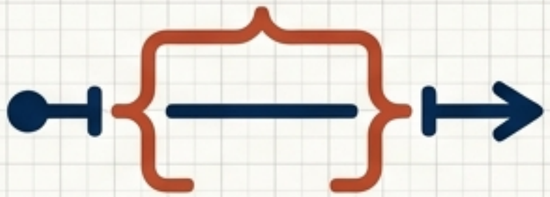
EXECUTIVE DIAGNOSTIC BRIEFING

CAISIによるDeepSeek V4 Proの評価：米中AIフロンティアの現在地

米国NISTによる最新のプライベート・ベンチマークに基づく、能力、コスト、および推論アーキテクチャの包括的分析

DATE: 2026年5月

エグゼクティブ・サマリー：主要な調査結果



能力の遅れ (The Capability Lag)

DeepSeek V4 Proはこれまで評価された中国製モデルの中で最も高性能だが、米国のフロンティアモデル（GPT-5クラス）と比較して約８ヶ月の遅れをとっている。

領域による非対称性 (Domain Asymmetry)

数学や基礎科学では高いスコアを記録する一方、サイバーセキュリティや複雑なソフトウェアエンジニアリングなど、行動を伴う「エエージェント機能」において深刻な性能低下が見られる。

コスト効率のパラドックス (The Cost Paradox)

トークン単価の安さにもかかわらず、複雑な推論タスクにおいては試行回数やトークン消費量が増加するため、最終的な「エンドツーエンドのタスク解決コスト」は米国モデルより最大41%高くなる場合がある。

グローバル・トラジェクトリー：米中フロンティアの推移

Field-note
CAISIの項目反応理論 (IRT) 評価により、DeepSeek V4 Proの総合Eloスコアは1260と算出。これは約8ヶ月前にリリースされたGPT-5とほぼ同等の水準であり、米国が依然として明確なリードを保っていることを示している。

評価の乖離：自己申告データとCAISIプライベートテストの違い

DeepSeekの自己申告 (Self-Reported)

主張:
Opus 4.6およびGPT-5.4とほぼ同等の性能。
問題点:
公開ベンチマークは学習データに混入（データ汚染）しているリスクが高く真の汎化能力を測定できない。

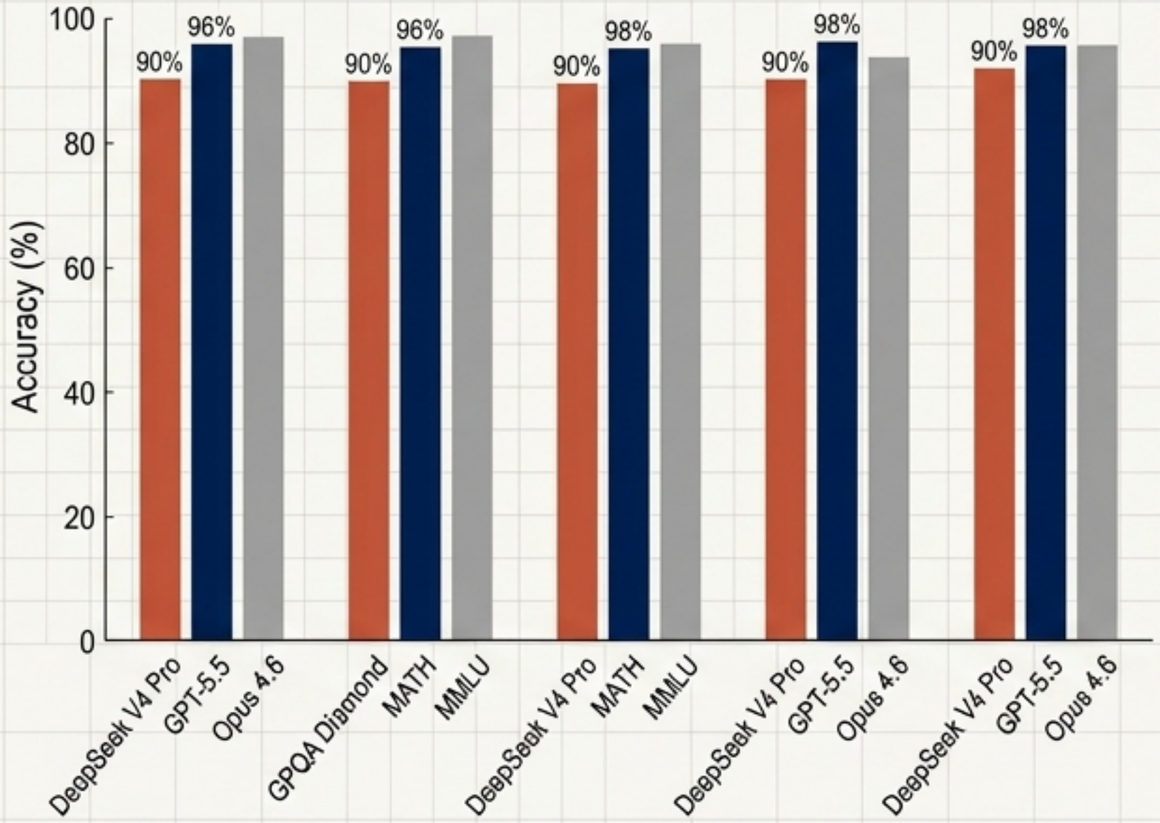
CAISIの現実 (CAISI Reality)

結果:
GPT-5と同等（8ヶ月遅れ）。

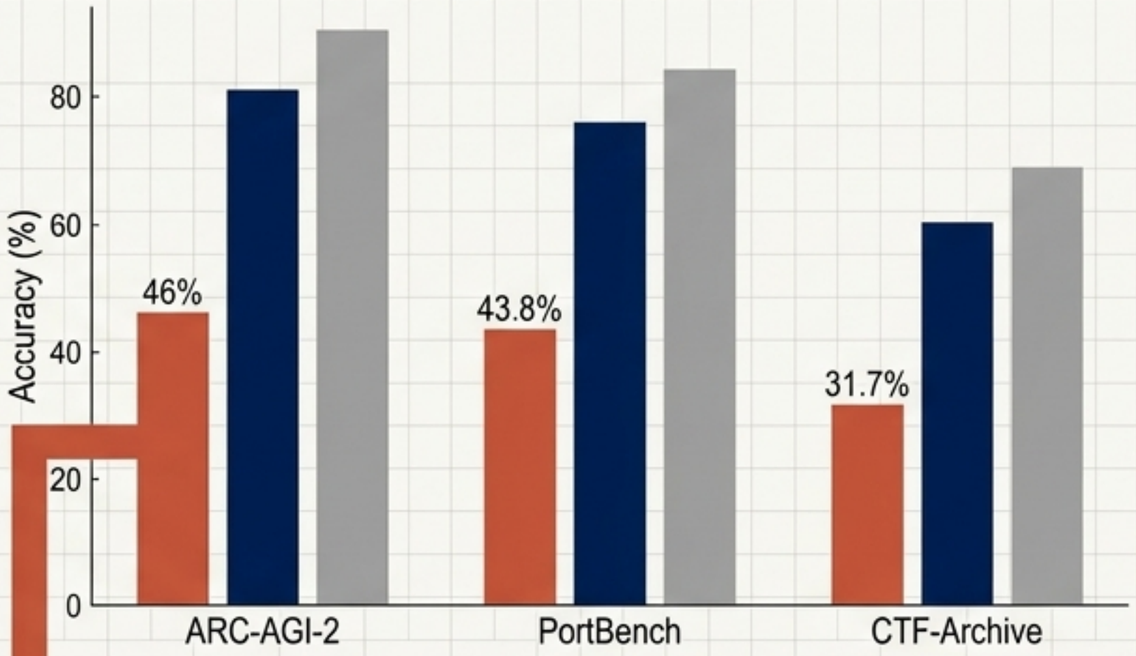
解決策:
外部データの影響を受けない、CAISI独自開発の完全非公開ベンチマーク (PortBench、CTF-Archive-Diamond等) による厳格なテストを実施。

能力の非対称性：エージェント機能における顕著な弱点

知識と推論 (Knowledge & Reasoning)



エージェント機能 (Agentic Capabilities)



数学（OTIS-AIME 97%）等では米国トップモデルに迫るが、推論や自律的行動が求められる環境（ARC-AGI-2やサイバーセキュリティ）では、性能が著しく低下する。

詳細分析：高度なサイバーおよびソフトウェアタスクの比較

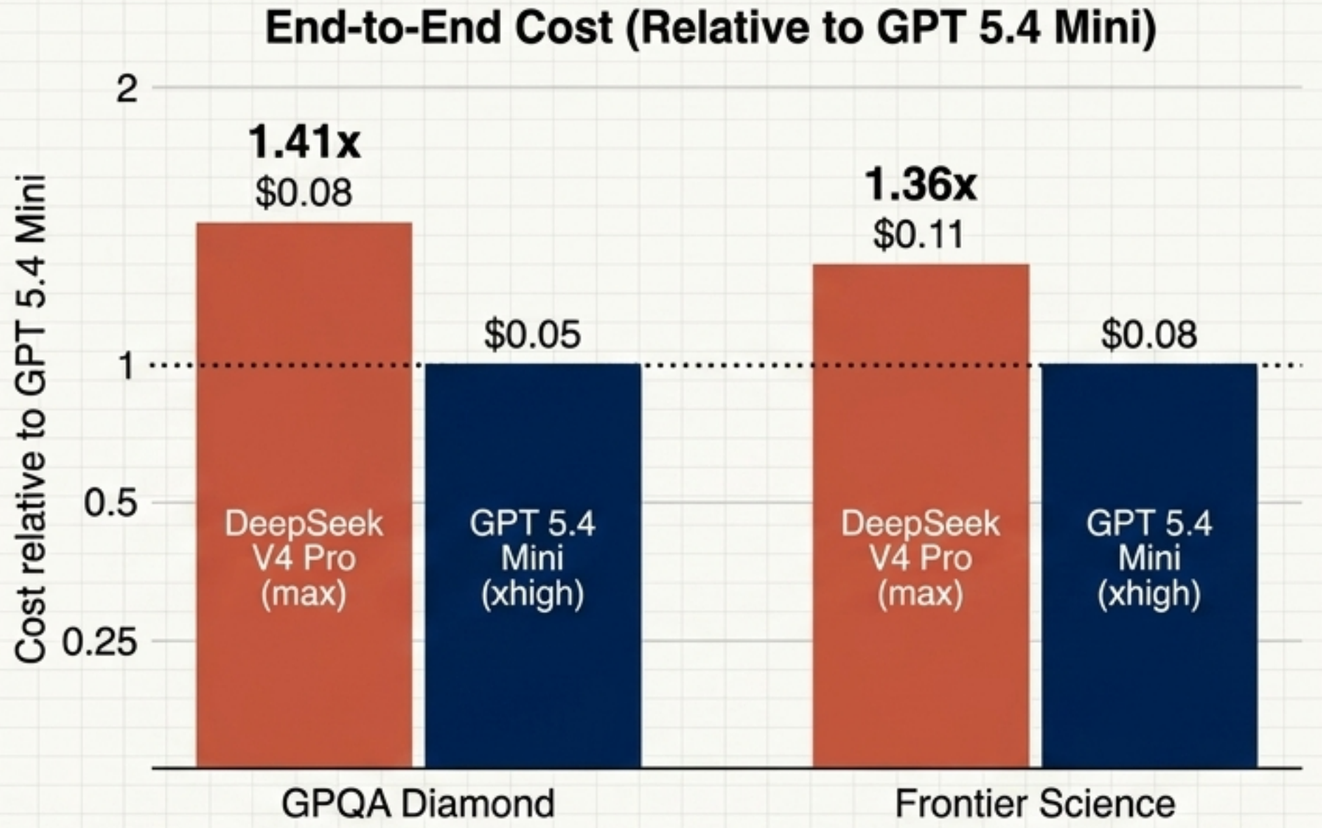
ベンチマーク (Benchmark)	OpenAI GPT-5.5 (xhigh)	Anthropic Opus 4.6 (max)	DeepSeek V4 Pro (max)
SWE-Bench検証済み	81%	79%	74%
ポートベンチ (PortBench)	78%	60%	44%
CTFアーカイブ・ダイヤモンド	71%	46%	32%

最も難易度の高い「CTFアーカイブ・ダイヤモンド」において、DeepSeek V4 Proの解決率はわずか32%。米国最先端のGPT-5.5 (71%) の半分以下の水準にとどまる。

経済的現実：入力コスト vs. エンドツーエンドのタスク完了コスト

入力コスト

DeepSeekの入力トークンは100万あたり$1.74。



トークン単価の構造にもかかわらず、複雑な推論タスク（GPQA等）では、GPT-5.4 miniと比較してタスク完了までの総コストが最大41%高騰する。正解に辿り着くためにより多くのトークン消費と試行を要するためである。

方法論の証明：CAISIはいかにして「真の知能」を測定するか

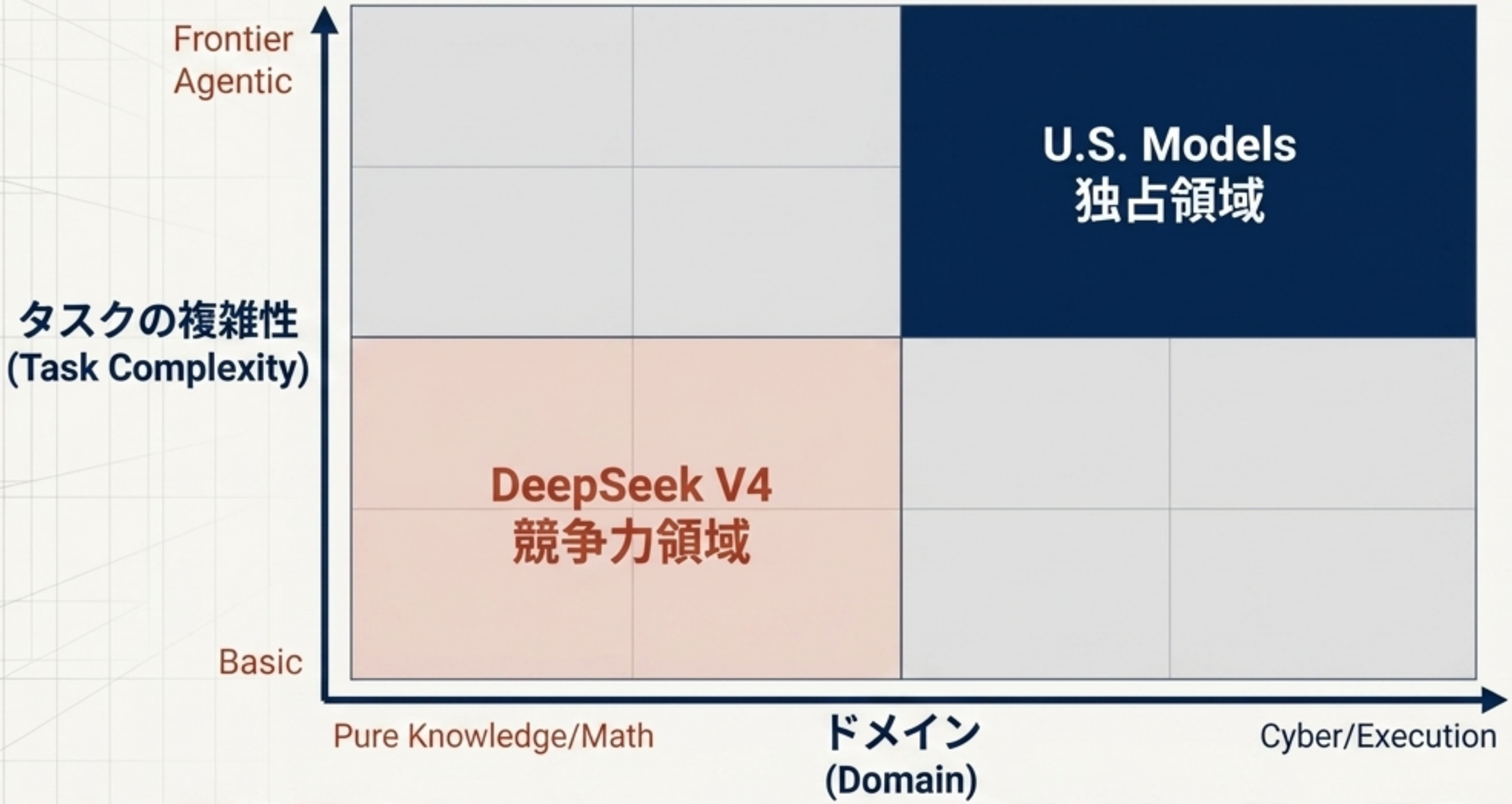
$$p_{ij} = \sigma(\theta_i - \delta_j)$$

- 単なる「正答率」の集計は、ベンチマークの難易度のばらつきにより誤解を招く。

- CAISIは項目反応理論（IRT）を採用。各タスク固有の難易度（δ）とモデルの真の能力（θ）を分離して確率モデル化（$p_{ij} = \sigma(\theta_i - \delta_j)$）。

- これにより、特定のテストへの過剰適合（ハッキング）を排除した、客観的なEloレーティングの算出が可能となる。

戦略的インプリケーション：AI導入における最適配置



基礎的な科学計算や数学的推論においては、DeepSeek V4は極めて競争力のある代替手段となり得る。

マルチステップのソフトウェア開発やサイバー防衛といった高度なエージェント機能においては、米国モデルが唯一の実用的な選択肢である。

評価環境およびソース検証

ハードウェア環境	クラウドベースのH200およびB200 GPUを使用し、開発者推奨のハイパーパラメータ（温度、top_p等）を厳格に適用。
エージェント評価	Inspectに組み込まれたReActエージェントを使用。PortBenchおよびCTF-Archive-Diamondでは、100万加重トークンの予算制限を設定。
Source	U.S. National Institute of Standards and Technology (NIST) - U.S. Center for AI Standards and Innovation (CAISI).