

Grok 4の評判まとめ

結論：Grok 4は「最高峰の推論能力」を誇る一方、**高価格やマスク氏寄りの偏り、応答遅延・安定性**といった課題も指摘されており、利用目的に応じた検討が求められる。

1. 技術的評価とベンチマーク

Grok 4は大規模強化学習を活用し、推論能力に特化した設計を採用。Ph.D.レベルの知的ベンチマーク (Humanity's Last Exam) で38.6%を達成し、OpenAI GPT-4やGemini 2.5 Proを上回った^[1]。Heavy版では44.4%にまで正答率を向上させ、人間レベルの多段階思考を実現している^[1]。

ベンチマーク	Grok 4	GPT-4 (o3)	Gemini 2.5 Pro
Humanity's Last Exam	38.6% ^[1]	~21%	~21.6%
Heavy版 (HLE)	44.4% ^[1]	—	—
ARC-AGI v1 (スコア)	66.6% ^[2]	—	—
ARC-AGI v2 (スコア)	15.9% ^[2]	—	—

2. 利用者のポジティブな声

- **高度な情報収集・要約能力**：X (旧Twitter) やウェブ情報をリアルタイム取得し、自動要約/可視化できる点が高評価^[2]。
- **ライティング支援**：指定文字数に正確に従い、レポートやコンテンツ制作を補助する能力に定評あり^[3]。
- **マルチエージェント機能** (Heavy版)：複数タスクの並行処理により、より高度なシナリオ分析や予測モデル構築が可能^{[2] [4]}。

3. コストと利用環境

月額料金はSuperGrokで約30ドル、Heavy版で300ドルと業界最高水準。API利用料金も入力3ドル/100万トークン、出力15ドル/100万トークンと高額^[5]。

「月額300ドルという高額設定にも関わらず、特定分野での明確な優位点が見えにくい」という指摘もある^[6]。

4. 指摘されるデメリット・懸念点

1. 偏り (バイアス)

- センシティブな話題ではマスク氏の投稿や見解を優先的に引用し、中立的回答を逸脱する挙動が報告されている^{[5] [7]}。
- 設計思想に「Musk寄り」のプリセットが隠されているのではという懸念が強い^[8]。

2. 応答速度・安定性

- Heavy版は推論時間が長く（数分～10分）、実務利用における即時性に課題がある^[2]。
- 複雑クエリでのメモリ不足やクラッシュ報告も散見される^[9]。

3. デザイン・クリエイティブ性能

- ゲームやビジュアル生成においては他モデル（Gemini、Claudeなど）に劣るとのレビューがある^[4]。

5. 総合的な評価と利用シーン

- **最先端の研究・専門的推論**を要する場面：Grok 4の強みが活きる。
- **コンテンツ制作・レポート作成**：文字数制御と要約力が高く、ライティング支援ツールとして有用。
- **コスト重視・クリエイティブ用途**：高額かつデザイン力に限界があるため、他モデル併用を検討すべき。
- **偏りや応答速重視**：センシティブ領域では注意が必要であり、即時応答を要する業務では遅延がボトルネックになる可能性が高い。

以上を踏まえ、Grok 4は「最高峰の知的推論エンジン」としての評価が確立している一方で、コスト・バイアス・安定性といった実務面の課題も明確である。使用目的を明確化し、必要に応じて他モデルとの併用検討が推奨される。

✻

1. <https://ai-market.jp/technology/grok-4/>
2. <https://shift-ai.co.jp/blog/29813/>
3. https://note.com/hitotsu_ai/n/n6eceff5ac1c1
4. <https://romptn.com/article/63770>
5. <https://gigazine.net/news/20250711-xai-grok-4/>
6. <https://zenn.dev/acntechjp/articles/963201c99e9c56>
7. <https://news.aibase.com/ja/news/19620>
8. <https://note.com/akikito/n/nac189f2b2b89>
9. <https://note.com/digistanet/n/n8b098b89243b>