情報が揃っている”Grok 4.1 Fast＋Agent Tools APIを基準に設計・評価を進め、Grok 4.20/420は公式仕様と価格、モデルカードが出た段階で段階導入（A/B・カナリア・ガードレール強化）を推奨**する。[8]

用語と調査範囲

命名の整理

ユーザーが指定した「Grok 4.20」は、外部報道・コミュニティ文脈では“4.20”と表記される一方、xAIのAPIドキュメント側では **「Grok 420」「Grok 420 Multi-Agent」**という表記で**「APIに近日提供（coming soon）」**として早期アクセス募集が存在する。[4]
したがって本報告書では、以下のように整理する。

- **確定（公式）**：API上の将来提供枠として「Grok 420」「Grok 420 Multi-Agent」が“coming soon”である。[1]
- **未確定（非公式／報道・SNS）**：コンシューマ向け“Grok 4.20（public beta）”のリリース時期や、4エージェント協調などの具体仕様。[9]

調査対象の情報源ポリシー

- **一次（最優先）**：xAI公式ニュース、xAI Docs（API仕様、FAQ、リージョン、ツール仕様）、モデルカード、RMF、法務文書（プライバシー、AUP等）、公式GitHub。 10
- **二次（補助）**：主要メディア報道、コミュニティ（Reddit/HN/日本語フォーラム）投稿・解説記事。 11
- **未確認情報の扱い**：推定・噂・推奨構成は「推定」と明示し、仕様・性能・価格の断定を避ける。 12

主要タイムライン

```
timeline
    title Grok系の主要更新（公式情報で確認できる範囲）
    2023-11-03 : Grok 発表（製品コンセプト提示）
    2024-03-17 : Grok-1 オープンリリース（314B MoE）
    2025-07-09 : Grok 4（RLを大規模化、Grok 4 Heavy）
    2025-09-19 : Grok 4 Fast（推論/非推論統合、2M文脈）
    2025-11-17 : Grok 4.1（実運用会話品質、幻覚低減）
    2025-11-19 : Grok 4.1 Fast＋Agent Tools API（ツール呼び出し特化）
    2026-02時点 : Grok 420 / 420 Multi-Agent（APIは "coming soon"）
```

（出典：xAI公式ニュース／公式Docs） 13

公式情報に基づく新機能・変更点

Grok 4を起点とした「推論＋ツール＋大規模RL」の一本道

xAIはGrok 4で、**200,000 GPUクラスター「Colossus」上で、推論能力を“事前学習スケールでのRL訓練”により強化**したと説明している。加えて、計算効率を6倍にしたインフラ・アルゴリズム面の改善と、検証可能データ（当初の数学・コーディング中心から多領域へ）拡張を述べている。 14
同時に、**ネイティブツール使用（Web/X検索やコード実行等）**をRLで学習した、と位置づけている。 14

Grok 4 Heavyと「並列テスト時算出」の方向性

Grok 4の同記事内で、xAIは **「parallel test-time compute」により複数仮説を同時に検討するモデルとしてGrok 4 Heavy を提示**している。UI上も「Agent 1/2/3」が並列に走る表現があり、“内部で複数エージェントが走り最後に統合する”設計が少なくともGrok 4 Heavyでは公式に語られている。 14
この系譜がユーザーが指す「Grok 4.20」文脈の“Multi-Agent（複数エージェント協調）”と整合する可能性は高いが、**4.20固有の実装（エージェント数、役割、計算量、開示範囲）は公式には未提示**である。 4

Grok 4 Fastの「統合アーキテクチャ＋2M文脈＋低コスト化」

Grok 4 Fastでは、**推論（長い思考）と非推論（即応）を“同一のモデル重み”で扱い、システムプロンプト等で挙動を制御する統合アーキテクチャ**が説明されている。これによりエンドツーエンド遅延やトークンコストを下げ、リアルタイム用途に向く、とされる。 15
また、2Mトークンのコンテキスト、Web/X検索などのエージェント能力を強調している。 15

Grok 4.1の「会話品質最適化（スタイル・人格・整合性）と幻覚低減」

Grok　4.1は、**創造性・感情理解・協調的対話など“実運用の会話品質”を改善**したモデルとして紹介され、Grok　4で用いた大規模RL基盤を、スタイル／人格／helpfulness／alignmentの最適化に適用したとしている。[16]
さらに、2週間のサイレントロールアウト中にライブトラフィックでブラインド比較評価を行い、旧本番モデルに対して**64.78%の選好**を得たと報告している。[16]

Grok 4.1 Fast＋Agent Tools APIの「ツール呼び出しの製品化」

xAIはAPI向けに、**Grok 4.1 Fast（2M文脈、ツール呼び出しに最適化）**と、**X検索・Web検索・コード実行・ファイル検索・MCP等を“サーバーサイドで提供するAgent Tools API”**を同時に発表している。[17]
重要なのは「ツールがxAIインフラ上で実行され、開発者が外部API鍵やサンドボックス等を運用せずに済む」設計思想で、エージェントをプロダクション実装する際の運用負債を削る方向性が明確である。[18]

Grok 420（= 4.20/420系の公式に確定できること）

現時点で公式Docsが確実に述べているのは、**「Grok　420」「Grok　420　Multi-Agent」がAPIに“coming soon”で、早期アクセス募集がある**ことまでである。[1]
したがって、「Grok　4.20」の新機能を公式に列挙することはできず、現段階では　**“既存のGrok　4系（4/4 Fast/4.1/4.1 Fast）で確認できる能力の延長線上に、420 Multi-Agentが位置づく”**という構造の提示が限界となる。[4]

API仕様として確定できる範囲

xAIのREST　APIリファレンスでは、チャット系エンドポイントとして　`/v1/chat/completions`　と　`/v1/responses`　が提示されている。[19]
また、**Batch API**や**動画生成系エンドポイント**が新規扱いでメニューに現れており、少なくともドキュメント体系上は動画生成を含む機能拡張が進んでいる。[19]
データ面では、デフォルトのAPIエンドポイントが　`https://api.x.ai`　で、要件がある場合にリージョン固定の　`https://<region>.api.x.ai`　を使える設計が説明されている。[20]

技術的特徴・アーキテクチャ

学習・最適化の中核は「大規模RL」と「検証可能データ拡張」

Grok 4では、**RL訓練を“事前学習スケール”で回す**という方針（Grok 3 Reasoningで見えたスケーリング則を踏まえる）が明示され、加えて計算効率向上（6x）とデータ収集拡張（検証可能データの多領域化）が述べられている。[14]
モデルカード側でも、学習パイプラインとして、**公開インターネット、第三者提供データ、ユーザー／契約者データ、内部生成データ**を含むデータレシピ、前処理（重複排除・分類等）、さらに　**人間フィードバック・検証可能報酬・モデル採点（model grading）を含む複数のRL手法とSFT**が言及される。[21]

Grok 4.1でも同様に、前学習→（能力改善のための）mid-training→SFT＋RLHF＋検証可能報酬＋モデルベース採点という組み合わせが記載され、加えて“新しいより強固な入力フィルタモデル”などデプロイ面の変更が言及される。[22]

アーキテクチャ詳細（パラメータ数等）の公開状況

- **Grok 4 / 4.1 / 4.1 Fast（商用フラッグシップ）**：本調査で参照した公式一次情報では、**パラメータ数・専門家数（MoE構成）・学習トークン総量の定量値は未公表**である（＝未指定）。[23]
- **参考：Grok-1（オープンリリース）**：xAIはGrok-1について、**314BパラメータのMixture-of-Expertsで、トークンあたり25%の重みがactive**、JAX＋Rust基盤でスクラッチ学習、など構成を明示している。[24]
  ただし、Grok-1の公開仕様をGrok 4.20へ外挿することはできない（世代も目的も異なる）。[25]

「Grok 4.20のMulti-Agent」は何を意味しうるか（推定を明示）

公式に確定できるのは、**420に“Multi-Agent”版が存在する**というラベルまでである。[1]
ただし、Grok 4 Heavyが「並列テスト時算出（parallel test-time compute）」として複数エージェント並列実行を示している点から、以下のような実装が“同系統の設計”として推定される（推定）。

```
flowchart LR
    U[User Prompt] --> R{Router / Policy}
    R -->|simple| S[Single-model response]
    R -->|complex| A1[Agent 1: hypothesis]
    R -->|complex| A2[Agent 2: hypothesis]
    R -->|complex| A3[Agent 3: hypothesis]
    A1 --> J[Judge / Aggregator]
    A2 --> J
    A3 --> J
    J --> O[Final Answer]
```

（根拠：Grok 4 Heavyが並列エージェントとテスト時算出を公式に述べる一方、420 Multi-Agentの具体は未公開）[26]

性能評価とコスト比較

ベンチマーク比較の前提と限界

ベンチマーク比較では、同名ベンチでも **（a）pass@1か、投票・多数決・並列TTCか、（b）ツール使用の有無、（c）スキャフォールド（エージェント用ツール、編集ツール等）、（d）問題サブセット** が異なると数値は直接比較できない。
たとえばxAIはGrok 4 Fastの表で「no tools」を明示するベンチを含める一方、OpenAIはGPT-4.1の記事で多数のベンチと一部脚注（例：tau-benchは複数回平均等）を提示している。[27]
このため、以下の表は「同一条件での厳密な横並び」ではなく、**“各社が公式に公開している評価値を、条件情報とセットで並記する”**形式を採る。

主要ベンチマーク（公式公開値を中心）

ベンチマーク	条件（公開情報ベース）	Grok 4 Fast	Grok 4	GPT-4.1	GPT-4o (2024-11-20)
GPQA Diamond	xAI表：pass@1（条件詳細は記事内表）/ OpenAI表：学術知識カテゴリ	85.7%	87.5%	66.3%	46.0%

ベンチマーク	条件（公開情報ベース）	Grok 4 Fast	Grok 4	GPT-4.1	GPT-4o (2024-11-20)
AIME	xAI表：**AIME 2025 (no tools)** / OpenAI表：**AIME '24**	92.0%	91.7%	48.1%	13.1%
MMLU	OpenAI表（xAI側は同表で未提示）	—	—	90.2%	85.7%
SWE-bench Verified	OpenAI表（xAI側は同表で未提示）	—	—	54.6%	33.2%

（出典：xAI「Grok 4 Fast」記事のベンチ表、OpenAI「GPT-4.1」記事のベンチ表） [27]

補足として、xAIはGrok 4 Fastの記事で **LiveCodeBench（Jan–May）**なども併記し、コーディング・数学系でGrok 4と同等近傍の性能を低コストで達成する主張をしている。 [15]

エージェント／ツール呼び出し性能（xAIが強く押す評価軸）

xAIはGrok 4.1 Fast＋Agent Tools APIで、ツール呼び出しを中心に複数のベンチとコストを提示している。

- **τ^2-bench Telecom：スコア100%／総コスト$105**（"Independent evaluation verified by Artificial Analysis"と注記） [18]
- **Berkeley Function Calling v4：Overall Accuracy 72%／総コスト$400** [18]
- 研究・検索系の内部／外部ベンチとして、Research-Eval / Reka FRAMES / X Browse等でスコアと平均コストを併記（GPT-5、Claude Sonnet 4.5等と比較） [18]

この設計思想は、「Grok 4.20/420 Multi-Agent」が仮に"内部協調"を強める方向であっても、xAIが公式に強化してきた評価軸（長文文脈＋ツール＋検索＋コード実行）と整合する、という意味で戦略的一貫性がある。 [28]

推論速度・スループット・メモリの比較（入手できた範囲）

ここは公式一次情報が乏しく、**推論速度はSLA/第三者測定/プロバイダ公称**が混在するため、条件を明示して比較する。

体系	指標	値	条件・測定法（公開情報）	ハードウェア	出典区分
Grok 4.1 Fast（Non-reasoning）	出力tps	122.5 tokens/s	"Artificial Analysis on the xAI API"に基づくとされる集計	未指定	第三者
Grok 4.1 Fast（Non-reasoning）	TTFT	0.52s	同上	未指定	第三者
OpenAI（Priority processing）	レイテンシSLA	GPT-4.1：**>80 tokens/sec**	p50 request latencyを5分単位で算出、と明記	未指定（OpenAI基盤）	公式

体系	指標	値	条件・測定法（公開情報）	ハードウェア	出典区分
OpenAI（Priority processing）	レイテンシSLA	GPT-4o：**>80 tokens/sec**	同上	未指定（OpenAI基盤）	公式
Llama 4 Maverick（Groq提供ページ）	TOKEN SPEED	~600 tps	プロバイダ（Groq）公称	未指定（Groq基盤）	第三者
Llama 4（Azure AIカタログ）	配置要件	Scout：BF16→int4で単一H100に収容 / Maverick：FP8で単一H100 DGX hostに収容	量子化・配置の説明	H100 / H100 DGX	第三者（MS）

（出典：Grok 4.1 Fastの第三者集計ページ、OpenAI Priority Processingページ、Groqのモデルカード、Azure AIカタログ） [29]

ここで、**Grok 4.20自身の推論速度・スループット・メモリ使用は公式未公開（未指定）**である。 [4]

コスト指標（API単価・ツールコール料金）

xAI（Grok 4.1 Fast / Grok 4 Fast）

- Grok 4.1 Fast：入力$0.20/1M、キャッシュ入力$0.05/1M、出力$0.50/1M、ツールコールは「$5/1000 successful invocations から」と記載。 [18]
- Grok 4 Fast：<128kで入力$0.20/1M・出力$0.50/1M、≥128kで入力$0.40/1M・出力$1.00/1M、キャッシュ入力$0.05/1M。 [15]
- ツール料金（例）：Web Search / X Search / Code Executionは $5/1K calls の記載がある。 [30]

OpenAI

- GPT-4.1：入力$2/1M、キャッシュ入力$0.50/1M、出力$8/1M（OpenAI公式）。 [31]
- GPT-4o：入力$2.50/1M、キャッシュ入力$1.25/1M、出力$10/1M（OpenAI公式）。 [32]
- Web searchツール：$10/1K calls（モデル料金とは別に、検索コンテンツトークン課金など条件付きの注意がある）。 [33]

Anthropic

- Claude Sonnet 4.6：入力$3/MTok、出力$15/MTok（公式ページ）。 [34]
- Claude Opus 4.6：入力$5/MTok、出力$25/MTok（公式ページ／ニュース）。 [35]

Llama（オープンウェイト）

Llamaはオープンウェイトであり、**“モデル自体の単価”ではなく、推論基盤（自前GPU／クラウド／推論特化ベンダ）によりコストが支配**される。Meta公式はLlama 4のMoE構成やパラメータ（アクティブ/総数）等を提示するが、推論単価は環境依存である。 [36]

コスト比較チャート（入手可能データの範囲）

```
xychart-beta
  title "主要モデルの出力単価（$/1M output tokens）"
  x-axis ["Grok 4.1 Fast","Grok 4 Fast(<128k)","GPT-4.1","GPT-4o","Claude Sonnet 4.6","Claude
Opus 4.6"]
  y-axis "$/1M output" 0 --> 25
  bar [0.5,0.5,8,10,15,25]
```

（出典：xAI Grok 4 Fast/4.1 Fastの価格、OpenAI GPT-4.1/GPT-4o価格、Anthropic Claude価格） [37]

```
xychart-beta
  title "推論スループット目安（tokens/sec）※混在指標"
  x-axis ["Grok 4.1 Fast(non)","GPT-4.1 priority SLA","GPT-4o priority SLA","Llama4
Maverick(Groq)"]
  y-axis "tokens/sec" 0 --> 650
  bar [122.5,80,80,600]
```

注：OpenAIは「>80 tokens/sec」というSLA表現、Llamaはプロバイダ公称、Grok 4.1 Fast(non)は第三者集計であり、同一条件のベンチではない。 [38]

安全性・有害出力対策、バイアス・フェアネス、プライバシー

xAIの安全性枠組み（RMF）と評価カテゴリ

xAIはRMFで、AIリスクを **「malicious use」と「loss of control」**の二大カテゴリとして整理し、評価すべき行動特性を **abuse potential / concerning propensities / dual-use capabilities**の3バケツに分類している。 [39]
また、一定規模（例：100人超死亡または$1B超被害など）を想定する“catastrophic malicious use events”に対しては、高度なセーフガード適用を述べつつ、**信頼された監査者や契約下の大企業等の“vetted users”には限定的に許可し得る**旨も明記している。 [40]
この設計は、研究用途（安全監査・能力測定）と一般提供を分離する考え方に近い。

Grok 4のモデルカード（定量結果の例）

Grok 4モデルカードでは、危害誘発リクエストへの拒否や、jailbreak／prompt injectionに対する頑健性、エージェント悪用（AgentHarm）・乗っ取り（AgentDojo）などを評価し、表形式で結果を提示している。 [41]
同カードの表（例）では、Refusals系のanswer rateが0に近い値として示され、AgentHarmやAgentDojoなど“エージェント文脈のリスク”も数値化されている。 [42]

また、concerning propensitiesとして **欺瞞（MASK dishonesty rate）、政治的バイアス（内部soft bias）、迎合（sycophancy）**を評価し、数値を提示する。 [43]
さらにdual-useとして、**WMDP（Bio/Chem/Cyber）、VCT、BioLP-Bench、CyBench、MakeMeSay**等を挙げ、特に生物領域の能力が高い点に注意を促している。 [44]

Grok 4.1のモデルカード（入力フィルタ強化と評価修正）

Grok 4.1モデルカードでは、Thinking/Non-thinkingの2構成を明示し、悪用評価（Refusals、AgentHarm、AgentDojo）に加え、**化学・生物の制限領域に対する入力フィルタのfalse negative rate**を表で示している。[45]
加えて、過去カードでは多言語評価設定に誤りがあり英語のみ評価していた点を今回修正し、“真の多言語結果”を報告すると注記している。[22]
こうした訂正は透明性として評価できる一方、旧数値との単純比較を難しくする（＝採用側は評価条件を固定して自前測定する必要がある）。

透明性とガードレールの“実装面”

xAIはGrokのシステムプロンプトを公式GitHubで公開し、Grok 4/4.1等のプロンプトやAPI向け安全プロンプトのファイル名まで列挙している。[46]
この方針は透明性を上げる一方、プロンプト注入やガードレール回避研究の材料にもなり得るため、モデルカードで扱う“jailbreak耐性”評価とセットで理解すべきである。[47]

プライバシー・データ処理・コンプライアンス

- プライバシーポリシーでは、ユーザ入力（Input）と出力（Output）を含む“User Content”が個人情報を含み得て、入力に含めた個人情報が出力に再現され得る旨が明記されている（旧版の明確な条項として確認）。[48]
- API側はリージョン固定エンドポイントを提供し、「特定リージョン内で処理したい」要件に対応する設計を説明している。[20]
- セキュリティFAQではSOC 2 Type 2準拠を述べ、HIPAAについてはBAA手続き案内を示している。[49]
- なお、外部報道として、xAIのAPIキー漏洩事案（GitHub上での露出）が指摘されており、運用管理・鍵管理の成熟度は導入前評価に含めるべきリスクとなる。[50]

外部研究から見た「エージェント化のセキュリティ課題」（参考）

学術的には、ツールや複数エージェント枠組み（AutoGen / CrewAI等）上で、モデルがどの程度攻撃シナリオを拒否できるかを比較する研究も出ており、エージェント化が従来のチャット安全策だけでは守り切れない可能性を示す。対象モデルはGrok 2などでありGrok 4.20直接ではないが、“エージェント運用のリスク一般”として重要である。[51]

実運用での評判と導入推奨

コミュニティの反応（肯定・否定の典型パターン）

日本語圏では、Zenn[52]の記事で「Grok 4.20自体は公式サイトだけでは断定できないことが多い」「APIモデル一覧に“4.20”が見当たらない」など、“一次情報不足”を前提に整理する投稿が見られる。[53]
英語圏ではHacker News[6]で「スレッドはGrokへのリンクだけで、仕様やベンチがないと議論しづらい」というコメントがあり、外部の技術者コミュニティからも情報開示不足が課題として捉えられている。[54]

一方、Grok 4系が公式に強調している差別化点――**X検索を含むリアルタイム情報アクセスと、ツール呼び出しのネイティブ統合**――は、投資・カスタマーサポート・調査など“外部情報の変化が速い領域”では魅力として語られやすい。[55]

また、報道レベルではイーロン・マスク[56]が“Grok 4.20 public beta”に言及したとされるが、X原文の一次確認が本調査環境では困難であり、ここは報道情報として扱うに留める。[57]

競合比較（性能・コスト・用途の比較軸を明示）

以下の比較軸を採用する。

- **能力（公式ベンチ）**：GPQA / AIME / MMLU / SWE-bench等（ただし条件差に注意） [27]
- **エージェント適性**：ツール呼び出しの製品化、検索・コード実行・RAG統合の容易さ [58]
- **コスト**：入出力単価、キャッシュ、ツールコール料金 [59]
- **スループット**：SLA/公称/第三者測定（混在） [60]
- **透明性・ガバナンス**：モデルカード、RMF、プロンプト公開、法務文書 [61]

主要モデルの“導入設計”観点まとめ

陣営	強み（一次情報ベース）	注意点（一次情報ベース）
xAI（Grok 4系）	2M文脈（Fast系）、Agent Tools API（X/Web検索、コード実行、ファイル検索、MCP）を“サーバーサイドで提供”し、エージェント実装の運用負債を削減する思想が明確。 [62]	Grok 4.20/420の仕様・性能・価格・モデルカードが現時点で確定しにくく、採用判断が“待ち”になりやすい。 [4]
OpenAI [63]	GPT-4.1は多数のベンチと価格を公式に提示し、Priority processing等のSLA設計もあり、運用・予算計画が立てやすい。 [64]	ツール料金（例：web search）は別体系で、トークン課金と合わせた総コスト設計が必要。 [65]
Anthropic [66]	Claude 4系はコーディング・エージェント文脈を強く打ち出し、価格体系（Sonnet/Opus等）も公式に整理。 [67]	速度・スループットの公式定量値は相対的に見えにくく、用途によりPoCで体感評価が必要。 [68]
Meta [69]（Llama 3/4）	オープンウェイトで自前運用・最適化が可能。Llama 4はMoEでアクティブ/総パラメータ、学習データ性質、知識カットオフ等を明示するモデルカードがある。 [70]	推論コスト・速度は提供基盤依存。ベンチ提出バリアント問題など、評価の再現性リスクが報道されている。 [71]
Google DeepMind [72]（Gemini）	ベンチの手法・結果をPDFで詳細に開示する枠組みがあり、評価方法の透明性が高い。 [73]	価格は利用API/プランで差が出るため、導入経路（Gemini API / Vertex AI等）を先に固定する必要がある。 [74]

用途別の結論と推奨

研究用途（安全性・能力評価、エージェント研究）

- **推奨**：Grok 4/4.1のモデルカードとRMFが揃っており、拒否・欺瞞・デュアルユースなどの評価指標も明示されているため、“評価設計の叩き台”としては有用。 [75]
- **注意**：Grok 4.20固有の評価は未公開のため、4.20での研究を主張する場合はデータ取得条件を厳密に固定し、再現可能なログ（プロンプト、ツール使用、モデルバージョン）を必ず残すべき。 [4]

プロダクト組込み（API）

- **推奨（現実解）**：当面は **Grok 4.1 Fast＋Agent Tools API** をベースラインにし、2M文脈＋ツール呼び出しの強みを活かす。価格とツールコール体系が公式に出ているため、原価設計が可能。 [76]
- **導入上の注意**：リージョン固定要件がある場合はリージョンエンドポイントを使う設計にする。 [20]

- **4.20/420待ちの判断基準**：公式に（1）モデル名、（2）価格、（3）コンテキスト長、（4）Multi-Agentの計算量（課金体系）、（5）モデルカード相当の安全性情報、が揃ってから段階導入。 1

チャットボット（コンシューマ／社内）

- **推奨**：会話品質・幻覚低減を狙うなら、Grok 4.1系の設計思想（スタイル最適化、幻覚率評価）に沿って、検索ツール併用＋幻覚監視を組み合わせる。 77
- **注意**：プライバシーポリシー上、入力に含めた個人情報が出力に再現され得る点は、個情法・GDPR対応の観点で設計時に明示（マスキング、PIIフィルタ、保持期間ポリシー）すべき。 78

生成コンテンツ（文章／画像／動画）

- **推奨**：APIドキュメント体系上、画像・動画生成エンドポイントが整備されつつあるため、xAIのマルチモーダル戦略は拡張局面にある。 79
- **注意**：生成系は権利・安全性・誤情報の統制が難しく、AUPや社内ポリシー（禁止領域、ラベリング、監査ログ）とセットで導入する。 80

出典URL一覧（主要な一次情報を優先）

```
■ xAI（一次）
https://docs.x.ai/developers/models
https://docs.x.ai/developers/rest-api-reference/inference/chat
https://docs.x.ai/developers/regions
https://docs.x.ai/developers/faq/security
https://x.ai/news/grok-4
https://x.ai/news/grok-4-fast
https://x.ai/news/grok-4-1
https://x.ai/news/grok-4-1-fast
https://data.x.ai/2025-08-20-grok-4-model-card.pdf
https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf
https://data.x.ai/2025-08-20-xai-risk-management-framework.pdf
https://x.ai/legal/privacy-policy
https://x.ai/legal/acceptable-use-policy
https://x.ai/legal/faq
https://github.com/xai-org/grok-prompts
https://x.ai/news/grok-os

■ OpenAI（一次）
https://openai.com/index/gpt-4-1/
https://developers.openai.com/api/docs/models/gpt-4o
https://openai.com/api/pricing/
https://openai.com/api-priority-processing/

■ Anthropic（一次）
https://docs.anthropic.com/en/docs/about-claude/pricing
https://www.anthropic.com/news/claude-4
https://www.anthropic.com/pricing
https://www.anthropic.com/news/claude-opus-4-6

■ Meta（一次＋準一次）
https://ai.meta.com/blog/llama-4-multimodal-intelligence/
```

```
https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/
https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E

■ Google（一次）
https://deepmind.google/models/model-cards/gemini-3-1-pro/
https://deepmind.google/models/evals-methodology/gemini-3-1-pro
https://ai.google.dev/gemini-api/docs/pricing
https://cloud.google.com/vertex-ai/generative-ai/pricing

■ 実運用の評判（参考：二次/コミュニティ）
https://news.ycombinator.com/item?id=46171596
https://news.ycombinator.com/item?id=47055701
https://zenn.dev/ainohogosya/articles/6a470f78e62c48
https://www.reddit.com/r/singularity/comments/1lw4e95/
question_why_isnt_grok_4_on_lmarena_or_devarena/
https://timesofindia.indiatimes.com/technology/social/elon-musk-says-grok-4-20-public-beta-
is-now-available-capabilities-of-ai-chatbot-offered-by-xai/articleshow/128499381.cms
```

（注：上記のうち、一次情報＝公式／モデルカード／公式Docs／公式ニュースを優先。コミュニティ・メディア情報は「評判」や「未確認情報の存在」を示す用途に限定し、仕様の断定には用いない。） [81]

---

[1] [4] [30] [81] https://docs.x.ai/developers/models
https://docs.x.ai/developers/models

[2] [5] [10] [14] [23] [26] https://x.ai/news/grok-4
https://x.ai/news/grok-4

[3] [21] [41] [42] [43] [44] [61] [75] https://data.x.ai/2025-08-20-grok-4-model-card.pdf
https://data.x.ai/2025-08-20-grok-4-model-card.pdf

[6] [15] [27] [69] https://x.ai/news/grok-4-fast
https://x.ai/news/grok-4-fast

[7] [54] [66] https://news.ycombinator.com/item?id=47055701
https://news.ycombinator.com/item?id=47055701

[8] [17] [18] [28] [37] [55] [56] [58] [59] [62] [63] [76] https://x.ai/news/grok-4-1-fast
https://x.ai/news/grok-4-1-fast

[9] [57] https://timesofindia.indiatimes.com/technology/social/elon-musk-says-grok-4-20-public-beta-is-now-available-capabilities-of-ai-chatbot-offered-by-xai/articleshow/128499381.cms
https://timesofindia.indiatimes.com/technology/social/elon-musk-says-grok-4-20-public-beta-is-now-available-capabilities-of-ai-chatbot-offered-by-xai/articleshow/128499381.cms

[11] https://news.ycombinator.com/item?id=46171596
https://news.ycombinator.com/item?id=46171596

[12] [53] https://zenn.dev/ainohogosya/articles/6a470f78e62c48
https://zenn.dev/ainohogosya/articles/6a470f78e62c48

[13] https://x.ai/news/grok
https://x.ai/news/grok

16 77 https://x.ai/news/grok-4-1
https://x.ai/news/grok-4-1

19 79 https://docs.x.ai/developers/rest-api-reference/inference/chat
https://docs.x.ai/developers/rest-api-reference/inference/chat

20 https://docs.x.ai/developers/regions
https://docs.x.ai/developers/regions

22 45 https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf
https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf

24 25 https://x.ai/news/grok-os
https://x.ai/news/grok-os

29 38 https://shadabchow.com/blogs/ai-models/grok-4-1-fast-non-reasoning
https://shadabchow.com/blogs/ai-models/grok-4-1-fast-non-reasoning

31 64 https://openai.com/index/gpt-4-1/
https://openai.com/index/gpt-4-1/

32 https://developers.openai.com/api/docs/models/gpt-4o
https://developers.openai.com/api/docs/models/gpt-4o

33 65 https://openai.com/api/pricing/
https://openai.com/api/pricing/

34 https://www.anthropic.com/claude/sonnet
https://www.anthropic.com/claude/sonnet

35 https://www.anthropic.com/claude/opus
https://www.anthropic.com/claude/opus

36 70 https://ai.meta.com/blog/llama-4-multimodal-intelligence/
https://ai.meta.com/blog/llama-4-multimodal-intelligence/

39 40 52 https://data.x.ai/2025-08-20-xai-risk-management-framework.pdf
https://data.x.ai/2025-08-20-xai-risk-management-framework.pdf

46 47 72 https://github.com/xai-org/grok-prompts
https://github.com/xai-org/grok-prompts

48 78 https://x.ai/legal/privacy-policy/previous-2024-12-20
https://x.ai/legal/privacy-policy/previous-2024-12-20

49 https://docs.x.ai/developers/faq/security
https://docs.x.ai/developers/faq/security

50 https://www.techradar.com/pro/security/doge-employee-with-sensitive-database-access-leaks-private-xai-api-key
https://www.techradar.com/pro/security/doge-employee-with-sensitive-database-access-leaks-private-xai-api-key

51 https://arxiv.org/abs/2512.14860
https://arxiv.org/abs/2512.14860

60 https://openai.com/api-priority-processing/
https://openai.com/api-priority-processing/

67 68 https://www.anthropic.com/pricing
https://www.anthropic.com/pricing

71 https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming
https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming

73 https://deepmind.google/models/model-cards/gemini-3-1-pro/
https://deepmind.google/models/model-cards/gemini-3-1-pro/

74 https://ai.google.dev/gemini-api/docs/pricing
https://ai.google.dev/gemini-api/docs/pricing

80 https://x.ai/legal/acceptable-use-policy
https://x.ai/legal/acceptable-use-policy