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

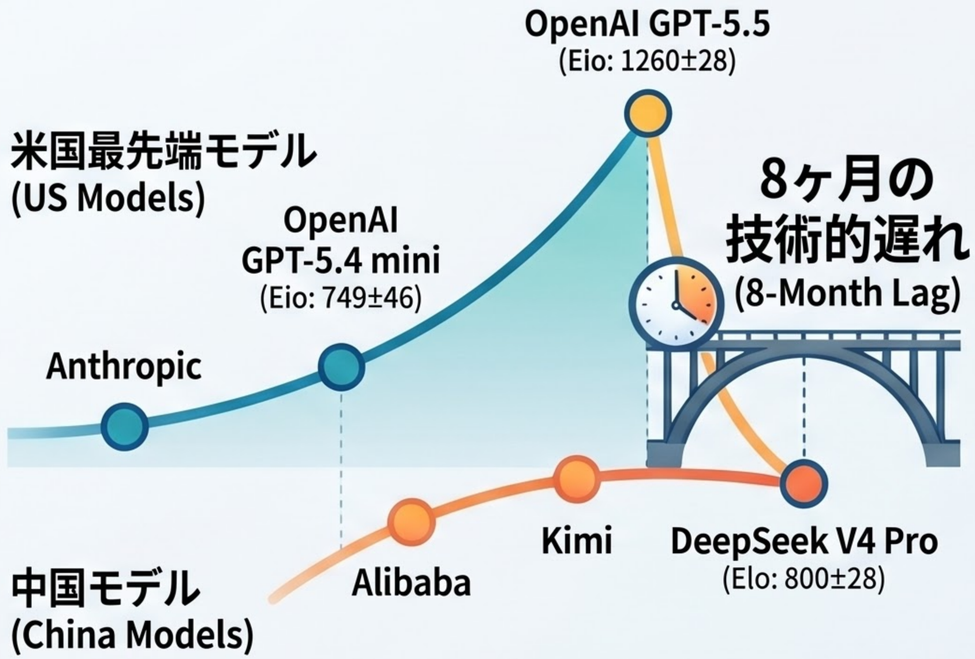
DeepSeek技術レポート：米国モデルに匹敵

非公開データ

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）



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

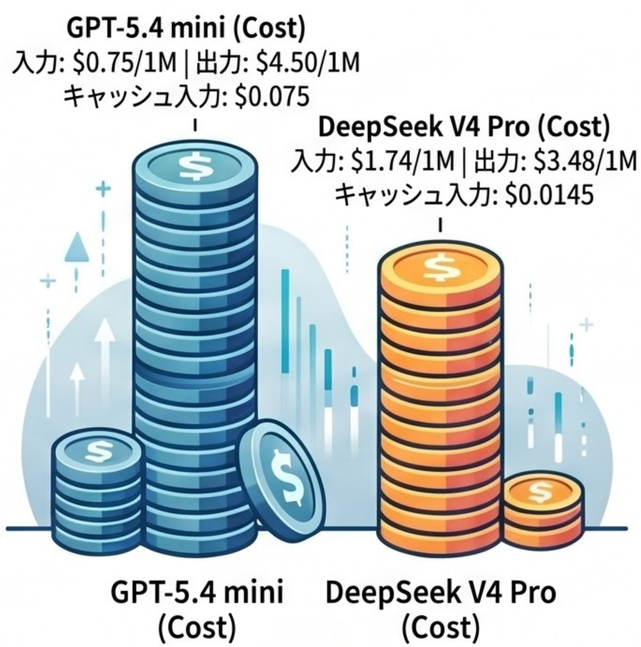
OTIS-AIME-2025:
97% 精度

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



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。