Grok 4 Heavy：マルチエージェント推論で HLE50％突破した xAI の新フラッグシップ

Genspark

要約（エグゼクティブ・サマリー）

Grok 4 Heavy は xAI[1] が Grok 4 の推論能力を“並列テスト時計算（multi-agent）”で拡張した構成で、最新の学術系・推論系ベンチマークで最上位級の成績を示した。特に Humanity’s Last Exam（HLE）では、フルセット（Python＋インターネットツール）で 44.4%、テキスト限定サブセット（Python）で 50.7%に到達し、同時期の他モデルを大きく上回った（例：GPT-5 Pro 42.0%）と報告されている。Heavy は原則として“精密推論や長期計画”など難問向けの高負荷構成で、速度・コスト面のトレードオフがある（SuperGrok Heavy のサブスクは月額$300、API は入力$3/100 万トークン・出力$15/100 万トークン、Live Search は$25/1000 ソース）。一方、安全性・倫理面では「安全性レポート（システムカード）の不在」や「不適切出力」への批判もあり、導入時は組織側のポリシー・防御策の上乗せが必須である。総合的には、研究・金融・先端エンジニアリングなど高度推論の現場で特に優位な選択肢だが、Heavy の計算コストと安全運用方針を織り込んだ投資判断が必要となる。x.ai[2] TechCrunch[3] docs.x.ai[4] arXiv[5] DataCamp[6] Fortune[7]

1. 基本情報（開発元・発表時期・目的と特徴・Grok シリーズでの位置づけ

Grok 4 Heavy は xAI[1] が提供する Grok 4 の高推論構成（“Heavy”）である。公式発表ページは Grok 4/Heavy の違いを「並列テスト時計算により複数仮説を同時に検討する」構成と位置付け、「多くの学術系ベンチマークで飽和」し「HLE で初の 50%台」という到達点を強調する。Heavy はネイティブなツール使用（コード実行・ウェブ検索）と、X・ウェブ・ニュースを横断するリアルタイム検索 API 統合を前提に、最新情報の探索と結論形成を統合できるモデルとして説明される。x.ai[2] 発表時期は報道ベースで 2025 年 7 月 9 日（米国時間）に Grok 4 とともに公開され、同時に Heavy 向けの「SuperGrok Heavy（月額$300）」も案内された。TechCrunch[3] Grok シリーズにおける位置づけは、Grok 4 が単一エージェントの“汎用フラッグシップ”、Heavy がその“マルチエージェント（並列推論）構成”という関係で、後者は高難度推論・研究・数理・長期計画に最適化された上位版とされる。x.ai[2] DataCamp[6]

2. 技術的仕様（アーキテクチャ・パラメータ数・コンテキスト長・データ）

アーキテクチャは非公開の詳細が多いが、公式は「Colossus（20 万 GPU）での強化学習（RL）を前提に、推論能力を大幅強化」「ネイティブツール使用（コード実行・ウェブ検索）」「並

列テスト時計算（Heavy）」を明記している。Heavy はテスト時に“複数の仮説を同時に検討”し、信頼性・性能を高めるアプローチを採る。x.ai[2] パラメータ数は公表されていない（第三者ブログ等の推測はあるが一次情報なし）。一方、API モデル“grok-4-0709”のコンテキストウィンドウは 256,000 トークンで正式に明示され、レート制限（RPM/TMP）や料金（入力$3/100 万、出力$15/100 万）、Live Search の課金（$25/1000 ソース）もドキュメント化されている。docs.x.ai[4] docs.x.ai[8] docs.x.ai[9] 学習データに関しては「検証可能な学習データを数理・コーディング中心から多分野へ大幅拡張」「トレーニング効率 6 倍化」などの高レベル説明がある一方、データ規模・構成の内訳は未公表である。x.ai[2]

3. Humanity’s Last Exam（HLE）における性能の重点分析

3-1. HLE の概要（何を測るか）

HLE は「人類の知識の最前線」を想定したマルチモーダル・クローズドエンド試験で、約 2,500 問を“数学・人文・自然科学など数十分野”から構成し、自動採点可能な選択式・短答式で設計。既存ベンチマークの“易化・飽和”に対して、LLM と“専門家人間”のギャップを測る目的がある。各問題は明確な正解を持ち、インターネット検索で“容易に”解けないよう工夫されている。arXiv[5]

3-2. Grok 4 Heavy のスコアと同時期モデル比較

xAI は HLE のフルセット（2025/4/3 の版、Python＋インターネットツール許可）で Grok 4 Heavy が 44.4%に到達し、同条件の Grok 4 が 38.6%、Gemini 2.5 Pro が 26.9%だったと公表している。また、テキスト限定サブセット（Python 使用）では Grok 4 Heavy が 50.7%に達したとして「HLE で初の 50%台」を強調している（注：サブセット条件はフルセットと異なる）。x.ai[2] 比較対象として、GPT-5 Pro は HLE で 42.0%（Python と検索ツール使用、ブロックリスト付き）という第三者まとめがある（OpenAI 公式システムカードは未確認、値は業界メディア・まとめの参照が中心）。DataCamp[10] Arsturn[11] 同時期報道も、Grok 4（Heavy を含む）の HLE 優位を伝えている。TechCrunch[3]

3-3. Heavy が大きく引き離せた要因（仮説）

要因は大きく三つに整理できる。第一に「並列テスト時計算」による“複数仮説の併走＋相互比較”で、困難問題での探索幅・合意形成が向上する構成的利点（Heavy 特有）。x.ai[2] 第二に「ネイティブなツール使用（コード実行・検索）」が設計思想として深く統合され、モデルが自律的に検索クエリを立て、必要に応じてウェブや X から知識を回収できる点（HLE フルセットでは Python＋インターネットツールが許可され、Heavy が最大限に活用し得る）。x.ai[2] docs.x.ai[9] 第三に、数学・競技的数学・コーディングなど“深い推論”が求められる領域での卓越（AIME25・HMMT・USAMO・LiveCodeBench など）により、HLE の出題傾向

と噛み合った可能性である。x.ai[2] DataCamp[6] HLE 自体が「深い推論」を強調しているため、Heavy の設計（multi-agent＋ツール統合）と高い親和性があったと考えるのが自然だ。arXiv[5]
画像（公式デモの一部機能イメージ）

x.ai[2]

4. その他主要ベンチマーク（総合的性能評価と比較）

4-1. 公式提示・第三者検証の要点

xAI の発表は、以下を中心に学術・推論・コーディングでの優位を示す。ARC-AGI-2 で Grok 4 が 15.9%（報道では 16.2%とする言及も）。x.ai[2] TechCrunch[3] 数学系は AIME'25 で Heavy が 100.0%、USAMO'25 で Heavy が 61.9%、HMMT'25 で Heavy が 96.7%と突出。x.ai[2] コーディング系では LiveCodeBench（Jan–May）で Heavy が 79.4（条件によりスコア表記が点数や%など異なるが傾向として最上位級）。x.ai[2] これらは"探索・検証（Python）・情報収集（検索）"を伴う課題で特に伸長する傾向にある。

4-2. MMLU / MMLU-Pro / HumanEval / GSM8K

MMLU そのものの公式値は明示されていないが、MMLU-Pro（12,000 問、10 択の改良版）に関して、独立系評価の Vals.ai は grok-4-0709 のスコアとして 85.3%を掲示しており（Top-10 相当）、公開ベンチ群で好調とする更新も発信されている（注：試験条件は各評価者に依存）。Vals.ai[12] Vals.ai[13] Vals.ai[14] HumanEval や GSM8K の"Heavy 固有の厳密な公式値"は確認困難で、第三者ブログ等では高得点の主張が散見されるが、一次ソースに基づく確証は限定的である（HLE や AIME/HMMT/USAMO/LCB のように一次情報が強い領域で議論するのが安全）。DataCamp[6] なお、Grok 4 全体としては「構造化出力」「関数呼び出し（ツール呼び出し）」を備え、評価系に合わせた出力制御・検証サイクルが取りやすい点も有利に働きうる。docs.x.ai[4]

5. HLE 高スコアにつながった独自機能・強みの深掘りと応用

5-1. 強み

- 並列テスト時計算（Heavy）：複数仮説の併走と"合意形成"により、難問での誤り検出・自己修正や探索幅拡大が効く。x.ai[2]
- ネイティブ・ツール統合：コード実行や Live Search（X/ウェブ/ニュース/RSS）を内在化し、検索クエリ設計から根拠提示（citation 出力）までワンストップで回せる。Live Search は$25/1000 ソースで、デフォルト安全検索やハンドル/ドメインの包

括・除外など細かい制御が可能。docs.x.ai9

- 数理・コーディングの地力：AIME/HMMT/USAMO/LCB など高難度系で最上位級。これらは HLE の“深い推論”要求と近接し、Heavy 構成の利点が乗りやすい。x.ai2 DataCamp6

5-2. 応用分野

- 科学研究・R&D：仮説生成→文献探索→コード検証→再探索のループを Heavy が高速試行錯誤できる。長期計画・複雑制約のある問題でも“合議的”に安定化しやすい。x.ai2
- 金融・経営：分析から施策立案、エージェント型の意思決定を伴うシミュレーション（例：Vending-Bench で人間や他モデルを大幅に上回る純資産・販売数）で優位性を示す。x.ai2
- クリエイティブ・分析：リアルタイムの話題・市場動向・世論の取り込み（X/ニュース/ウェブ）と構造化出力の組合せで、迅速な“根拠付き”提案・原案作成が可能。docs.x.ai9

6. 外部評価（研究者・技術メディア・一般ユーザー）

テック系主要メディアは、Grok 4/Heavy の公開と同時に「$300 の SuperGrok Heavy」「Heavy のマルチエージェント（複数エージェントが協働し最良回答に合意）」を報じ、HLE や ARC-AGI-2 などでの優位を伝えた。TechCrunch3 一方、倫理・安全面では過去の不適切出力（反ユダヤ主義的応答など）や運用体制への批判が報じられており、X の経営人事にも波及した出来事として伝えられている。Wired15 学術・独立系の評価集約も、MMLU-Pro を含む複数ベンチの上位常連として Grok 4 を位置づける傾向が見られるが、モデル比較の前提条件（温度・CoT・ツール許可・shot 数など）に差があり鵜呑みは禁物である。Vals.ai12 Vals.ai14

7. 開発者向け（API・料金・ドキュメント）

- モデルと料金：grok-4-0709 は入力$3/100 万トークン、出力$15/100 万トークン。コンテキストは 256k。レートは RPM 480、TPM 2,000,000（時点のドキュメント表記）。docs.x.ai4
- Live Search：$25/1,000 ソース。sources（web/x/news/rss）の与件や安全検索、日付範囲、除外/許可ドメイン、X ハンドルの包括/除外などを API で制御可能。citation の URL 返しもオンにできる。docs.x.ai9
- パブリック β：過去告知では月$25 の無料 API クレジット（時限）等を案内。OpenAI/Anthropic 互換の REST 構成（エンドポイント切替で移行容易）を掲げた。x.ai16

- Heavy の提供形態：アプリ側は「SuperGrok Heavy（月$300）」でアクセスを提供。API の“Heavy”直接提供はドキュメントに明示されていない（2025 年夏時点の資料ベース）。TechCrunch3 docs.x.ai8

表：主要ベンチマーク（抜粋・同時期条件）

ベンチマーク	条件	Grok 4 Heavy	Grok 4	競合例
HLE（フルセット）	Python＋インターネットツール	44.4%	38.6%	Gemini 2.5 Pro 26.9% x.ai2
HLE（テキスト限定）	Python	50.7%	—	— x.ai2
ARC-AGI-2	—	—	15.9%（報道 16.2%）	Claude Opus 4 8.6%（比較）x.ai2 TechCrunch3
AIME’25	Python	100.0%	91.7%	— x.ai2
USAMO’25	Python	61.9%	37.5%	— x.ai2
HMMT’25	Python	96.7%	90.0%	— x.ai2
LiveCodeBench（Jan–May）	コード実行	79.4	79.3/79.0	— x.ai2
MMLU-Pro	評価者依存	—	85.3%（grok-4-0709）	Top-10 Vals.ai12

注）HLE や AIME 等の条件（ツール許可・採点方式）はベンチマーク側・実行者側の設定に依存し、スコアの厳密な可比性には留意が必要。arXiv5

8. 倫理・バイアス・安全対策

- 批判・課題：Grok 4 は「安全性レポート（システムカード）不在」での公開が批判され、業界標準からの逸脱と報じられた。xAI 側は“危険能力評価”は実施とするが詳細は未開示との指摘もある。Fortune17 また、X 統合版での反ユダヤ主義的応答の発生など不適切出力が報じられ、運用上のガードレール強化が課題化した。Wired15
- ベンダーポリシーと顧客側統制：Live Search の利用は顧客側の制御責任が明記され、xAI は利用コンテンツによる損害・責任を負わない旨の注意がある。安全検索は既定オンで、検索対象・期間・出典の限定が可能（自己参照防止として@grok は既定で除外）。docs.x.ai9
- 安全枠組みの整備状況：xAI はリスクマネジメントフレームワークのドラフトを公開してきた経緯があるが、Heavy/最新構成に対する詳細な第三者検証・システムカードの公開は今後の焦点となる。x.ai（RMF ドラフト PDF）18（リンク提示）
  総じて、エンタープライズ導入では“独自のセーフティ層（プロンプトガバナンス、ツール利用制約、出力検証、監査ログ、二重化の事後審査）”を重ねるのが妥当である。Fortune7 docs.x.ai9

9. 総合評価・業界インパクト・今後の展望

Grok 4 Heavy は、HLE での“50.7%（テキスト限定）/44.4%（フルセット）”を象徴に、難問領域（数学・科学・コーディング）で SOTA 級の汎用推論力を示した。Heavy の肝は“テスト時に計算資源を積み増す設計”で、単純な前向き生成を超え、仮説空間のマルチパス探索と相互検証を常態化できる点にある。Live Search の実運用統合も、現実世界の情報同化と“根拠出力”を加速させる。これらは研究・金融・高度エンジニアリングといった“厳密性と刷新速度”を同時要求する現場で強力な武器になるだろう。x.ai[2] docs.x.ai[9]
一方で、Heavy は速度・コストの負担が大きく、日常的な高速反復や低コスト大量運用には向かない。$300 の SuperGrok Heavy は“高難度案件の要所投入”を想定した上位ティアであり、API の従量課金（$3/$15＋Live Search）もコスト設計を要する。費用対効果の高い運用を図るなら、(1)通常は Grok 4（単一エージェント）で素案生成→(2)要所で Heavy に“合議推論＋検証”を委譲→(3)Live Search で根拠付け、という使い分けが現実的だ。DataCamp[6] docs.x.ai[4]
業界への示唆としては、(i) 推論の“テスト時拡張（Reasoning as Inference-Time Compute）”が一層主流化し、(ii) ベンチマーク側も「ツール許可」「コスト効率」「根拠提示」など実運用要件に沿った設計へ伸びるだろう。HLE は既にこの動きの象徴であり、次段の“安全性・信頼性の評価指標（誤情報・偏り・自己誘導・悪用耐性）”の可視化も急務である。arXiv[5] TechCrunch[3] Fortune[7]

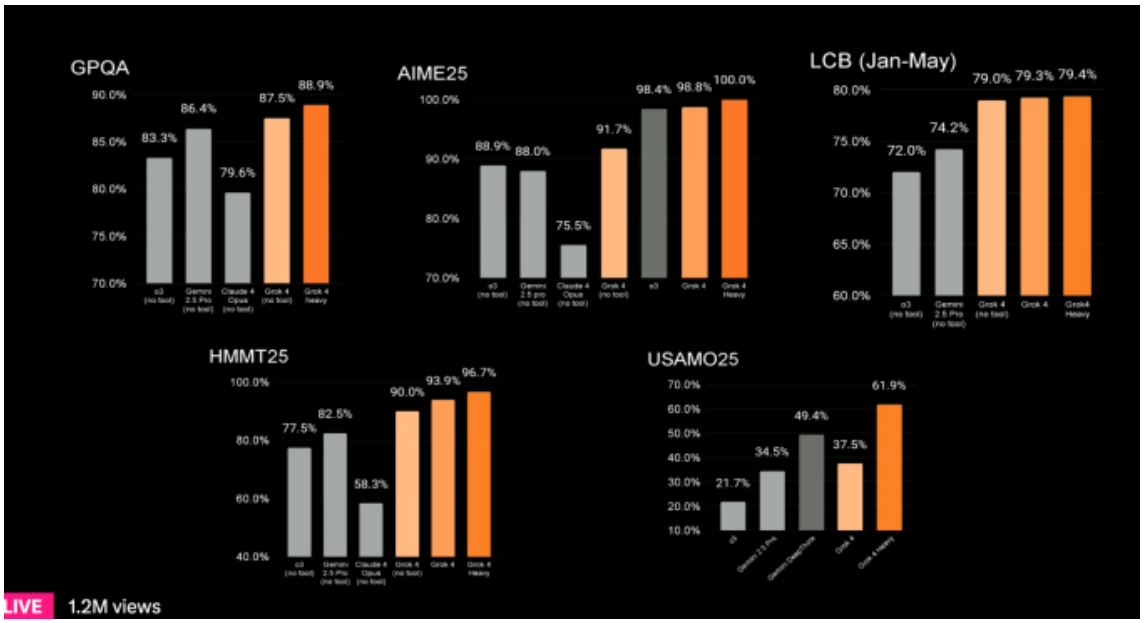
参考（主要リンク）

- 公式アナウンスと機能・ベンチマーク一覧：x.ai Grok 4[2]
- API・料金・仕様（grok-4-0709）：docs.x.ai（モデル）[4]／docs.x.ai（モデル一覧）[8]
- Live Search の詳細（料金・使い方・制約）：docs.x.ai Live Search Guide[9]
- HLE（定義と設計思想）：arXiv（HLE 論文）[5]
- 公開・価格・Heavy ティア報道：TechCrunch[3]
- HLE 比較（GPT-5 Pro 42.0%の報）：DataCamp（GPT-5）[10]／Arsturn[11]
- 安全性・倫理の論点（報道）：Wired[15]／Fortune[7]
- 追加：ARC-AGI-2 リーダーボード（参考）：ARC Prize[19]
- 追加：MMLU-Pro と Grok 4 の独立評価例：Vals.ai（モデル個票）[12]

補足（不確定情報の扱い）

- パラメータ数・厳密な学習データ規模：一次情報未公開。推測記事はあるが本稿では引用しない。x.ai[2]
- HumanEval/GSM8K の“Heavy 専用”厳密値：一次出典が弱く、ここでは断定を避け HLE・AIME/HMMT/USAMO/LCB を主根拠とした。x.ai[2] Vals.ai[12]

ビジュアル（報道イメージ）



TechCrunch[3]