DeepSeek V4 Proの実力評価：米国最先端モデルとの「8か月の差」を読み解く

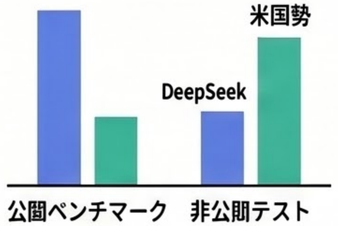
評価の結論：総合能力と非公開タスク

総合能力は「GPT-5相当」に到達

CAISIの分析によれば、DeepSeek V4 Proは中国モデルで過去最高だが、米国勢の最新水準（GPT-5.5等）には約8か月及はないとされる。

非公開タスク（held-out）での性能低下

公開ベンチマークでは極めて高いスコアを出す一方、事前に公開されていない問題（ARC-AGI-2 semi-private等）では米国勢との差が顕著になる。



領域別の実力：追随と停滞

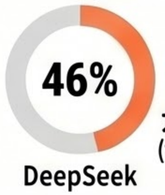
数学
(OTIS-AIME-2025)

米国勢を凌駕・肉薄
(vs 90-95%銘)

自然科学
(GPQA-Diamond)

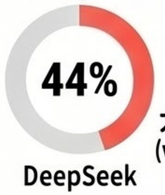
ほぼ同等
(vs 91% Opus 4.6)

抽象推論
(ARC-AGI-2)



大きな差が存在
(vs 63% Opus 4.6 / 79% GPT-5.5)

ソフトウェア移植
(PortBench)



大きな差が存在
(vs 60% Opus 4.6 / 78% GPT-5.5)

サイバーセキュリティ
(CTF)

課題が残る
(vs 46% Opus 4.6)

技術的優位性とコスト構造

超長文脈と高効率アーキテクチャ

1.6T(有効49B)のMoE設計に加え、Compressed Sparse Attention等により、1Mトークン文脈での計算量を劇的に削減。

圧倒的な破壊的低価格

V4-ProのAPI価格は、キャッシュ価格が通常価格の1/10まで下がるなど、米国勢の同等能力モデルと比較して極めて安価。

オープンエコシステム戦略

MITライセンスでのオープンウェイト配布により、Hugging Face等での採用率と粘着性を高め、開発者コミュニティを囲い込む。

導入企業が直視すべきリスク

94%のハルシネーション率

Artificial Analysisの報告によれば、「知らない時でも答えようとする」傾向が非常に強く、情報の正確性が求められる業務では注意が必要。

データガバナンスと所在

入力データやログが中国国内で処理・保管される可能性があり、規制産業や地政調達における追加精査が不可欠。

「自社Held-out評価」の必要性

公開リーダーボードの順位を鵜呑みにせず、自社の非公開データを用いた検証を行って初めて、導入の妥当性が判断できる。