Humanity's Last Exam における各AIモデルの評価結果

概要

Humanity's Last Exam (HLE) は、AIモデルの真の能力を測定するために設計された最も困難なベンチマークの一つです。Center for AI SafetyとScale AIが共同開発し、世界50カ国、500以上の機関から約1,000人の専門家が参加して作成された2,500問の問題で構成されています。このテストは既存のベンチマークが90%以上のスコアに達する「ベンチマーク飽和」問題に対応するため、大学院レベル以上の極めて高難度な問題を含んでいます。[1]

各モデルの評価結果

最高性能モデル群（40%以上）

Grok 4 Heavy: 50.7%[2] [3]

- xAIが開発したマルチエージェント型システム
- 複数のAIエージェントが協調して問題解決
- 現在HLEで最高スコアを記録
- テキストのみのサブセットでの評価結果

GPT-5 Pro (with tools and reasoning): 42.0%[4]

- OpenAIの最新フラグシップモデル
- ツールとリーズニング機能を併用した構成
- 従来のo3モデルから大幅な性能向上を実現

上位性能モデル群（25-40%）

Grok 4 (with tools): 38.6%[5]

- 外部ツールアクセス機能付きのGrok 4
- スタンダード版から大幅な性能向上

Gemini 2.5 Deep Think: 34.8%[6] [7]

- Googleの並列思考技術を採用したモデル
- 2025年国際数学オリンピック（IMO）で金メダルレベルを達成[7]
- 複数のアイデアを同時に検討し最適解を導出

Gemini 2.5 Pro (with tools): 26.9%[8]

- ツール使用機能を有効にしたGemini 2.5 Pro
- LMArenaリーダーボードで1位を獲得[9]

中位性能モデル群（20-25%）

Grok 4: 25.4%[10]

- xAIのスタンダードモデル
- ツール不使用での評価結果

OpenAI o3 (with tools): 24.9%[8]

- OpenAIの推論特化モデル（ツール使用時）

GPT-5: 24.8%[4]

- 推論機能有効、ツール無しでの評価
- ベースGPT-5の6.3%から大幅改善

標準性能モデル群（18-22%）

Gemini 2.5 Pro Preview: 21.64%[1]

- 2025年3月リリースのプレビュー版
- ツール使用なしでの最高スコア

OpenAI o3: 20.32%[1]

- スタンダード版のo3モデル

Gemini 2.5 Pro Experimental: 18.8%[11] [9]

- 実験版でありながら他モデルを上回る性能
- ツール使用なしでの結果

下位性能モデル群（10-12%）

Claude Opus 4.1 (Thinking): 11.52%[12]

- 拡張思考機能付きClaude Opus 4.1
- Anthropicの最新モデル

Claude Opus 4.1: 10.7%[13] [14]

- Anthropicのフラグシップモデルのスタンダード版
- コーディング性能では74.5%のSWE-bench Verifiedスコアを達成

重要な分析ポイント

人間との性能差

人間の専門家はHLEで80-90%のスコアを達成する一方、最高性能のAIモデルでも50.7%に留まっており、**依然として大きな性能差**が存在しています。これはAIが表面的なパターン認識から真の推論能力への移行段階にあることを示しています。[15]

ツール使用の効果

多くのモデルでツール使用により性能が大幅に向上しており、**外部リソースとの統合**がAIの能力拡張において重要な役割を果たしています。

マルチエージェント手法の優位性

Grok 4 Heavyの最高スコアは、**複数のAIエージェントが協調**することで単体モデルを超える性能を実現できることを実証しています。

推論能力の重要性

GPT-5において推論機能の有無で18.5ポイントの差（6.3% → 24.8%）が生じており、**深い推論処理**がHLEのような高難度タスクで決定的な要因となっています。[4]

HLEは単なるベンチマークを超え、AIが人間レベルの専門知識と推論能力に到達するための重要な指標として機能しており、各モデルの真の能力を明らかにする「人類最後の試験」としての役割を果たしています。[1]

⁂

1. https://openai.com/ja-JP/index/introducing-gpt-5/
2. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/
3. https://www.reddit.com/r/singularity/comments/1kudxtf/claude_4_opus_thinking_scores_107_on_humanitys/
4. https://note.com/taku_sid/n/neba39ca7a701
5. https://note.com/npaka/n/nc204b8d8d9ad
6. https://patmcguinness.substack.com/p/claude-opus-41-openais-gpt-oss-and
7. https://www.vellum.ai/blog/gpt-5-benchmarks
8. https://www.adcal-inc.com/column/gemini-2-5-pro/
9. https://www.linkedin.com/posts/mattasmall_gpt-5-vs-claude-opus-41-blog-braintrust-activity-7360353299272798210-YKwP
10. https://zenn.dev/galirage/articles/gpt-5-release
11. https://japan.zdnet.com/article/35230953/
12. https://www.anthropic.com/news/claude-opus-4-1
13. https://www.watch.impress.co.jp/docs/series/nishida/2038398.html
14. https://www.lifehacker.jp/article/2504-gemini-25-pro-is-googles-most-powerful-ai-model-yet-and-its-already-free/

15. https://artificialanalysis.ai/models/claude-4-1-opus

Model	HLE Score (%)	Notes
Grok 4 Heavy	50.7	Multi-agent system, text-only subset
GPT-5 Pro (with tools and reasoning)	42.0	With tools and reasoning enabled
Grok 4 (with tools)	38.6	With external tools
Gemini 2.5 Deep Think	34.8	Parallel thinking approach
Gemini 2.5 Pro (with tools)	26.9	With tools enabled
Grok 4	25.4	Standard version
OpenAI o3 (with tools)	24.9	With tools enabled
GPT-5	24.8	With thinking enabled, no tools
Gemini 2.5 Pro Preview	21.6	Preview version from March 2025
OpenAI o3	20.3	Standard version
Gemini 2.5 Pro Experimental	18.8	Experimental version without tools
Claude Opus 4.1 (Thinking)	11.5	With extended thinking
Claude Opus 4.1	10.7	Regular version