

DeepSeek V4 Pro 評価レポート：米国最先端モデルとの実力差とコスト効率

米国国立標準技術研究所 (NIST) 傘下のCAISIによる評価

ベンチマーク評価の真実

自己申告 vs 第三者評価の乖離



公開データ

非公開データ

DeepSeek技術レポート：
米国モデルに匹敵

CAISI独自の非公開ベンチ
マーク (PortBench等)：
米国モデルに劣る

エージェント能力の検証

PortBench (コード移植能力)

GPT-5.4
xHigh:
60.3%
成功率

DeepSeek
V4 Pro:
43.8%
成功率



米国最先端モデル
(US Models)

中国モデル
(China Models)

OpenAI GPT-5.5
(Eio: 1260±28)

OpenAI
GPT-5.4 mini
(Eio: 749±46)

8ヶ月の
技術的遅れ
(8-Month Lag)

Anthropic

Alibaba

Kimi

DeepSeek V4 Pro
(Elo: 800±28)

米国の最先端技術から約8ヶ月の遅行

CAISIの分析によると、DeepSeek V4 Proの性能は、約8ヶ月前にリリースされた米国の主要モデルと同等水準に留まっています。

数学・科学分野での強み



OTIS-AIME-2025:
97% 精度



Frontier Science:
74% 精度

特定の専門ドメインでは米国のトップモデルに迫る実力

コスト効率の比較

GPT-5.4 miniと比較して5つの指標で低コスト

GPT-5.4 mini (Cost)
入力: \$0.75/1M | 出力: \$4.50/1M
キャッシュ入力: \$0.075

DeepSeek V4 Pro (Cost)
入力: \$1.74/1M | 出力: \$3.48/1M
キャッシュ入力: \$0.0145



GPT-5.4 mini
(Cost)

DeepSeek V4 Pro
(Cost)

最大53%のコスト削減

SWE-Bench等で圧倒的なコスト優位性

ソフトウェア修正タスク (SWE-Bench) では、米国モデルより53%低いコストで実行可能ですが、一部の科学系タスク (GPQA) では逆に41%高くなる場合があります。